McGRAW-HILL

ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY

www.MHEST.com

18 SUR-TYP

Surface (geometry)

A two-dimensional geometric figure (a collection of points) in three-dimensional space. The simplest example is a plane—a flat surface. Some other common surfaces are spheres, cylinders, and cones, the names of which are also used to describe the three-dimensional geometric figures that are enclosed (or partially enclosed) by those surfaces. In a similar way, cubes, parallepipeds, and other polyhedra are surfaces. Often only the context indicates whether a surface or a solid is being referred to, but in modern mathematical usage such words refer only to surfaces. *See* CUBE; POLYHEDRON; SOLID (GEOMETRY).

Any bounded plane region has a measure called the area. If a surface is approximated by polygonal regions joined at their edges, an approximation to the area of the surface is obtained by summing the areas of these regions. The area of a surface is the limit of this sum if the number of polygons increases while their areas all approach zero. *See* AREA; CALCULUS; INTEGRATION; PLANE GEOMETRY; POLYGON.

Methods of description. The shape of a surface can be described by any of several methods. The simplest is to use the commonly accepted name of the surface, such as sphere or cube, if such a name exists. Information about a surface's shape might be only partially conveyed by the name, other information being necessary for a complete description. In mathematical discussions, surfaces are normally defined by one or more equations, each of which gives information about a relationship that exists between coordinates of points of the surface, using some suitable coordinate system. *See* COORDINATE SYSTEMS.

An equation (or equations) that defines a surface may have any of several forms. Relative to a three-dimensional rectangular coordinate system, a surface might be defined implicitly by an equation such as $F(x,y,z) = 0$ or explicitly by an equation such as $z = f(x,y)$ [here z was chosen as the dependent variable] or by parametric equations $x = x(u,v)$, $y = y(u,v)$, and $z = z(u,v)$, in which u and v are independent variables. Some surfaces are best defined by using a cylindrical or spherical coordinate system. *See* ANALYTIC GEOMETRY.

Some surfaces are conveniently described by explaining how they might be formed. If a curve, called the generator in R_3 (three-dimensional space), is allowed to move in some manner, then each position the generator occupies during this motion is a collection of points, and the set of all such points constitutes a surface that can be said to be swept out by the generator. In particular, if the generator is a straight line, a ruled surface is formed. If the generator is a straight line and the motion is such that all positions of the generator are parallel, a cylindrical surface (or just cylinder) is formed. A plane, therefore, is a ruled surface as well as a cylinder. If the generator is a straight line and all positions of the generator have a common point of intersection, a conical surface (or just cone) is formed. (To restrict the generator of a cylinder or cone from filling all

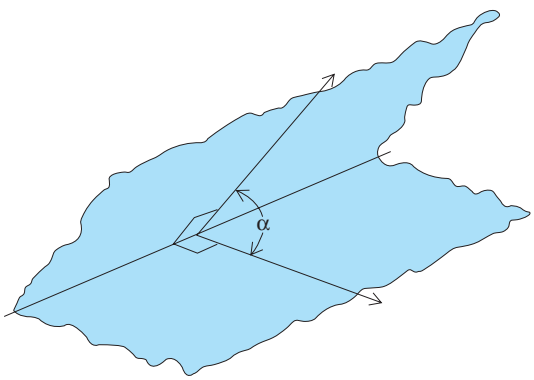


Fig. 1. Part of a dihedron (the rest has been torn away), showing dihedral angle.

of space, another condition is imposed: The generator must always contain a point of some particular curve in space.) A ruled surface that could be bent to lie in a plane (the bending to take place without stretching or tearing) is called a developable surface. Examples of developable surfaces include cylinders and cones, as well as other types. *See* CONE; CYLINDER.

Dihedron. A dihedron is the surface formed by bending a plane along a line in that plane. More formally, a dihedron is the union of two half-planes that share the same boundary line (**Fig. 1**).

A third plane, perpendicular to the boundary line, intersects the dihedral, forming an angle called the dihedral angle. Two lines that intersect generally form two supplementary pairs of equal-measure angles. The three-dimensional analog is the intersection of two planes, which define four dihedrals. Each of the four dihedral angles is either equal to, or supplementary to, the other three dihedral angles.

Quadric surfaces. A surface whose implicit equation $F(x,y,z) = 0$ is second degree is a quadric surface (**Fig. 2**), a three-dimensional analog of a conic section. A plane section of a quadric surface is either a conic section or one of its degenerate forms (a point, a line, parallel lines, or intersecting lines). With the proper choice of a rectangular coordinate system, an equation describing a quadric surface can have one of several basic forms (see **table**). *See* CONIC SECTION; QUADRIC SURFACE.

Surfaces of revolution. When a plane curve (the generator) is revolved about a line in that plane (the axis of revolution, or just axis), a surface of revolution can be said to be swept out (**Fig. 3**). The resulting surface will be symmetric about the axis of

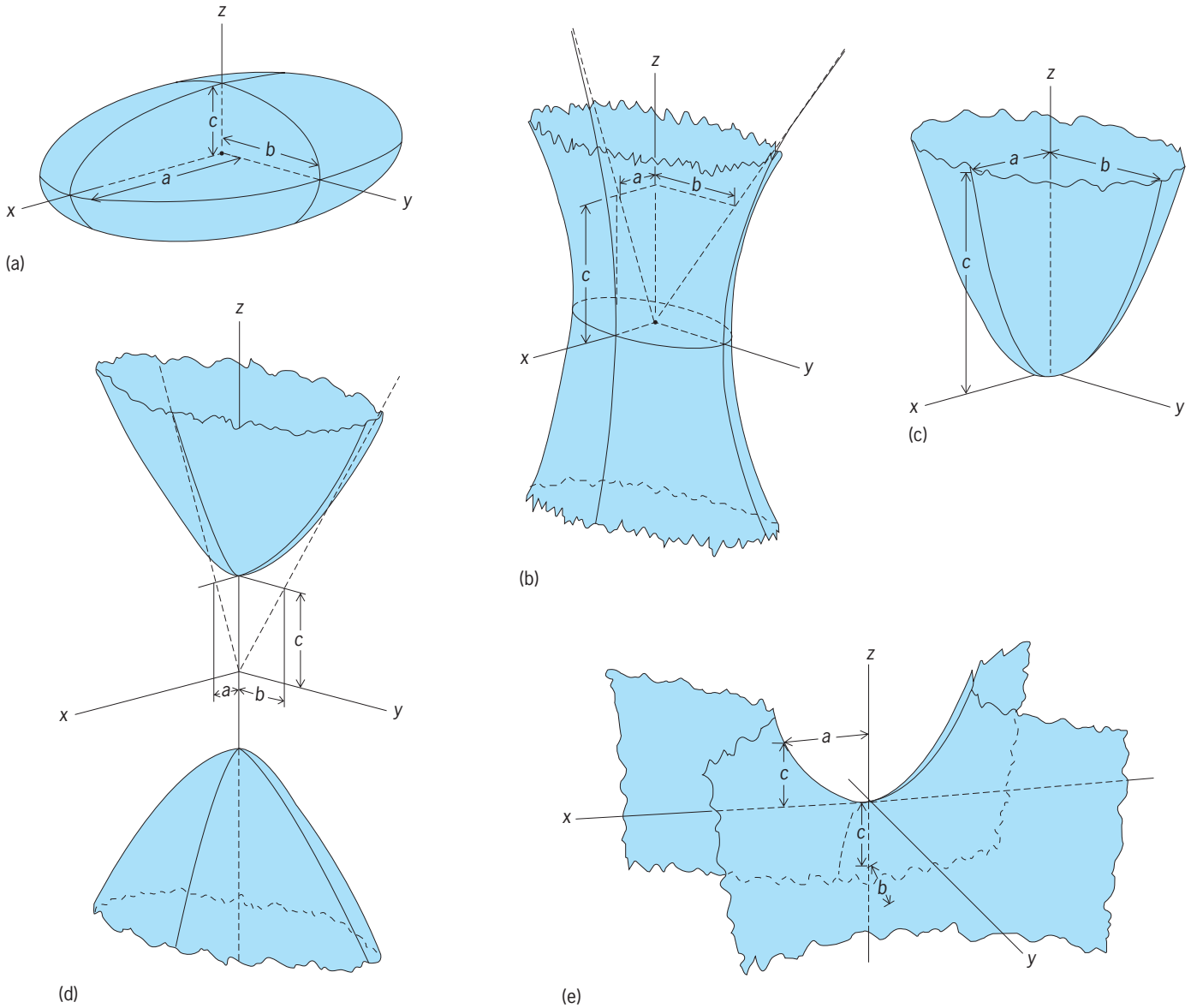


Fig. 2. Some quadric surfaces, with parameters *a*, *b*, and *c*. (*a*) Ellipsoid; $0 < c < b < a$. (*b*) Hyperboloid of one sheet; $0 < a < b < c$. (*c*) Elliptic paraboloid; $0 < a < b$ and $c > 0$. (*d*) Hyperboloid of two sheets; $0 < a < b < c$. (*e*) Hyperbolic paraboloid; $c > 0$.

Quadric surfaces

Surface type	Equation ($a, b, c \neq 0$)	Comments
Ellipsoid	$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$	If $a = b = c$, the surface is a sphere. If any two of a, b, c are equal, the surface is a spheroid.
Hyperboloid of one sheet	$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$	A surface of revolution if $a = b$. A ruled surface.
Hyperboloid of two sheets	$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1$	A surface of revolution if $a = b$.
Elliptic cone	$z^2 = \frac{x^2}{a^2} + \frac{y^2}{b^2}$	A surface of revolution if $a = b$. A ruled, and developable, surface.
Elliptic cylinder	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$	A surface of revolution if $a = b$. A ruled, and developable, surface.
Parabolic cylinder	$y = ax^2$	A ruled, and developable, surface.
Elliptic paraboloid	$\frac{z}{c} = \frac{x^2}{a^2} + \frac{y^2}{b^2}$	A surface of revolution if $a = b$.
Hyperbolic paraboloid	$\frac{z}{c} = \frac{x^2}{a^2} - \frac{y^2}{b^2}$	A ruled surface.

revolution, and this property leads to a more formal definition: A surface of revolution is any collection of points in three-dimensional space that, with a suitably chosen coordinate system, might be represented by an equation (or equations) in cylindrical coordinates (r,θ,z) in which the θ coordinate is absent. In such a representation, the z axis will be the axis of revolution. Planes perpendicular to the axis that intersect a surface of revolution will form parallel circles (or just parallels) whose centers lie on the axis; planes that contain the axis intersect the surface of revolution along congruent curves called meridian sections (or just meridians). If the generator lies in the xz plane (or yz plane) and has the equation $f(x,z) = 0$ [or $f(y,z) = 0$], an equation of the surface of revolution in cylindrical coordinates will be $f(r,z) = 0$.

There exist an unlimited number of surfaces of revolution. Mentioned below are only some of the common ones encountered in pure and applied mathematics.

Circular cylinder. A circular cylinder (a quadric surface) is formed when the generator and the axis of revolution are distinct parallel lines. A circular cylinder is an unbounded surface and so has infinite area.

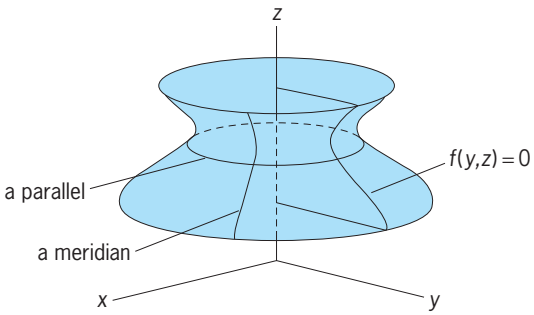


Fig. 3. Surface of revolution. The generating curve (in the *yz* plane) is $f(y,z) = 0$. The axis of revolution is the *z* axis.

However, if the generator is only a segment of a line (rather than the entire line), a bounded circular cylinder is generated. If R is the distance between the segment and the axis (this is called the radius of the cylinder) and h is the length of the segment (the height of the cylinder), then the area A of the surface is given by Eq. (1).

$$A = 2\pi Rh \tag{1}$$

Circular cone. A circular cone is a quadric surface formed when a straight-line generator intersects the axis of revolution at an acute angle. The cone consists of two parts, the nappes, joined at the point of intersection, which is the vertex of the cone. Although a cone is unbounded, a related bounded surface is formed when only a segment of a line is revolved about an intersecting line. If the segment has an end point on the axis, has length L, and makes an acute angle α with the axis, then the area A of this surface is given by Eq. (2).

$$A = \pi L^2 \sin \alpha \tag{2}$$

Sphere. A sphere (a quadric surface) is usually defined as a collection of points in three-dimensional space at a fixed distance (the radius, here denoted by R) from a given point (the center). However, a sphere can also be defined as the surface of revolution formed when a semicircle (or the entire circle) is revolved about its diameter.

The intersection of any plane with a sphere will be a circle (except for tangent planes). Such a circle is called, respectively, a great circle or a small circle, depending on whether or not the plane contains the center of the sphere. The plane area enclosed by a great circle is πR^2, and the area A of the sphere is four times that of a great circle, as in Eq. (3).

$$A = 4\pi R^2 \tag{3}$$

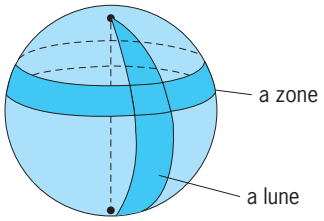


Fig. 4. Sphere, with a zone and a lune.

If only part of a semicircle is revolved about the diameter, a part of a sphere called a zone is formed (**Fig. 4**). This surface will be the same as that part of a sphere between two parallel planes that intersect the sphere. In the distance between these planes is h, the area A_z of the zone is given by Eq. (5).

$$A_z = 2\pi Rh \tag{5}$$

If a semicircle is revolved about its diameter through an angle less than one revolution, the surface swept out is a lune (Fig. 4). If θ is the radian measure of the angle through which the semicircle is rotated, the area A_l of the lune is given by Eq. (6).

$$A_l = 2\theta R^2 \tag{6}$$

See SPHERE.

Spheroid. A spheroid (also called an ellipsoid of revolution) is the quadric surface generated when an ellipse is revolved about either its major or minor axis. If the revolving is about the minor axis of the ellipse, the surface can be thought of as a flattened sphere, called an oblate spheroid. (The Earth, being slightly flattened at the poles, has roughly the shape of an oblate spheroid.) If the revolving is about the major axis, the surface can be thought of as a stretched sphere, called a prolate spheroid. (A watermelon has roughly the shape of a prolate spheroid.) *See* ELLIPSE.

Paraboloid. A circular paraboloid is the quadric surface formed when a parabola is revolved about its axis. Rays emanating from the focus of a parabola will be parallel to the axis after reflection off the parabola. This property is used to advantage by some reflecting telescopes and radio telescopes, which have surfaces with the shapes of circular paraboloids. Incoming rays parallel to the axis of revolution aimed at the concave side of the surface will be brought to a focus at the focus of the paraboloid. *See* PARABOLA; RADIO TELESCOPE.

Hyperboloid. A circular hyperboloid is the quadric surface formed when a hyperbola is revolved about either its transverse axis or its conjugate axis. The surface will be, respectively, a hyperboloid of one sheet (Fig. 2*b*) or two sheets (Fig. 2*d*), depending on whether the revolving is about the conjugate axis or the transverse axis of the hyperbola. *See* HYPERBOLA.

If one of two skew lines (lines in three-dimensional space that are not coplanar) is selected to be the axis of rotation and the other line is the generator, the surface swept out is a hyperboloid of one sheet. (The usual definition of a surface of revolution must be extended, by allowing the axis and the generating curve to lie in different planes.) Every point on a hyperboloid of one sheet is the point of intersection of two lines, each point of which lies on the hyperboloid. Either of these lines (which define equal angles with the axis of revolution) could be used as the generator.

Torus. A torus is generated when a circle is revolved about a line that does not intersect the circle. This doughnut-shaped surface has the property that not all points on the surface have the same sign of curvature. If R is the distance between the axis and the center of the generating circle, then the torus has respectively negative curvature, zero curvature, or positive curvature at points on the torus closer to the axis than R, at a distance R, or farther from the axis than R. *See* DIFFERENTIAL GEOMETRY; TORUS.

Catenoid. A suspended uniform slack rope has the shape of the curve called a catenary. If the axis of revolution is a horizontal line that underlies the suspended rope, then, using the catenary as the generator, the surface of revolution formed is a catenoid (**Fig. 5**). *See* CATENARY.

A soap film tries to attain a shape that minimizes the surface area subject to necessary constraints. An airborne soap bubble is spherical because that is the surface of minimum area that encloses a given volume of air. If two wire circles define the bases of a frustrum of a circular cone, and if the height of the frustrum is small, then when withdrawn from a soapy solution the soap film along the side of the wire-frame frustrum will form a surface of revolution having minimum area: a catenoid. *See* MINIMAL SURFACES.

Pseudosphere. A tractrix is the path followed by an object dragged slowly through a resisting medium by a constant-length cord, where the other end of the cord follows a specified path. If the cord is one unit long, and the end follows a straight-line path perpendicular to the initial position of the cord, and if this curve is revolved about that line, then a pseudosphere is generated. This surface has the property that its curvature is a negative constant at every nonsingular point of the surface. *See* TRACTRIX.

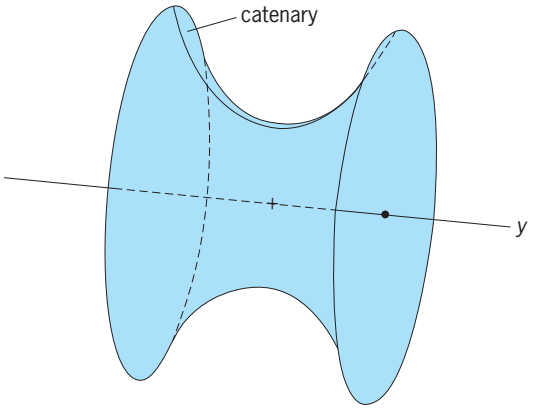


Fig. 5. Catenoid. The axis of revolution is the *y* axis.

Area. If ds is the element of arc length of a generator, and R represents the distance of that element of length from the axis of revolution, then the surface area A is given by Eq. (7), where a and b define

$$A = \int_a^b 2\pi R ds \tag{7}$$

the end points of the generator. When attempting to evaluate an area, the best variable of integration to use is normally determined by how the generator is described relative to the coordinate system chosen.

An alternative method for determining the areas of some surfaces of revolution is by using the surface theorem of Pappus: If a plane curve is revolved about a line that does not intersect the curve (except perhaps at an end point), then the area of the surface of revolution generated is equal to the length of the generating curve multiplied by the distance that the centroid of the curve moves during one revolution. To visualize the centroid of a plane curve, it may be imagined that the curve is made of a uniform thin wire, and a thin massless membrane encloses the wire. If the wire-membrane assembly is held with its containing plane horizontal, then the centroid is the so-called balance point, where the assembly could be balanced on a pencil point without tipping in any direction. It is necessary to know the distance of the centroid from the axis of revolution in order to use this theorem of Pappus, and usually it is just as much work to find this distance as it is to compute the surface area by integration. However, in some cases this distance can be easily found by using symmetry arguments. For example, the surface area of a torus can be computed with this theorem of Pappus by recognizing that the centroid of the generating circle is at the center. If d is the radius of the generating circle, the arc length is $2\pi d$; if R is the distance between the centroid and the axis of revolution, the distance traveled by the centroid during one revolution is $2\pi R$. The surface area A of a torus, therefore, is given by Eq. (8). Areas of other surfaces of revolution can be

$$A = 4\pi^2 dR \tag{8}$$

found by similar arguments. Harry L. Baldwin, Jr.

Bibliography. H. G. Ayre and R. Stephens, *A First Course in Analytic Geometry*, 1956; J. S. Frame, *Solid Geometry*, 1948; R. R. Middlemiss, J. L. Marks, and J. R. Stewart, *Analytic Geometry*, 3d ed., 1968; D. C. Murdoch, *Analytic Geometry, with an Introduction to Vectors and Matrices*, 1966; C. Smith, *An Elementary Treatise on Solid Geometry*, 1910; D. J. Struik, *Lectures on Classical Differential Geometry*, 1950, reprint 1988.

Surface-acoustic-wave devices

Devices that employ surface acoustic waves in the analog processing of electronic signals in the frequency range 10^7–10^9 Hz. Surface acoustic waves are mechanical vibrations that propagate along the surfaces of solids. In 1885, Lord Rayleigh discovered a type of surface acoustic wave that contains both compressional and transverse components 90° out of phase with one another. Since that time, other types of surface acoustic waves have been discovered and are an active area of intense research. A few notable examples include a wave propagating along a layer on a surface (Love wave), a wave propagating along an interface between two solids (Stoneley wave), and transverse guided waves on solids (Bleustein-Gulyaev-Shimizu waves). Love waves are shear-horizontal (SH) waves that have displacement only in a direction perpendicular to the plane of propagation.

Piezoelectric materials provide the required coupling between electrical signals and mechanical motion to generate surface acoustic waves. Crystalline piezoelectric materials, such as quartz, lithium niobate, and lithium tantalate, exhibit low attenuation and dispersion, and are therefore ideal for acoustic propagation. Surface acoustic waves in such a material are generated through a localized electric field at the surface that is created by applying voltage to an array of metal electrodes or fingers. This electrode array is known as an interdigital transducer (IDT). The IDT can also be used to detect surface waves, producing electrical output and hence an overall response. *See* PIEZOELECTRICITY.

Surface-acoustic-wave (SAW) devices have led to a versatile technology for analog signal processing in the frequency range 10^7–10^9 Hz. The much slower propagation velocity of acoustic waves as compared to electromagnetic waves permits time delays in SAW devices, as compared to electrical delay lines, that are crucial for signal processing applications. Notable devices include band-pass filters, resonators, oscillators, pulse compression filters, fast Fourier transform processors, and more recently chemical and biological sensors. Consumer application areas include mobile phones, television and satellite receivers, keyless entry systems (garage doors, cars, and so forth), and wireless applications. Commercial applications include fiber-optic communication, oscillators, local-area networks (LANs), test equipment, and chemical and biological detection systems, with military applications in radar, sonar, and advanced communications. *See* ELECTRIC FILTER; LOCAL-AREA NETWORKS; MOBILE COMMUNICATIONS; OPTICAL-FIBER COMMUNICATIONS; OSCILLATOR; RADAR; SIGNAL PROCESSING; SONAR; TELEVISION RECEIVER.

SAW transduction. A basic SAW device known as a delay line is shown in **Fig. 1***a*. A piezoelectric substrate has a polished upper surface on which two IDTs are deposited using photolithographic methods. The left-hand input transducer is connected, via fine bonded leads, to the electric source (V_s) through an electrical matching network (Z) and source resistance (R_s). The right-hand output transducer drives the load (R_L), usually 50 ohms, through another electrical matching network (Z). Advances in computer modeling have led to the development of 50-ohm IDT designs that do not require external electrical matching networks. The center

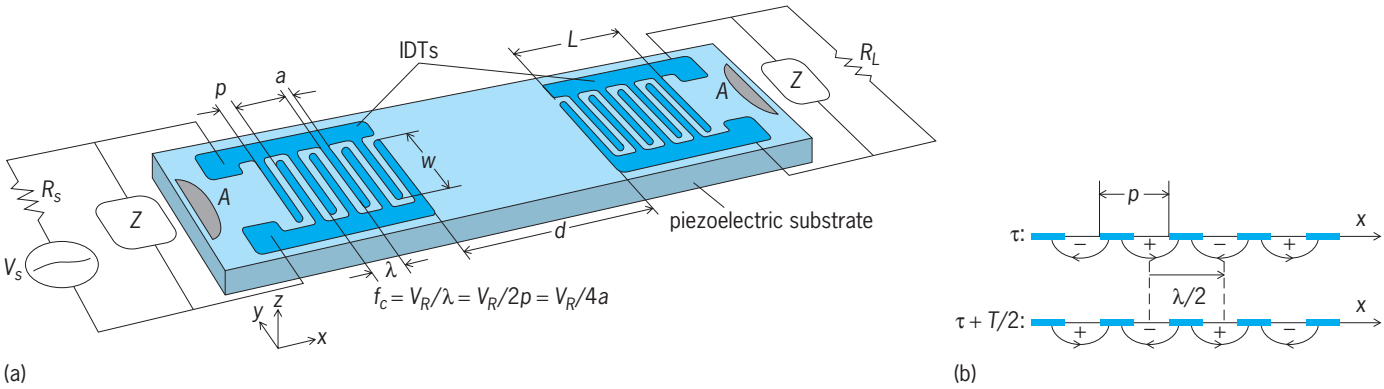


Fig. 1. Operating principle of a surface-acoustic-wave delay line. (*a*) Device layout. (*b*) Cross section of substrate and transducers, showing electric fields at times τ and $\tau + T/2$.

frequency (f_c) is governed by the Rayleigh wave velocity (V_R) on the piezoelectric substrate and the electrode width (a) of a single finger, according to Eq. (1). For SAW devices, the velocity of the wave de-

$$f_c = \frac{V_R}{4a} \tag{1}$$

pends on the properties of the piezoelectric crystal and its crystallographic orientation. Computer models have proven essential to iterate through numerous crystallographic orientations to search for the existence and type of a suitable acoustic wave. *See* DELAY LINE.

For filter applications, the width of the passband, which is governed by length of the IDT (L), is critical for frequency selection. Increasing L by the addition of more electrode pairs sharpens the filter response and reduces noise. IDT-to-IDT spacing (d) is used to select the delay time or phase slope of the filter. The IDT aperture (w) governs the diffraction behavior and determines transducer output power. Owing to symmetry, each transducer generates acoustic waves equally in two opposite directions, so that it is bidirectional. In this case, half of the power propagates in an unwanted direction, giving a loss of −3 dB [that is, $10 \cdot \log_{10}(0.50)$], and in a delay line with two IDTs this propagation contributes to −6 dB of insertion loss in the passband. Though waves are terminated by using an absorber (A), bidirectional emission is undesirable. (Absorber efficiency depends on the acoustic absorption properties of the material and the type of acoustic wave. For Rayleigh SAW devices, silicon rubber is a very effective absorber, reducing the amplitude of the backward-traveling waves by over 30 dB. However, SAW devices that use shear-horizontal waves require unidirectional transducers since such waves are not attenuated by absorbers.) In advanced SAW devices, through the use of unidirectional transducers, acoustic waves propagate preferentially in one direction, which dramatically reduces overall acoustic loss and eliminates the need for absorbers. Insertion losses for unidirectional transducers are around −4 dB or better, depending on the substrate type.

The voltage across the electrodes produces an electric field (Fig. 1*b*), which generates compressions and expansions near the surface. These give rise to various elastic waves. A sinusoidal supply voltage (V_s) produces vibrations that add constructively only if the distance (p) equals half the wavelength (λ). Stresses produced at time τ by a pair of fingers, for a given polarity of the voltage, travel distance $\lambda/2$ during half-period $T/2$ at the speed (V_R) of the Rayleigh wave (Fig. 1*b*). At time $\tau + T/2$, the stress arrives under the neighboring pair of fingers, where the voltage has just changed sign, producing a stress with the same phase. The stress due to the second pair of fingers adds constructively to the first.

The acoustic response at frequency f for the SAW delay line in Fig. 1*a* can be calculated approximately by regarding each IDT as having N electrodes or sources. Through summation of sources for their amplitude and phase, the frequency response can be determined as proportional to $|(\sin x)/x|$ [that is, $|\sin c(x)|$], where x is given by Eq. (2), providing a

$$x = \frac{N\pi(f - f_c)}{f_c} \tag{2}$$

band-pass filter characteristic. The electrical matching networks are normally arranged to minimize filter loss without compromising the acoustic response. The optimum number of periods, N, is inversely proportional to the piezoelectric coupling as is the filter band-pass width. *See* TRANSMISSION LINES.

Numerous IDT geometries have been designed to achieve optimal performance on a wide variety of piezoelectric materials. The simplest designs use single split electrode geometries (**Fig. 2***a*). A slightly more complex design uses a double split electrode configuration (Fig. 2*b*) to reduce reflection effects in the passband. In Fig. 2*a*, waves reflected from corresponding edges of two neighboring fingers of width $\lambda/4$, separated by an interval of $\lambda/4$, add constructively because the path difference causes a phase shift of 2π. The design in Fig. 2*b* suppresses reflections since each finger is composed of two strips of width $\lambda/8$. In this case the distance between two neighboring fingers produces a path difference of $\lambda/2$, resulting in destructive interference. However, the double split electrode configuration demands

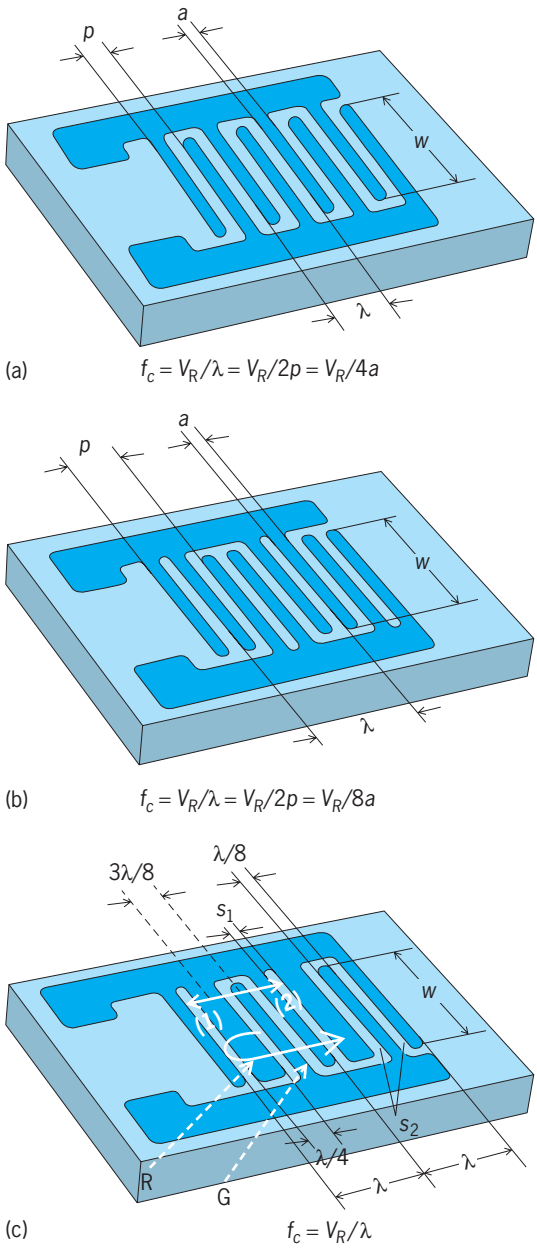


Fig. 2. Transducer arrangements based on (*a*) single split electrode, (*b*) double split electrode, and (*c*) Electrode-Width-Controlled/Single-Phase Unidirectional Transducer (EWC/SPUDT).

a higher photolithographic resolution to fabricate, since the IDT finger width (a) must reduced by a factor of 2 to achieve the same operating frequency as the layout of Fig. 2*a*. For example, for a 500-MHz filter on ST-quartz (where ST is the cut plane of the quartz) having Euler angles of (0°, 132.75°, 0°), with a velocity of $V_R = 3158$ m/s for x propagation in the x direction, the width (a) for the single split IDT is 1.6 μm, whereas for the double split IDT the width (a) reduces to 0.8 μm. [The angles given as (φ, θ, ψ) are referred to as Euler angles and used to designate the piezoelectric crystal cut. Euler angles describe the rotation of the crystal surface referenced to the crystal axes (X, Y, Z). Three main properties of the crystal must be transformed to the rotated axes using Euler angles to determine the SAW propagation behavior. These three material parameters are the elastic, piezoelectric, and permittivity constants of the crystal. The crystal density is not dependent on the rotation. By using the equation of motion and Laplace's equation, the SAW propagation velocity and wave type can be calculated. *See* CRYSTALLOGRAPHY; EULER ANGLES.

Figure 2*c* shows a common unidirectional transducer design that preferentially launches acoustic waves to the right. This IDT uses variable pitch (that is, finger-to-finger spacing) and finger width to achieve a sophisticated response. By using a reflection center (R) located at the middle of the $\lambda/4$ electrodes, a -90° phase shift is introduced. Each generation center (G) launches waves designated (1) and (2) in both directions. By setting the distance between the reflector (R) and the launching electrode (G) to $3\lambda/8$, the waves moving to the left return in phase with the rightward-moving waves. In practice, the spaces s_1 and s_2 and the distance from G to R are optimized using computer simulation methods to achieve the best performance (for example, a highly linear phase response). *See* TRANSDUCER.

Piezoelectric substrates. SAW devices are fabricated using methods found in standard semiconductor processing facilities, such as photolithography, thin-film evaporation, and chemical etching techniques. Common piezoelectric substrates are ST-cut (0°, 132.75°, 0°) quartz, X-propagating for temperature stability; Y-cut Z-propagating (YZ) lithium niobate (0°, 90°, 90°) for high piezoelectric coupling; 128° Y-cut X-propagating lithium niobate (0°, 38°, 0°) for reduced bulk-wave excitation as compared with YZ lithium niobate; and gallium arsenide (GaAs) for compatibility of SAW devices with integrated circuits. *See* INTEGRATED CIRCUITS; MICROLITHOGRAPHY.

The **table** lists some common substrates for Rayleigh and leaky-SAW substrates. Rayleigh waves are called true surface waves since the propagation velocity is a real number. However, leaky waves refer to propagation velocities that are complex numbers, where the imaginary term contributes to additional loss within the crystal. In either case, the coupling efficiency is determined by the electromechanical coupling coefficient K^2, which is a measure of the efficiency for a specific piezoelectric substrate and orientation to convert an applied electrical signal into mechanical motion.

Current commercial photolithography is limited to approximately 0.3–0.5-μm line resolution, which limits operating frequency of Rayleigh-wave devices to about 2 GHz. This limitation has prompted the study of alternative wave-propagation mechanisms for use with existing photolithography. When acoustic waves with a higher velocity are used, the line resolution is not as critical. Leaky-SAW (LSAW) velocities can be much higher than SAW ones to the extent that LSAW devices can be designed for operation at frequencies up to 1.6 times higher than their SAW counterparts using the same photolithographic

Parameters of common piezoelectric substrates for SAW devices

Material	Crystal cut	SAW axis	Velocity, m/s	K^2, %	Temperature coefficient, ppm/°C	Applications
			Rayleigh-wave devices			
Quartz	ST	*X*	3158	0.11	~0	Precision oscillators, low-loss radio-frequency (RF) resonators
Lithium niobate	*Y*	*Z*	3488	4.5	94	Wideband intermediate-freqeuncy (IF) filters
Lithium niobate	128°	*X*	3992	5.3	75	Wideband IF filters
Gallium arsenide	(100)	<110>	2481	0.06	35	Semiconductor integrated circuit compatibility
Bismuth germanium oxide	(110)	<001>	1681	1.4	120	Long delay lines
			Leaky-wave devices			
LST-quartz	ST, $\theta = 15.7°$	*X*	3948	0.11	~0	Gigahertz-range precision oscillators, low-loss RF resonators
Lithium niobate	64°	*X*	4478	11.3	−81	Band-pass RF filters, mobile transceivers
Lithium niobate	41°	*X*	4840	21	−65	Band-pass RF filters, RF resonators
Lithium tantalate	36°	*X*	4112	4.7	−32	Ladder filters for antennas, mobile transceivers

geometry. Leaky surface acoustic waves propagate just beneath the piezoelectric surface, in the form of surface-skimming bulk waves (SSBW) for a free surface (that is, for an electrically open surface), and in the form of shear-horizontal waves for an electrically shorted surface. The type of propagation is a complex function of the crystal type, cut, and symmetry. LSAW devices have found extensive application in radio-frequency-filter devices for mobile and cellular phone systems into the 2-GHz regime. Some common LSAW piezoelectric substrates are LST-quartz (0°, 15.7°, 0°), 64° *YX*-lithium niobate (0°, 64°, 0°), 41° *YX* lithium niobate (0°, 41°, 0°), and 36° *YX* lithium tantalate (0°, 36°, 0°). In **Fig. 3**, an example of an optimized LSAW device is given using 36° *YX* lithium tantalate. This design implements an IDT similar to Fig. 2*c*, and the use of unidirectional transducers enables it to achieve an insertion loss of −4.8 dB at the center frequency, an improvement over the insertion loss of −6 dB when bidirectional IDTs are used.

Band-pass filters. Conventional (lumped-element) linear-phase passive *LC* filters all have some inherent degree of phase nonlinearity over their prescribed frequency range, and this nonlinearity increases with the order of the filter (that is, the number of reactive components). The resultant size of a passive *LC* filter, combined with cost and complexity, is often not suitable for integrated and mobile systems. However, the size of a SAW filter decreases with increasing frequency, with only the limitation of operating as a band-pass filter as opposed to a low-pass filter. In their band-pass operation, SAW filters have much more design versatility than *LC* band-pass filters. In contrast to conventional *LC* filters, SAW filters have the property that linear phase response can be achieved independently of the amplitude response and the amplitude response can be shaped to be asymmetric about its center frequency. Further, SAW filters are transversal filters in which the signal is repetitively delayed and added to itself, as in antenna arrays and digital filters. A design procedure used to model SAW filters is to compute the inverse Fourier transform of the prescribed filter response, giving the impulse response, which is the desired spatial image for the transducer. Due to finite piezoelectric substrate lengths, an infinitely long time-duration impulse response is not realizable. This limitation has necessitated the use of weighting functions to multiply and truncate the impulse response. *See* ANTENNA (ELECTROMAGNETISM); FOURIER SERIES AND TRANSFORMS.

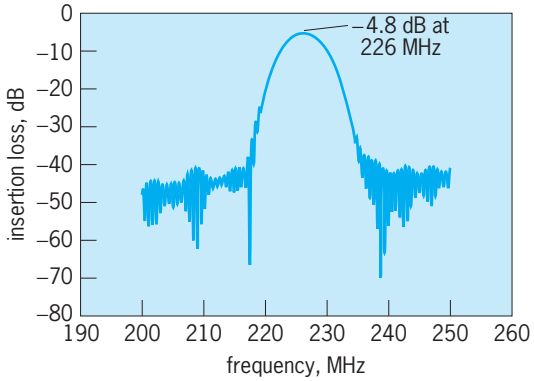


Fig. 3. Insertion loss for leaky-SAW device using 36° *YX* lithium tantalate.

A crucial factor in SAW filters is amplitude weighting, that is, alteration of the amplitude-frequency response to improve filter response. The SAW design in Fig. 1*a* has sidelobe suppression of about −24 dB below the main peak. Instead of using the uniform IDT with constant apodization (that is, finger-length overlap), the overlap of the fingers can be variable. If the apodization of the IDT is changed to a sinc function, the inverse Fourier transform will now approximate a rectangular band-pass response from the filter. This increases the sidelobe suppression to −30 dB,

but produces undesirable amplitude and phase ripple in the passband from the Gibbs phenomena as a result of having finite electrode length for the SAW aperture (w) in Figs. 1 and 2. The use of finite electrodes translates to termination of an infinite series expansion of a mathematical expression used to describe the IDT. Early termination of the infinite series results in the Gibbs phenomena, which occur in all orders of approximation. The result overshoots the original function.

To achieve a rectangular response without using excessively long apodized IDTs, windowing function techniques can be used to modify the IDT apodization pattern. Window functions are widely used in digital filter design to improve the shape of the passband response. Finite impulse response length corresponds to the truncation of an infinite Fourier series. This abrupt truncation of a Fourier series causes the Gibbs ripple phenomena. Window function techniques circumvent the problem of using excessive impulse response lengths by convolution of the IDT sinc function with a chosen window function (such as, Kaiser, Hamming, or cosine). Desirable window functions are those that have a narrow main frequency response and sidelobes that decrease rapidly with frequency. As a result, windowing function methods nearly eliminate passband ripple, while achieving high sidelobe suppression (greater than 45 dB). *See* DIGITAL FILTER.

Resonators. SAW resonators consist of either one or two IDTs bounded by two reflection gratings. One-port SAW resonators employ a single IDT for input and output, in conjunction with two SAW reflection gratings (**Fig 4**). Two-port SAW resonators can be formed using separate IDTs for input and output between two reflection gratings. In one-port SAW resonators, surface waves emitted from both sides of the excited IDT are constructively reflected at the center frequency by reflection gratings (R), which give rise to standing surface acoustic waves within the IDT.

Elements of the reflection grating are both periodically spaced and normal to the SAW propagation direction, to create narrow-band performance. Since the SAW grating reflects surface waves from both sides of an excited bidirectional IDT, the device insertion loss can be less than −6 dB. In practice, the reflection gratings can be either open or shorted metal strips, shallow grooves in the substrate, or shallow grooves filled with a metal such as gold to increase reflection efficiency. The key operation principle for SAW resonators is the phase of the grating reflection coefficient. Acoustic resonance occurs in a SAW resonator when the total phase shift φ of the surface wave is $2n\pi$ ($n = 1,2,\ldots$) within the cavity bounded by the reflection gratings. This is analogous to a microwave resonant cavity, where energy oscillation is between electric and magnetic fields, except in the acoustic case where the oscillation is between mechanical stress and strain fields. The resonator quality factor depends on the position of the IDTs, the reflection efficiency, and the cavity losses from material properties. Both resonant and antiresonant behavior

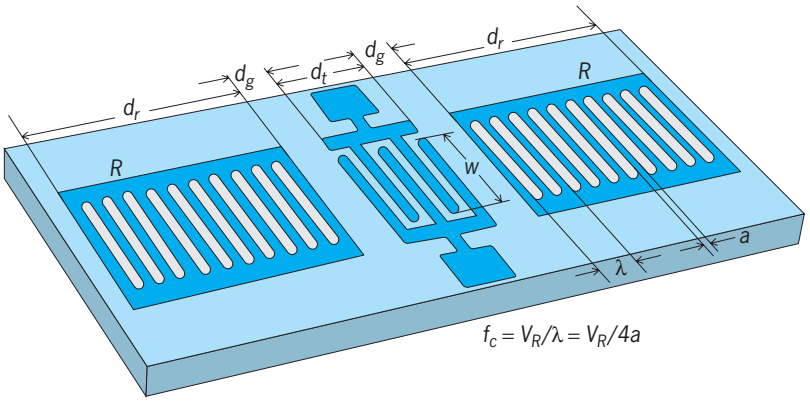


Fig. 4. One-port SAW resonator design. The key dimensional parameters are d_r, d_g, d_t, a, and w. The grating reflectors (R) are designed to create standing waves, forming a resonant cavity.

can be obtained, depending on whether the spacing between the arrays is an even or odd integral number of half-wavelengths of the surface acoustic wave. *See* CAVITY RESONATOR.

In SAW resonator applications, it is crucial to maintain short- and long-term frequency stability. For this reason, special care is required to isolate a SAW device from environmental vibration and temperature change. A critical aspect in packaging SAW resonators is the use of materials that isolate the device from vibration. Packaging methods often employ special support structures to achieve a high degree of vibration isolation. For ST-cut quartz SAW resonators, the vibrational sensitivity is about $\Delta f/f = 1 \times 10^{-9}/g$ [that is, fractional change in frequency per g (9.8 m/s^2 or 32.2 ft/s^2)]. For commercial resonators, the frequency tolerance at 25°C (77°F) ranges from about ±20 ppm to ±300 ppm depending on cost. For one-port SAW resonators in the range 100–1200 MHz, matched insertion loss is in the range 0.5–2.5 dB, with quality factor (Q_L) ~ 1600–7000. *See* VIBRATION ISOLATION.

In precision resonators, the effect of noise is a serious concern, and efforts are made to minimize noise through use of optimized designs and circuits. Flicker ($1/f$) noise is a type of noise that occurs in high-precision oscillators, causing the central frequency peak to broaden as a result of phase noise. One of the most common ways to measure phase noise is with a spectrum analyzer. Noise and spurious signals appear as jitter on either side of the output signal's center frequency on a spectral density plot. On this type of plot, pure frequencies would appear as a single spectral line, but this is not physically possible and instead phase noise causes broadening of the spectral line. Phase noise is specified in decibels relative to the carrier at a given frequency offset from the carrier (dBc/Hz). For example, if we shift 1 Hz from the carrier peak to the right, the dB value describes how far down the noise is compared to the signal. Attainable flicker noise ($1/f$) levels are −140 dBc/Hz at a frequency offset of 1 Hz. Over the same, frequency range, two-port resonator losses are 1–4 dB, and Q_L is ~3000–13,000. Two-port SAW

resonators have been reported at 2.6 and 3.3 GHz, with insertion loss less than 11 dB, $Q_L = 2000$, and insertion loss of 17 dB, $Q_L = 1600$, respectively. *See* SPECTRUM ANALYZER.

For commercial applications, one-port SAW resonators are useful for replacing bulk-wave resonators operating at high-overtone modes and for applications in narrow-band intermediate-frequency (IF) stages for analog cellular phone systems. Two-port resonators are used in oscillator applications for both fixed-frequency and tuning applications. The majority of SAW resonators are Rayleigh-wave designs fabricated on cuts of quartz for temperature stability.

SAW devices for chemical and biological detection. A more recent use of SAW devices is for chemical and biological detection. Since the acoustic energy is confined at the surface of the piezoelectric substrate, surface acoustic waves are highly sensitive to surface perturbations of the propagating medium. The boundary conditions at the solid or liquid interface govern the wave amplitude and velocity, allowing the surface wave device to respond to any shift in mechanical and electrical properties of the contacting medium. Chemical detection systems based on Rayleigh-wave SAW devices have largely used ST-quartz for its excellent temperature stability. To achieve specificity to a particular chemical agent, select polymer films are applied to the surface of the SAW device. Signal transduction occurs when chemicals partition into the polymer film, thereby altering the film properties, causing a change in the electrical response (that is, phase, frequency, time delay, or amplitude). For SAW resonators, acoustic loss (for example, due to fluids) at the surface contributes to a reduction of the Q, causing broadening of the response, and therefore degrades the detection limit.

SAW fluid detection systems require the use of piezoelectric substrates that minimize coupling of acoustic energy into fluids. To achieve this, a substrate must support shear-horizontal SAW (SH-SAW) excitation. A widely used substrate is the leaky-SAW substrate of 36° *YX* lithium tantalate, which can be converted to an SH-mode device by electrically shorting the surface where the acoustic wave propagates by application of a thin metal film. To increase sensitivity and electrically insulate the surface of SH-SAW devices for fluid-based biodetection systems, a thin layer or waveguide (often silicon dioxide or polymers) is applied. This creates waves that are dispersive or frequency-dependent, known as Love waves. Love-wave biosensors have been reported with detection limits of ∼10–800 picograms/cm^2. *See* SEISMOLOGY.

Darren W. Branch

Bibliography. C. K. Campbell, *Surface Acoustic Wave Devices for Mobile and Wireless Communications*, Academic Press, San Diego, 1998; D. Royer and E. Dieulesaint, *Elastic Waves in Solids I: Free and Guided Propagation*, Springer-Verlag, Berlin, 2000; D. Royer and E. Dieulesaint, *Elastic Waves in Solids II: Generation, Acousto-optic Interaction, Applications*, Springer-Verlag, Berlin, 2000.

Surface and interfacial chemistry

Chemical processes that occur at the phase boundary between gas–liquid, liquid–liquid, liquid–solid, or gas–solid interfaces.

The chemistry and physics at surfaces and interfaces govern a wide variety of technologically significant processes. Chemical reactions for the production of low-molecular-weight hydrocarbons for gasoline by the cracking and reforming of the high-molecular-weight hydrocarbons in oil are catalyzed at acidic oxide materials. Surface and interfacial chemistry are also relevant to adhesion, corrosion control, tribology (friction and wear), microelectronics, and biocompatible materials. In the last case, schemes to reduce bacterial adhesion while enhancing tissue integration (**Fig. 1**) are critical to the implantation of complex prosthetic devices, such as joint replacements and artificial hearts. Selected technologies that involve surfaces and interfaces are listed in **Table 1**. *See* BIOMEDICAL CHEMICAL ENGINEERING; CRACKING; HETEROGENEOUS CATALYSIS; PROSTHESIS; SURFACE PHYSICS.

Interactions with the substrate may alter the electronic structure of an adsorbate. Those interactions that lower the activation energy of a chemical reaction result in a catalytic process. Adsorption of reactants on a surface also confines the reaction to two dimensions as opposed to the three dimensions available for a homogeneous process. The two-dimensional confinement of reactants in a bimolecular event seems to drive biochemical processes with higher reaction efficiencies at proteins and lipid membranes. *See* ADSORPTION.

A limitation in the study of surfaces and interfaces rests with the low concentrations of the participants in the chemical process. Concentrations of reactants at surfaces are on the order of 10^{-10} to 10^{-8} mole/cm^2. Such low concentrations pose a sensitivity problem from the perspective of surface analysis. Experimental techniques with high sensitivity are required to examine the low concentrations of a surface species at interfaces (**Fig. 2**).

Electron surface analysis techniques. Electron spectroscopy methods are widely used in the study of

TABLE 1. Some typical surface and interfacial processes

Interface	Processes	Significance
Liquid–liquid	Solute partitioning	Solvent extraction
Gas–solid	Adsorption/desorption Surface-mediated reactions	Catalysis
Liquid–solid	Electrochemistry Molecular recognition	Solar energy conversion Energy transfer and storage in cell membranes DNA sequencing

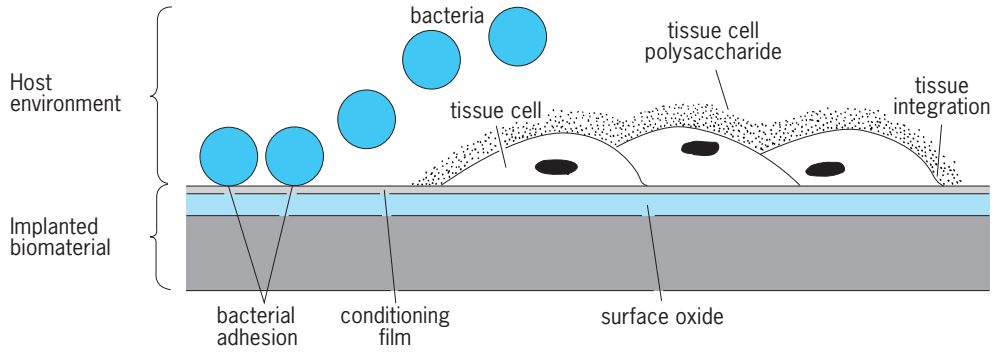


Fig. 1. Scheme for reducing bacterial adhesion at an implanted biomaterial.

surfaces because of the small penetration depth of electrons through solids. This attribute makes electron spectroscopy inherently surface-sensitive, since only a few of the outermost atomic layers are accessible. Electron spectroscopy requires propagation over distances of ∼10 cm without collisions with gas-phase molecules. The mean free path of an electron, λ_e, through a gas is given by Eq. (1), where k

$$\lambda_e = \frac{4kT}{\pi \xi^2 P} \tag{1}$$

is Boltzmann's constant, T is absolute temperature, ξ is molecular diameter of the ambient gas, and P is ambient pressure. If an electron propagates 10 cm (a typical path length between the sample and detector) at room temperature, a vacuum of 7×10^{-2} pascal is required. Electron spectroscopy is typically carried out at much lower pressures of ultrahigh vacuum ($<1 \times 10^{-7}$ Pa) to ensure sample integrity.

The methods of electron spectroscopy used in surface studies have several common characteristics (**Table 2**). A source provides the incident radiation to the sample, which can be in the form of electrons, x-radiation, or ultraviolet radiation. Electron beams are generated from the thermionic emission of metal filaments or metal oxide pellets. The incident radiation induces an excitation at the surface of the sample, which alters the energy distribution of electrons that leave the surface. This distribution provides a diagnostic of the composition or structure of the interface. Typical electron energy analyzers include retarding-field, cylindrical-mirror, and hemispherical-mirror analyzers. *See* ELECTRON SPECTROSCOPY.

Auger electron spectroscopy (AES). This technique is used to determine the elemental composition of a surface species. As shown in **Fig. 3***a*, the incident electron beam (2–3 keV) ionizes a K-shell electron of an atom of energy, E_K, creating an electron vacancy. An incident x-ray can also initiate the same event. An electron in a higher-energy L-shell with energy E_L subsequently drops down to fill the vacancy, releasing the energy $E = E_K - E_L$. The excess energy can either be emitted as an x-ray or be released to an outer-shell electron. Known as an Auger electron, it is ejected with kinetic energy $E \sim (E_K - E_L - E_M)$. This is a KLM transition, based on the energy levels involved in this particular process. Other transitions involving different electrons also occur (for example, KLL and LMM). The kinetic energy of the Auger electrons provide a so-called fingerprint for elemental identification, as found in numerous reference compilations. *See* AUGER EFFECT; ELECTRON-HOLE RECOMBINATION.

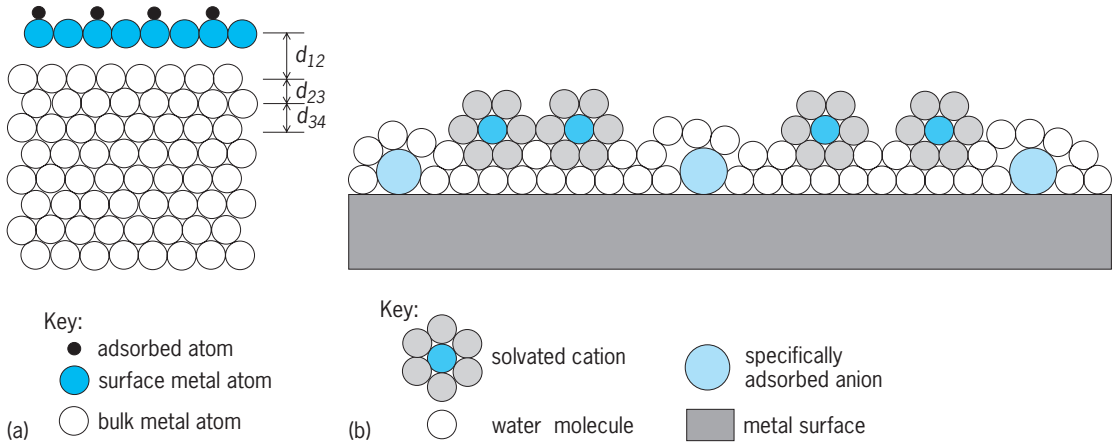


Fig. 2. Microscopic surface structures. (*a*) Surface atoms occupying positions other than the bulk-terminated locations. The first interlayer spacing d_{12} is greater than the bulk interlayer spacing d_{34}. The second interlayer spacing d_{23} is smaller than the bulk interlayer spacing. (*b*) Electrical double layer. The plane of closest approach is the inner Helmholtz plane. The adjacent layer of solvated cations is the outer Helmholtz plane.

TABLE 2. Surface-sensitive experimental techniques

Technique	Source*	Detectors	Level of Information
Auger electron spectroscopy (AES)	Electrons 2–3 keV	Cylindrical mirror or retarding field	Elemental composition
X-ray photoelectron spectroscopy (XPS)	X-rays 1254 eV (Mg) 1487 eV (Al)	Hemispherical or cylindrical mirror	Elemental composition and oxidation state
Ultraviolet photoelectron spectroscopy (UPS)	UV radiation 21 eV He(I) 41 eV He(II)	Hemispherical or cylindrical mirror	Electronic properties of adsorbate and/or bulk material
Energy loss spectroscopy (ELS)	Electrons 50–1000 eV	Electron energy analyzer	Electronic structure of surface
High-resolution electron energy loss spectroscopy (HREELS)	Electrons 1–10 eV	Electron energy analyzer	Vibrational losses
Low-energy electron diffraction (LEED)	Electrons 20–500 eV	Retarding fields and phosphorescent screen	Surface structure or periodicity
Infrared spectroscopy (IRS)	Photons	Mercury-cadmium-telluride or indium antimony	Molecular identity
Optical ellipsometry	Photons	Photomultiplier	Adsorbate layer thickness
Scanning tunneling microscopy (STM)	Tunneling current	Ammeter	Substrate roughness and texture

*Mg = magnesium; Al = aluminum; He = helium.

X-ray photoelectron spectroscopy (XPS). This technique is used to determine the oxidation state as well as the elemental composition of a surface species. In x-ray photoelectron spectroscopy, an inner-shell electron or photoelectron is ejected as a result of excitation by an incident x-ray photon (Fig. 3*b*). The x-ray source is usually a magnesium or aluminum anode. The ejected electrons have an energy $E = h\nu - E_i$, where $h\nu$ and E_i are the energies of the x-ray and the inner electron, respectively. The resulting vacancy can be filled by the Auger relaxation process. The energy of the photoelectron reflects both the oxidation state and elemental composition of the surface species. *See* X-RAY SPECTROMETRY.

Ultraviolet photoelectron spectroscopy (UPS). This is similar to x-ray photoelectron spectroscopy except that ultraviolet photons are used for excitation. The helium I and II lines are the most commonly used sources. Since ultraviolet radiation is less energetic than x-radiation, ultraviolet photoelectron spectroscopy probes only the valence electron levels of the surface species. The electron is ejected from the other outer shell with energy $E = h\nu - E_\upsilon$, where E_υ is the energy of the valence electron (Fig. 3*c*).

Electron energy loss spectroscopy (EELS). This technique probes the electronic and vibrational states of the surface species. It is divided into two categories, based on the energy E_p of the incident electron. Electronic levels are accessed with E_p = 50–1000 eV, whereas vibrational levels are probed with E_p = 1–10 eV (1 meV = 8.066 cm^{-1}). The technique that probes electronic levels is known as energy loss spectroscopy (ELS); the technique that probes vibrational levels is known as high-resolution electron energy loss spectroscopy (HREELS). Higher resolution is required to resolve the vibrational bands because the energy transitions are smaller, as shown in Fig. 3*d*. Most of the electrons are scattered elastically, and

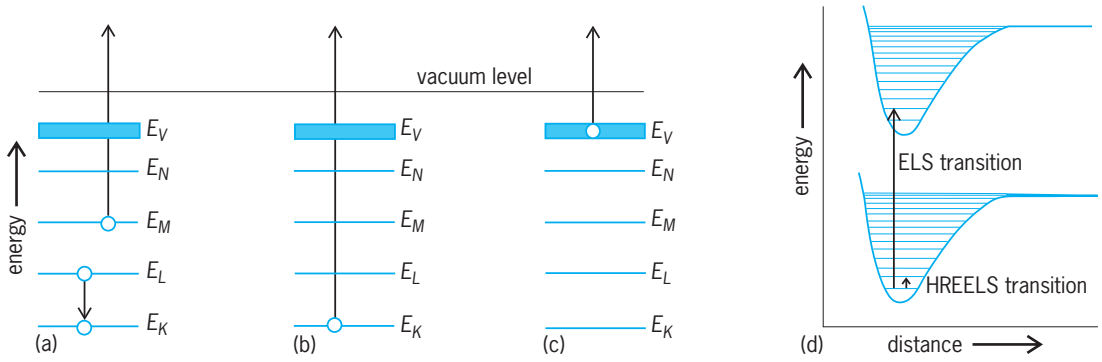


Fig. 3. Processes of electron spectroscopies. (*a*) Auger electron spectroscopy. A core-level electron is ejected by excitation from the incident high-energy electron. A second electron fills the vacancy left by the ejected electron. The third electron, the Auger electron, is ejected from the atom with the energy $E = E_K - E_L - E_M$. (*b*) X-ray photoelectron spectroscopy. A core-level electron is ejected from the atom as a result of excitation by the incident x-ray beam. Note that the Auger process, described in *a*, can also occur as a result of x-ray stimulation of the sample. (*c*) Ultraviolet photoelectron spectroscopy. Ultraviolet radiation excites the sample, causing ejection of weakly bound valence electrons. (*d*) Electron energy loss spectroscopy. An electron is excited into a higher electronic (ELS) or vibrational (HREELS) state by the incident electron beam. The scattered electrons lose the energy that was required for excitation. E_v = energy of the valence electrons.

they fail to induce an electronic or vibrational excitation of the surface species. However, in a few cases the incident electron undergoes an inelastic collision, losing energy via excitation of a vibration or electronic transition. The energy of the reflected beam is $E = E_p - E_{loss}$ (Fig. 3*d*). For high-resolution electron energy loss spectroscopy, only those vibrations that have a component of their transition dipole moment perpendicular to the surface can be excited. With this surface selection rule, the loss spectrum often provides information pertaining to the orientation of the surface species with respect to the surface normal.

Low-energy electron diffraction (LEED). This technique is the surface analog of x-ray diffraction; it probes the two-dimensional periodicity of the surface species. Applicable primarily to studies of single-crystal surfaces, low-energy electron diffraction arises from the elastic scattering of electrons incident on the surface. The electrons scatter in phase or out of phase, depending on the incident energy (wavelength) and the position of the atoms relative to one another. In-phase scattering, which interferes constructively to yield the low-energy electron diffraction pattern, occurs when the Bragg condition is satisfied. The Bragg condition is given by Eq. (2), where a is the distance

$$a(\sin\phi - \sin\phi_0) = n\lambda \tag{2}$$

between surface atoms, ϕ and ϕ_0 are the backscattered and incident angles of the electron beam, λ is the wavelength of the electron beam, and n is an integer corresponding to the diffraction order. The diffracted electrons are accelerated and focused by a series of grids onto a phosphorescent screen for visualization of the low-energy electron diffraction pattern. *See* ELECTRON DIFFRACTION.

Optical surface analysis techniques. Optical spectroscopy techniques (visible and infrared) are also useful for probing the chemical composition and molecular arrangement of surface species. Typical application configurations are the transmission and reflection (both external and internal) modes as shown in **Fig. 4**. Transmission spectroscopy relies on the passage of the probe beam through the sample. External and internal reflection spectroscopies involve the reflection of the probe beam from a medium with a lower refractive index to a medium with a higher refractive index, and from a higher to lower refractive index, respectively. The sample support must be optically transparent to the probe beam for the internal reflection mode. In both cases, the substrates are polished to a smooth, mirrorlike finish. *See* SPECTROSCOPY.

Diffuse reflection spectroscopy, which is essentially reflection from a rough surface, has been utilized to examine supported-metal catalysts and high-surface-area oxides. In all cases, the power lost by the interaction of the incident radiation with the surface species gives rise to an absorption spectrum. Comparison with the spectrum of the adsorbate starting material provides insights into the nature of the surface species. Propagation distances are much longer than those for electron spectroscopy, allowing the application of these techniques to in-situ surface studies such as at the gas–solid interface of a high-pressure chemical reactor as well as the liquid–solid interface of an electrochemical cell. Raman spectroscopy also provides information about the molecular composition of the surface species, although at a lower sensitivity. *See* RAMAN EFFECT.

Optical probes are less surface-sensitive than electron spectroscopy because of the longer wavelength of the incident energy. This can result in the sampling of both bulk and surface species, complicating the interpretation of the observed spectrum. Penetration depths for optical spectroscopies are typically a fraction of the wavelength of the incident light, and they are a function of the optical properties of the sample and of the support and the angle of incidence. For visible radiation, penetration depths in a reflection experiment are tens to hundreds of nanometers, whereas those for infrared radiation are a few hundred to thousands of nanometers. For internal reflection spectroscopy, the penetration depth d_p is given by Eq. (3), where λ_1 is the wavelength of the incident radiation, θ is the angle of incidence, and n_{21} is the ratio of the refractive indices of media 2 and 1, respectively.

$$d_p = \frac{\lambda_1}{2\pi(\sin^2\theta - n_{21})^{1/2}} \tag{3}$$

Infrared spectroscopy (IRS). This is a useful technique for the determination of the molecular composition and orientation of the surface species. In an external reflection mode, a beam of infrared light is reflected at a smooth metal, semiconductor, or insulator surface, as shown in Fig. 4*b*. For films with thicknesses

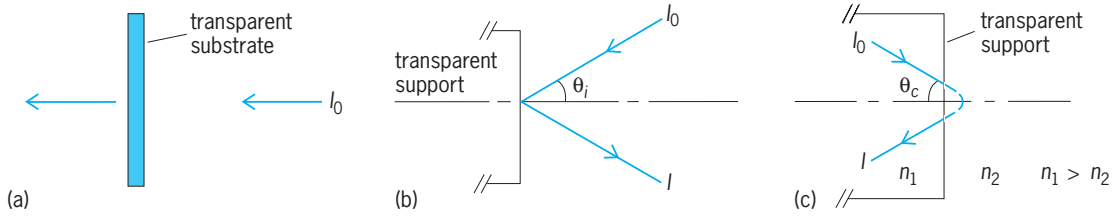


Fig. 4. Typical configuration for application of optical spectroscopies to surface analysis. (*a*) Transmission. (*b*) External or specular reflection. (*c*) Internal reflection. I = reflected light intensity. I_0 = initial light intensity. θ_i = angle of incidence at phase 1–phase 2 boundary. θ_c = angle of refraction at phase 1–phase 2 boundary. n_1 = refractive index of phase 1. n_2 = refractive index of phase 2.

greater than ~10 nm, conventional dispersive spectrometers can be used. However, studies of monolayers require a Fourier transform spectrometer and high-sensitivity, low-noise, photoconductive detectors, for example, mercury-cadmium-telluride.

The band shapes and intensities of reflection spectra can be calculated quantitatively with classical electromagnetic theory. Variables in the calculation, which is a reflection spectroscopy analog to the Beer-Lambert law in transmission spectroscopy, include the optical function of each phase and the polarization and angle of incidence of the incident beam. Comparison of the intensities of the observed and calculated spectra, assuming that the force constants of a given vibrational mode are unchanged by immobilization, provides an estimate of the average spatial orientation of the adsorbate with respect to the substrate. As with high-resolution electron energy loss spectroscopy, such interpretations at highly reflective surfaces such as metals are aided by the preferential excitation of those vibrational modes with a transition dipole moment perpendicular to the surface, according to the infrared surface selection rule. An orientational analysis at a surface with a low reflectivity can also be accomplished, but it is hindered by a complex variation of the relative degree of excitation of vibrations with modes both perpendicular and parallel to the surface. *See* INFRARED SPECTROSCOPY.

Optical ellipsometry. This is a variation of the external reflection measurement in that the state of polarization of the reflected light is measured as opposed to the reflected power. The change in polarization is caused by the difference in the intensity and phase change for the components of the incident electromagnetic wave that are parallel and perpendicular to the surface. Such changes, which are a function of the optical properties of the adsorbate, can be used to determine its thickness. More detailed treatments with effective medium theory can also provide insights into the density of the surface layer. Computerized data acquisition is advantageous, because numerical analysis requires the collection of large data sets and extensive complex algebraic manipulation. *See* ELLIPSOMETRY.

Scanning tunneling microscopy (STM). This is a surface analysis technique that provides information about the three-dimensional topography of a surface with resolution at the atomic level. Operation in ultrahigh vacuum, at liquid–solid interfaces, and in laboratory ambient (air) has been demonstrated.

The two general operation configurations are the constant-current and constant-height modes. In the constant-current mode (**Fig. 5***a*), a bias voltage (~10–10,000 mV) is applied between a sharp metallic tip and a conducting sample. A small tunneling current (~0.2–10 nanoamperes) flows when the tip is positioned within a few atomic diameters of the sample. The tip is then scanned laterally with a piezoelectric translator over a surface while a computer-controlled feedback mechanism adjusts the height of the tip to maintain a constant current. An image of the surface contour is constructed by plotting a weighted average of the tip height z versus the lateral positions x and y.

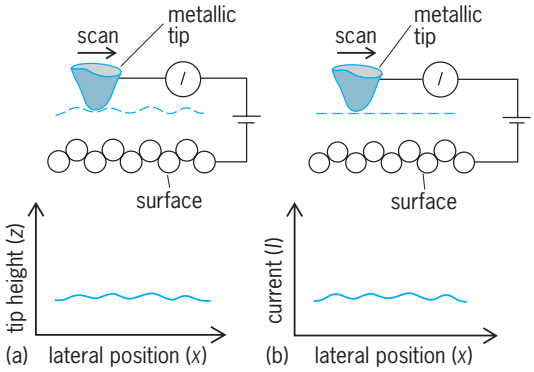


Fig. 5. Scanning tunneling microscopy and images for (*a*) constant-current mode and (*b*) constant-height mode; shown only in *x* dimension for simplicity.

In the constant-height mode, a tip is scanned across a surface at constant height while monitoring the current I (Fig. 5*b*). A surface image is derived from a plot of multiple scans displaced from each other.

Each mode of operation has its own specific advantages. The constant-current mode provides topographic pictures of surfaces that are not atomically flat. This mode operates at scan rates of only a few hertz and is limited by the vertical movement of the tip by the z translator. In contrast, the constant-height mode can image atomically smooth surfaces at rates up to ~1 kHz. These high imaging rates are attainable because only the electronics for monitoring the tunneling current—and not the vertical movement of the z translator—must respond to the atoms passing under the tip. A high imaging rate allows the examination of real-time processes, reduces data acquisition time, and minimizes image distortion caused by thermal drift and piezoelectric creep. *See* INTERFACE OF PHASES; SCANNING TUNNELING MICROSCOPE.

Mary M. Walczak; Marc D. Porter

Bibliography. C. Bai, *Scanning Tunneling Microscopy and Its Application*, 2d ed., 2000; H. Bubert and H. Jenett (eds.), *Surface and Thin Film Analysis: A Compendium of Principles, Instrumentation, and Applications*, 2002; H. P. Hughes and H. Starnberg (eds.), *Electron Spectroscopies Applied to Low-Dimensional Materials*, 2000; A. J. Milling, *Surface Characterization Methods: Principles, Techniques, and Applications*, 1999; G. A. Somorjai, *Introduction to Surface Chemistry and Catalysis*, 1994.

Surface condenser

A heat-transfer device used to condense a vapor, usually steam, by absorbing its latent heat in a cooling fluid, ordinarily water. Most surface condensers consist of a chamber containing a large number of 0.5–1-in.-diameter (1.25–2.5-cm) corrosion-resisting alloy tubes through which cooling water flows. The vapor contacts the outside surface of the tubes and is

condensed on them. The tubes are arranged so that the cooling water passes through the vapor space one or more times. About 90% of the surface is used for condensing vapor and the remaining 10% for cooling noncondensable gases. Air coolers are normally an integral part of the condenser but may be separate and external to it. The condensate is removed by a condensate pump and the noncondensables by a vacuum pump. *See* HEAT TRANSFER; STEAM CONDENSER; VAPOR CONDENSER. Joseph F. Sebald

Surface hardening of steel

The selective hardening of the surface layer of a steel product by one of several processes which involve changes in microstructure with or without changes in composition. Surface hardening imparts a combination of properties to the finished product not produced by bulk heat treatment alone. Among these properties are high wear resistance and good toughness or impact properties, increased resistance to failure by fatigue resulting from cyclic loading, and resistance to surface indentation by localized loads. The use of surface hardening frequently is also favored by lower costs and greater flexibility in manufacturing.

Hardening processes. The principal surface hardening processes are (1) carburizing; (2) the modified carburizing processes of carbonitriding, cyaniding, and liquid carburizing; (3) nitriding; (4) flame hardening, induction hardening, and laser hardening; and (5) surface working.

Carburizing introduces carbon into the surface layer of low-carbon steel parts and converts that layer into high-carbon steel, which can be quench-hardened by appropriate heat treatment. Carbonitriding, cyaniding, and liquid carburizing, in addition to supplying carbon, introduce nitrogen into the surface layer; this element permits lower case-hardening temperatures and has a beneficial effect on the subsequent heat treatment. In nitriding, only nitrogen is supplied, and reacts with special alloy elements present in the steel.

Whereas the foregoing processes change the composition of the surface layer, flame hardening, induction hardening, and laser hardening depend on a heat treatment applied selectively to the surface layer of a medium-carbon steel. These processes are based on the same principle, but differ in the source of heat; also, in flame hardening and induction hardening, the heated layer is rapidly cooled (quenched) by an external cooling medium, whereas the layer heated by a laser beam is quenched (rapidly cooled) by conduction to the underlying cold metal. Surface working by shot peening, surface rolling, or prestressing improves fatigue resistance by producing a stronger case, compressive stresses, and a smoother surface.

Table 1 lists the surface-hardening processes for steel and their major characteristics. Processes related to the surface hardening of steel are hard-facing, metal spraying, electroplating, and various diffusion processes involving elements such as aluminum, silicon, and chromium. *See* CLADDING; ELECTROPLATING OF METALS; METAL COATINGS.

Carburizing. The oldest method of surface hardening steel, carburizing, introduces carbon into the surface layer of a low-carbon steel by heating above the transformation range in contact with a carbonaceous material. The carbon diffuses into the steel from the surface and thus converts the outer layer into high-carbon steel. The composite is then heat-treated by the procedures generally applicable to steels. In particular, it must be cooled from above the transformation temperature at a rate sufficiently fast to transform the high-carbon surface layer into a hard martensitic case while the low-carbon core remains tough and shock-resistant. The quench is usually followed by a low-temperature stress-relief anneal. *See* HEAT TREATMENT (METALLURGY).

Pack carburizing. In pack carburizing, carbon is supplied by charcoal or coke to which carbonates or organic materials are added; the mixture is known as the carburizing compound. Parts to be hardened are packed in a steel box with the carburizing compound and heated to the carburizing temperature, usually 1700–1750°F (925–955°C). Carbon is transferred to the steel by the formation of carbon monoxide at the compound surface and by its decomposition to carbon and carbon dioxide at the steel surface according to the reaction given below. The carbonates

$$C + CO_2 \rightleftharpoons 2CO$$

or organic materials in the carburizing compound decompose and increase the concentration of carbon oxide gases in the box required for the transfer of carbon to the steel. The depth of penetration of the carbon depends upon the time and temperature at which the treatment is carried out.

TABLE 1. Surface-hardening processes for steel

Process	Elements added	Hardening mechanism
Carburizing	Carbon	Formation of martensite
Carbonitriding	Carbon and nitrogen	Formation of martensite
Cyaniding	Carbon and nitrogen	Formation of martensite
Liquid carburizing	Carbon and nitrogen	Formation of martensite
Nitriding	Nitrogen	Precipitation of alloy nitrides
Flame hardening	None	Formation of martensite
Induction hardening	None	Formation of martensite
Laser hardening	None	Formation of martensite
Surface working	None	Work hardening

The principal advantage of pack carburizing is its simplicity; no expensive equipment is required. The results are almost certain to be satisfactory with proper temperature control and hardening practice. The hardening treatment usually requires that the parts, after removal from the box, be heated again above the transformation temperature, and that this process be followed by quenching.

Gas carburizing. In gas carburizing the parts are heated in contact with carbon-bearing gases, commonly carbon monoxide and hydrocarbons. The hydrocarbons may be methane, propane, butane, or vaporized hydrocarbon fluids. They are usually diluted with an inert carrier gas to control the amount of carbon supplied to the steel surface and to prevent the formation of soot.

The carbon monoxide and hydrocarbons are decomposed at the steel surfaces, the carbon thus liberated being absorbed by the steel. Close control of gas composition is required because the rate at which carbon is supplied to the steel surface controls the concentration of carbon in the carburized case. This control is an important advantage of gas carburizing. The process is cleaner and entails lower labor costs than pack carburizing. Direct quenching from the furnace is possible in gas carburizing; hence the process is particularly well suited to volume output for the mass production industries. In many installations large continuous furnaces with attached quenching and tempering equipment are used; the parts are charged into one end and leave at the other in the carburized and hardened state. Batch furnaces are also used for gas carburizing.

Steels for carburizing. The selection of steels for carburizing primarily concerns grain growth characteristics, carbon and alloy content, machinability, and cost. Steels that retain a fine-grained structure at the case-hardening temperature such as aluminum deoxidized steels are desirable because they permit simple heat-treatment procedures, in particular, hardening by a direct quench. A fine-grained structure in the finished product is essential for maximum shock resistance.

Plain carbon steels are satisfactory for many applications in which low distortion is not a critical requirement and for which optimum core properties are not required. The most common carbon contents are 0.20% in plain carbon steels and 0.08–0.20% in alloy steels. Steels with relatively high sulfur content are frequently used for improved machinability.

The most common alloy elements in carburizing steels are nickel (0.5–3.5%), chromium (0.5–1.6%), and molybdenum (0.1–0.25%). The nickel-molybdenum steels are particularly popular for strength and toughness of the core and toughness of the case. Alloy steels have less tendency to develop coarse-grained structures at the carburizing temperature. They also permit slower quenching rates in hardening, thus reducing distortion and the tendency to crack during quenching. Their higher cost is frequently more than offset by lower finishing costs due to reduced distortion. Improved fatigue resistance is another consideration in the selection of alloy steels.

Typical parts surface-hardened by carburization, as described above, or by modified carburizing processes, described below, include gears, ball and roller bearings, piston pins, sprockets and shafts, clutch plates, and cams.

Modified carburizing processes. Alloying the steel with nitrogen lowers the transformation temperature and reduces the transformation rate. Thus modification of the carburizing process by the diffusing of both carbon and nitrogen into the surface layer of the steel enables the process to be carried out at lower temperatures than with carbon alone. The sources of carbon and nitrogen distinguish the several processes from each other.

Carbonitriding. The carbonitriding process is the same as gas carburizing except that ammonia is added to the furnace atmosphere to provide the source of nitrogen. The amount of nitrogen absorbed by the steel can be controlled by the concentration of ammonia in the furnace atmosphere and the temperature. Thus the nitrogen content of carbonitrided cases may vary from small concentrations to the relatively high level characteristic of cyaniding.

Both because carbonitriding is conducted at lower temperatures than carburizing and because the slower transformation rate permits slower quenching rates, less distortion results from the hardening process. Consequently, plain carbon steel can frequently be substituted for alloy steels in carbonitrided parts.

Figure 1 shows a typical carbonitrided case. The microstructure is similar to a carburized case except that it contains a larger fraction of retained austenite near the surface.

Cyaniding and liquid carburizing. If the parts are immersed into molten baths consisting of solutions of cyanides and other salts, the cyanides supply both carbon and nitrogen to the steel. Thus the results from cyaniding and liquid carburizing are similar to

Fig. 1. Microstructure of typical carbonitrided steel. Dark border is specimen mount. Martensite needles (gray) and retained austenite (whitish) can be seen near surface. Microhardness indentations (black) are smaller in the case than in the core, indicating the greater hardness of the case. (*Courtesy of A. J. Gregor*)

those from carbonitriding. By controlling the composition and temperature of the bath, the amounts of carbon and nitrogen absorbed by the steel can be controlled within limits. The term cyaniding is usually applied to processing at temperatures of 1400–1550°F (760–845°C). Liquid carburizing is carried out at 1600–1750°F (870–955°C); at these temperatures the nitrogen absorption is lower so that the process, as far as the product is concerned, approaches pack and gas carburizing.

Nitrogen in cyanided cases results in the same advantages as those obtained from nitrogen in carbonitrided cases, but the control of nitrogen concentration is less precise. Cyaniding and liquid carburizing have the advantage of rapid heating of the charge because of the good heat transfer from the bath to the steel. These processes have considerable flexibility in that selective hardening may be accomplished by partial immersion and that different parts may be treated in the same bath for different times.

One disadvantage of these processes is that the case for a given treatment time is less deep when low operating temperatures are used. The size of salt baths is limited by the necessity to obtain uniform temperatures throughout. The baths are, therefore, less well adapted to quantity production than furnaces employing gaseous atmospheres. The nitrogen dissolved in the steel may cause an appreciable fraction of the steel to remain untransformed as retained austenite in the final product, particularly in the cyaniding process, but also in carbonitriding, especially at low temperatures.

Nitriding. This process is carried out by heating steels of suitable composition in contact with a source of active nitrogen at temperatures of 925–1100°F (495–595°C) for periods of 1–100 h, depending upon the steel being treated and the depth of case desired. Under these conditions, nitrides form if the steel contains alloying elements such as aluminum, chromium, molybdenum, vanadium, and tungsten. The formation of alloy nitrides at the nitriding temperature accounts for the hardened case. The microstructure of a typical nitrided steel is shown in **Fig. 2**.

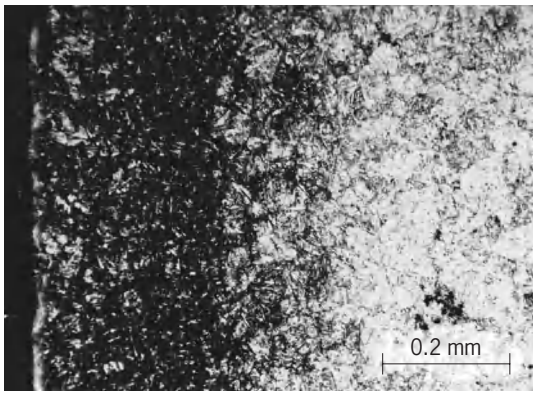


Fig. 2. Microstructure of a nitrided steel, which was heated to 1700°F (925°C) for 1 h, oil-quenched, tempered at 1250°F (675°C) for 2 h, air-cooled, and nitrided for 48 h at 975°F (525°C). Dark border is specimen mount.

The usual source of active nitrogen for nitriding is ammonia. However, mixtures of molten cyanide salts are also used, with similar advantages and disadvantages as in cyaniding and liquid carburizing. Salt mixtures with cyanates (which are less toxic than cyanides) are coming into use. A typical nitriding installation using ammonia consists of a reservoir of ammonia, a furnace, a retort containing the parts to be case hardened, and equipment to control the temperature and gas flow. At the nitriding temperature, part of the ammonia decomposes at the surface of the steel, liberating active nitrogen, some of which diffuses into the steel. The remainder passes into the molecular form, which is inert. For successful nitriding it is therefore necessary to control the gas flow in such a way as to continuously supply fresh ammonia to all steel surfaces. However, an oversupply of ammonia results in the formation of an excessively thick iron nitride layer on the surface. This so-called white layer can be controlled by regulation of the degree of ammonia dissociation, or it must be removed by grinding if it exceeds a depth of 0.0005 in. (12.7 micrometers).

In ion nitriding, the workpieces are cathodically charged to a potential of 400 to 1000 V in an evacuated furnace chamber. When a nitrogen-containing gas at low pressure is introduced into the chamber, it becomes ionized (a plasma). The positive nitrogen ions bombard the surface of the workpiece under the influence of the potential drop and cause the nitriding reaction which introduces nitrogen into the surface layer of the steel. The bombardment by the plasma discharge also generates the heat necessary to raise the temperature to the level required for the reaction to proceed. The basic features of the nitriding reaction are essentially the same as in conventional gas or liquid nitriding: nitrogen is absorbed at the surface, it diffuses into the workpiece and, given sufficient nitriding potential and time, a compound layer is formed at the surface. Ion nitriding is best suited for the production of shallow cases. It requires special equipment and appropriate expertise. Among the advantages claimed for ion nitriding, the use of which is still growing, are its flexibility, ability to be controlled, and low energy and gas consumption.

Steels that produce the hardest nitrided cases consistent with optimum case depth contain about 1% aluminum, 1.5% chromium, and 0.3% molybdenum. Alloy steels containing only chromium and molybdenum as nitride-forming elements are also popular. Stainless steels and high-speed cutting steels are nitrided for improved wear resistance. Among the parts hardened by nitriding are camshafts, fuel injection pump parts, gears, cylinder barrels, boring bars, spindles, splicers, sprockets, valve stems, and milling cutters.

Induction, flame, and laser hardening. The principle of both flame and induction hardening is to heat quickly the surface of the steel to above the transformation range followed by quenching rapidly. In induction hardening, heat is generated within the part by electromagnetic induction. The part (such

as a crankshaft, camshaft, axle, gear, or piston rod) is usually placed inside a copper coil or coils through which a rapidly alternating current is flowing. High-frequency currents are used because they confine the induced currents to the surface of the part to be heated; the higher the frequency, the shallower the case. Short heating cycles minimize conduction of heat to the interior and thus further restrict the heating to the surface layer. *See* INDUCTION HEATING.

In flame hardening, the steel to be surface-hardened is heated by direct impingement of a high-temperature gas flame. The surface layer is quickly heated to a temperature above the transformation range, followed by a quench. In both induction and flame hardening the quenching action is a combination of heat extraction by the cold metal beneath the case and by an external quenching medium.

Induction and flame hardening require close control of time and temperature of heating. In general, induction heating is used when large numbers of symmetrically shaped parts are to be processed. Flame hardening can be readily applied to large parts, such as large gears or lathe ways, and to parts of intricate design, such as camshafts. It is more economical than induction hardening when only a few parts are to be treated.

Steels for induction or flame hardening usually contain 0.4–0.75% carbon. Because no change in composition is involved, the steel is selected for both case and core properties. During hardening the core is not affected and, consequently, the core properties must be developed by proper heat treatment before surface hardening. Cast irons are also induction- or flame-hardened for certain applications. *See* CAST IRON.

In laser hardening, as in induction and flame hardening, a surface layer of the workpiece is heated above the transformation range without a change in composition, but laser hardening does not require an external quenching medium. The heat source is a laser which generates an intense energy flux and rapidly raises the temperature. The laser beam is manipulated by lenses and mirrors, and travels at a controlled speed across the surface area to be hardened. In order to reduce the energy loss by reflection, absorbing coating materials are applied or the surface is roughened. Since the heated surface layer is shallow, it loses heat rapidly to the interior (self-quenching) and hardens by the formation of martensite. The case depth depends mainly on the power density and traveling speed of the beam and usually is less than for induction and flame hardening. Advantages of laser hardening are its ability to harden selectively small areas as well as complex shapes and small bores. Drawbacks are the limited case depth and the capital costs of the equipment. *See* LASER.

Surface-working processes. The selective cold working of the surface layer of parts of steel and other metals increases surface hardness. The working requires a force exceeding the compressive yield strength of the material. Such a force can be applied by various methods, for example, hammer peening, mechanical peening, shot peening, and surface rolling. The surfaces must be accessible to the peening or rolling operation, but the processes can be applied to selected critical areas, such as the fillets of shafts. Prestressing involves stressing a part beyond its yield strength.

Surface-working processes result in substantial improvement of the properties, especially the fatigue resistance. The processes are applied to coil and leaf springs, shafts, gears, and steering knuckles. They have the advantage of being comparatively inexpensive. They must be carefully controlled to avoid overworking or underworking of the surface, both of which fail to give the desired improvement in the fatigue properties.

Properties. The characteristic properties of surface-hardened steels depend on the properties of both the case and the core. Case properties are determined mainly by composition, microstructure, and case depth. Core properties of carburized steels depend primarily on the transformation characteristics of the core during the hardening of the case. In nitriding and also in induction and flame hardening, the core properties are developed before the case hardening treatment.

Microstructure. The microstructures of the case and core are controlled by their composition and heat treatment. The structure of carburized, carbonitrided, and cyanided cases is typical of heat-treated high-carbon steel, and that of induction- and

TABLE 2. Typical case properties

Process	Typical case depth, in.*		Typical hardness
Carburizing	< 0.020	Shallow	55–65 RC
	0.020–0.040	Medium	
	0.040–0.060	Heavy	
	> 0.060	Extra deep	
Carbonitriding	0.003–0.020		55–62 RC
Cyaniding	0.001–0.010		
Nitriding	0.005–0.025		85–95 R-15N
	0.001–0.003	High-speed steel	
Induction hardening	0.010–0.25		50–60 RC
Flame hardening	0.030–0.25		50–60 RC
Laser hardening	0.008–0.050		
Surface working	0.020–0.040		

*1 in. = 25 mm.

flame-hardened cases is typical of medium-carbon steel. The desired microconstituent in each instance is martensite, which usually is tempered. The retained austenite, which is promoted by nitrogen in carbonitrided and cyanided cases, and by some other alloy elements, is generally undesirable. Austenite lowers the hardness and its subsequent transformation in service may cause brittleness and dimensional changes of the case-hardened part. *See* TEMPERING.

The microstructure of nitrided cases consists of finely dispersed alloy nitrides precipitated in a preexisting hardened and tempered structure. In induction- and flame-hardened cases, the fast rate of cooling from the austenitizing temperature results in the formation of martensite. The characteristic feature of surface-worked cases is their cold-worked structure and resulting increase in hardness. *See* PLASTIC DEFORMATION OF METAL.

Case depth. Typical case depths of surface-hardened steels are shown in **Table 2**. There has been a trend toward thinner cases, which are acceptable because of stronger cores and less distortion during processing. However, for some types of service, especially those involving contact of two loaded parts, the critical stress occurs below the surface; the case depth should preferably be sufficient to allow for this condition. Thick cases are essential in some wear applications.

Residual stresses. Compressive stresses at the surface are desirable because they provide protection against fatigue failure. The expansion accompanying the formation of martensite in carburizing and, similarly, that resulting from the precipitation reactions in nitriding cause these residual compressive stresses. Shot peening or other surface-working processes also cause residual compressive stresses at the surface.

Dimensional changes. In heat-treating carburized steels, distortion occurs primarily as a result of uneven quenching, which causes different parts of the surface to transform at different times. In nitriding, dimensional changes result primarily from the increase in volume of the case as the alloy nitrides are precipitated.

Michael B. Bever; Carl F. Floe

Bibliography. American Society for Metals, *Metals Handbook*, desk edition, 1985; C. R. Brooks, *Heat Treatment of Ferrous Alloys*, 1979.

Surface mining

A mining method used to obtain valuable minerals from the earth by first removing the overlying soil and rock (overburden) and subsequently recovering the valuable mineral. Surface mining is done until the accumulated overburden removed becomes too thick (and expensive to remove) for the economic recovery of the mineral. For mineral deposits too deep to be surface-mined, underground mining methods are used. *See* MINING.

History. Large-scale surface mining in the United States started in the early 1900s, as the use of large mechanical excavators made the recovery of lower-grade ores economical. The Bingham Canyon copper mine in Utah, which opened in the late 1800s, was converted to an open-pit mine around 1905. Full-revolving steam shovels, purchased from the Panama Canal project, greatly improved the mining efficiency at this mine and allowed the open-pit concept to be implemented. Technology advancements in drilling and blasting methods also contributed to successful open-pit mining.

Surface coal mining became a significant contributor to coal production around 1910. A surface coal mine on Grapevine Creek near Danville, Illinois, may have been the first large-scale strip mine in the United States. Track-mounted steam shovels from the Panama Canal project were also used at this mine. It was not until the 1930s that steam shovels were replaced by electric shovels. Around the same time, trucking became the dominant transportation method, replacing steam-powered track haulage.

Mining methods. Surface mining methods include placer mining, open glory-hole mining or milling, open-pit mining, strip mining, and quarrying.

Placer mining. These methods are used to recover unconsolidated mineral-rich (placer) deposits, often in streams, rivers, or their associated flood plain. Hydraulic methods, such as high-pressure water, can be used to loosen dry, unconsolidated material. The water–ore mixture then flows by gravity to an associated facility for processing into a salable product. Due to environmental concerns, this mining method is seldom used today. Another method of mining placer deposits is to use a dredge for excavating unconsolidated material underwater. Sand and gravel mining often use dredges. *See* PLACER MINING.

Open glory holes or milling pits. These are combined surface and underground mining operations, where a large opening or shaft is excavated vertically from the surface through a mineral-rich deposit to an underground mine. Drilling and blasting of the ore is directed downward from the surface around the shaft. The broken ore falls through the shaft and is recovered in the underground mine. These methods are typically used for precious-metal ore deposits such as gold, silver, or copper.

Open-pit mining. This method is used to recover a large ore deposit by removing the overlying earth and rock in conjunction with removing the ore. An open-pit mine is designed in a circular or oval shape, creating a deep pit. Steep, unstable slopes are avoided since they could trigger a landslide and endanger the miners. The pit is kept dry by draining water to a sump and pumping it to a treatment facility outside the pit. The overburden and processed waste rock are placed in a landfill area. The mining operation can continue until the cost of removing the overburden exceeds the payment received for the recovered mineral. Open-pit mining typically is used at very large ore bodies, with the economic life sometimes exceeding 50 years. Open-pit mining is also called open-cut mining since the overburden is not returned to the pit and an open cut remains after mining is completed. *See* OPEN-PIT MINING.

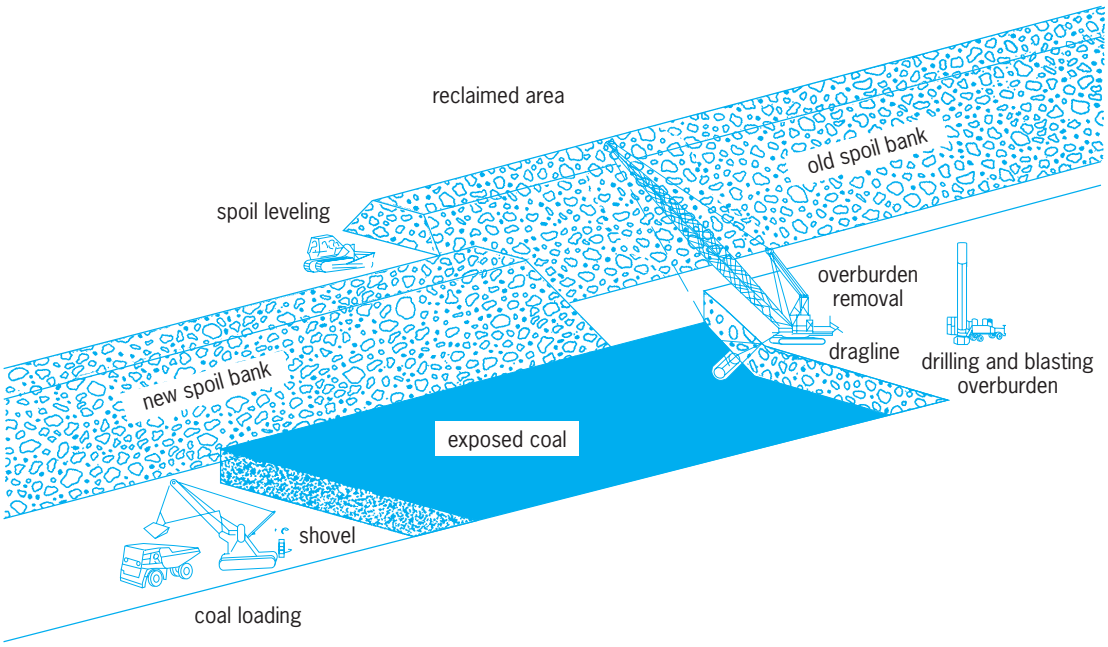


Fig. 1. Area mining.

Strip mining. Strip mining is the surface mining of coal. Before mining begins, ditches and ponds are constructed for containing sediment from erosion caused by rainstorms. The area is then cleared of trees, and the topsoil is removed and stockpiled. The three types of strip mining are area mining, contour mining, and mountaintop removal. *See* COAL; COAL MINING.

Area mining occurs on large sites that are flat or with rolling hills (**Fig. 1**). An initial opening is made by removing the overburden from a small area to uncover the coal. The opening is called a box cut since it is rectangular. It can also be called the first cut. Overburden from the first cut is stockpiled. After the coal is removed, overburden from a second cut is placed over the area of the first cut and graded to approximately the same contour as before mining. The area is covered with the stockpiled topsoil and reclaimed by planting grasses and trees. This operation continues until the limit of the mine is reached. The overburden from the first cut is then hauled to the last or final cut and used to restore that area. In this manner, the mined zone in a surface coal mine is returned to a condition similar to the one that existed before mining.

When a coal seam is located along a steep hillside and only one cut into the hill can be economically made, the contour mining method is used (**Fig. 2**). As the overburden cut advances, the excavated material is hauled past the exposed coal to the previously mined area and used for reclamation.

The mountaintop removal mining method is used to recover multiple coal seams that lie within the geologic strata on a mountain ridge (**Fig. 3**). The ridge's top is excavated, exposing the coal seams for recovery. Overburden is sequentially replaced to restore the ridge to approximately the same configuration as before mining. The overburden has a larger volume than the previously undisturbed strata because broken rock occupies more space than consolidated

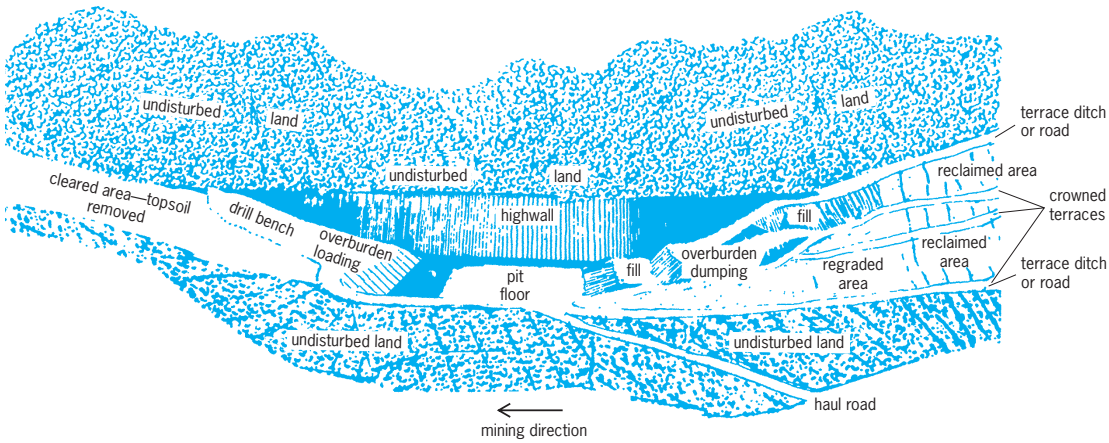


Fig. 2. Contour mining.

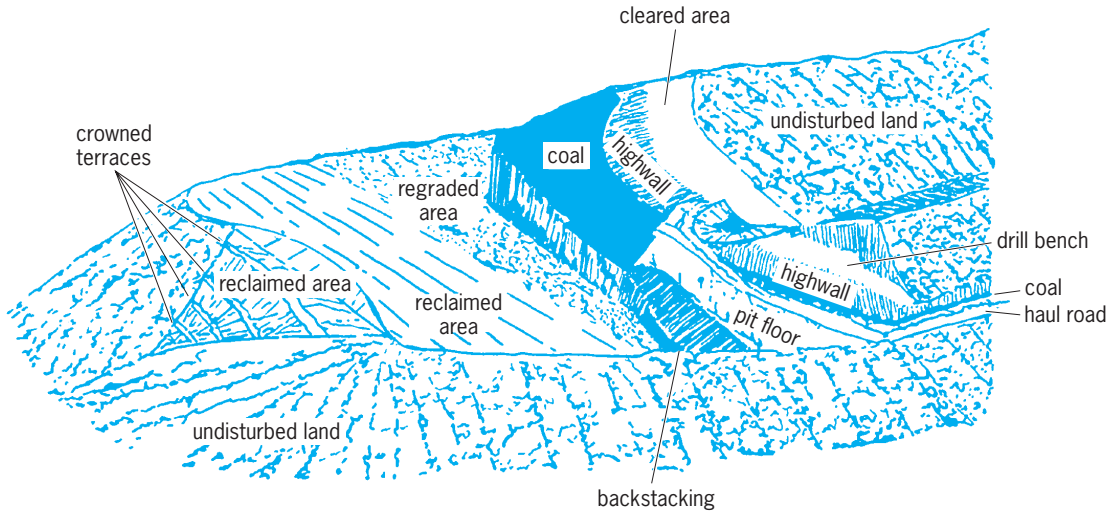


Fig. 3. Mountaintop mining.

rock, so excess overburden is placed in a landfill near the mine site.

Engineers plan the limits of strip mines by calculating economic strip ratios. The economic strip ratio is the thickness of the overburden divided by the thickness of the coal. Mining is economically feasible until the cost of excavation and reclamation becomes higher than the coal's worth. For example, if the price of coal restricts the mining of a coal seam that is 5 ft (1.5 m) thick to removing 100 ft (30 m) of overburden or less, the economic strip ratio is 20 to 1.

Quarrying. Surface mining of valuable stones, such as building stone and limestone, are called quarries (**Fig. 4**). A typical quarry operation removes only a small amount of overburden, while most of the rock mined is sold for crushed rock. The rock is typically broken by blasting, loaded into trucks, and hauled to a processing facility, often a crusher. Crushers are mechanical devices used to break rocks into smaller sizes. Large screens are used to separate the various rock sizes. Ornamental rock quarries, such as marble, may cut the rock to dimensions requested by a purchaser. Most of the rock mined is sold, with a large opening remaining after mining, as well as a near-vertical rock slope called a highwall. To restore a quarry, rock and soil are placed against the highwall to create a stable slope, which is subsequently covered by topsoil and revegetated. *See* QUARRYING.

Equipment. Productivity improvements in the surface mining industry have resulted principally from the increased size and efficiency of the equipment used. Steam-powered shovels were primarily responsible for creating the large-scale surface mining industry. They have been replaced by electric- or diesel-powered shovels. *See* POWER SHOVEL.

Surface mining sites use the same type of equipment as the construction industry. The major types of equipment include bulldozers, drilling and blasting equipment, wheel loaders, hydraulic excavators, draglines, off-road trucks, mechanical

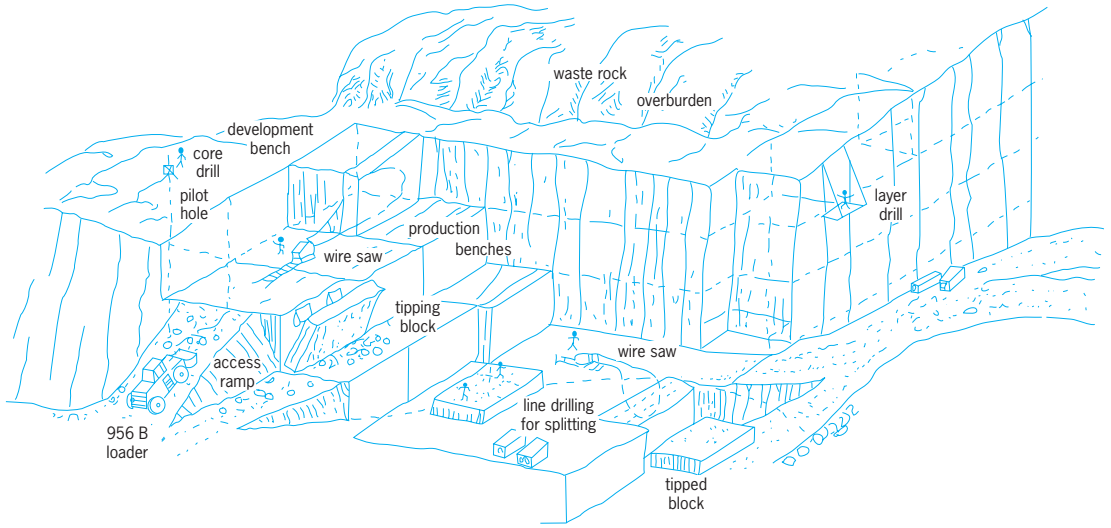


Fig. 4. Quarrying.

shovels, bucket wheel excavators, and scrapers. *See* CONSTRUCTION EQUIPMENT.

A bulldozer is a tractor with a large blade attached to its front, and is propelled using crawler tracks. Bulldozers can be used to relocate soil and rock over very short distances (less than 1000 ft or 300 m) and are typically used for leveling and grading soil during reclamation.

Large drilling machines are used to create vertical boreholes 6–12 in. (15–30 cm) in diameter in the overburden. Explosives inserted into the boreholes are detonated to break up the overburden so that it can be easily excavated.

Wheel loaders are commonly called front-end loaders. They are rubber-tired tractors with a large bucket in front. The bucket is filled by pushing it into overburden or ore, lifted and maneuvered by a hydraulic mechanism; and then its contents are dumped into a truck bed.

A hydraulic excavator is a backhoe with large bucket attached to a long boom. The bucket is filled with overburden or ore by being pulled back toward the operator by a hydraulic mechanism. These are quite useful when working on flat terrain and excavating material lower than the base of the excavator.

Draglines are machines that have a very large bucket suspended from a boom by cables (**Fig. 5**). The bucket is cast a considerable distance from the machine by swinging and dropping it in one motion. The bucket is filled with overburden by dragging it back toward the machine by retracting the cables. The filled bucket is then lifted and rotated to the disposal area and dumped by lifting the rear of the bucket, again by retracting the cables. Draglines have bucket capacity up to 220 yd^3 (168 m^3), and some weigh over 10,000 tons (9070 metric tons).

Haulage trucks used at mine sites have much higher capacity than trucks built for highway use, with some able to carry over 300 tons (272 metric tons).

Bucketwheel excavators can be used to excavate overburden that is unconsolidated and does not contain large rocks (**Fig. 6**). They continuously excavate using a rotating wheel of buckets (similar to a waterwheel) that scoops up and drops overburden onto a conveyor.

Scrapers are rubber-tired machines that can be used to remove and transport unconsolidated overburden. They are loaded by pushing forward a metal plate that scrapes and lifts dry, unconsolidated overburden into a hopper in the center of the machine. After hauling, the overburden is unloaded through an opening at the bottom of the hopper while moving forward. A scraper can be used, for example, to deposit overburden at an area being restored.

Industry statistics. According to the Mine Safety and Health Administration, there were about 11,000 surface mines in the United States in 2001. The sand and gravel industry had the most operations with 7126 mines, including 745 dredging operations. In the United States, all the sand and gravel is produced by surface mining. Stone quarries are the second most numerous type of surface mine. Coal-, nonmetal-, and metal-producing surface mines numbered less than 1000 mines each. In comparison, the total number of underground mining operations

Fig. 5. Large dragline operating from the top of the highwall.

in the United States during 2001 was 851, including 730 underground coal mines. *See* UNDERGROUND MINING.

In 2001, approximately 115,000 people were employed by surface mining operations in the United States (see **table**). The sand and gravel industry had the most employees. The largest surface mining operations, based on the average number of employees per mine, were metal-producing mines, principally the large western United States open-pit mines.

In the United States, the electric utility industry produces about 50% of its power from coal and consumes most of the coal produced there. Over 1 billion tons (900 million metric tons) of coal is produced annually in the United States, of which over 700 million tons (635 million metric tons) comes from surface mines. In 2001, the four states producing the most coal by surface mining were Wyoming, 350 million tons (318 million metric tons); West Virginia, 60 million tons (54 million metric tons); and Kentucky and Texas (primarily lignite) at 48 million tons (44 million metric tons) each.

Regulations. The surface mining industry is governed by safety and environmental laws, enacted by federal, state, and local governments. Agencies created by the governmental bodies typically issue regulations to enforce the laws. Federal agencies that regulate surface mining include the Mine Safety and Health Administration, Environmental Protection Agency, Office of Surface Mining, Bureau of Land Management, U.S. Forest Service, National Park Service, and Bureau of Indian Affairs.

Safety. Federal health and safety standards applicable to all surface mines are enforced by the Mine Safety and Health Administration. The Office of Surface Mining enforces the federal laws at coal mines concerning water and air pollution standards, while allowing individual states to have primacy over the industry if they enact laws that mostly mirror the federal laws. For these states, the Office of Surface Mining would merely oversee the state enforcement program. Many mining states have enacted such laws.

Environmental protection. Prior to 1977, individual states regulated the environmental practice of the surface coal mining industry. In 1977, a federal Surface Mining Law was enacted. This law established the Office of Surface Mining and provided program regulations, including mining and reclamation standards and the requirements to obtain written permits to mine.

Early surface mining in the United States did not employ the industry standards that are now used. Mining sites were generally not restored after the mining was completed. This resulted in unsightly sites, as well as hazardous conditions, poor vegetation, and water pollution. Modern surface mines have high reclamation standards. They restore the mined area to approximately the same contour as previously existed, revegetate with native grasses and trees, and comply with water pollution standards. Industry and environmental organizations such as the American Society of Mining and Reclamation were formed to promote reclamation practices for lands disturbed during mineral extraction.

Surface mining statistics

	Number of surface mines	Number of employees	Average number of employees per mine
Sand and gravel	7,126*	37,508*	5
Coal	852	28,304	33
Stone	2,412	32,311	13
Metal	89	11,076	124
Nonmetal	418	5,561	13
Total	10,897	114,760	11

*Includes 745 dredging operations.
SOURCE: Mine Safety and Health Administration.

Fig. 6. Bucketwheel excavators. (*a*) Large type, daily capacity 144,000 yd^3 (110,000 m^3). (*b*) Smaller model.

Previously unreclaimed surface mine sites are also being restored by the surface mining industry. Some are restored directly in conjunction with the mining of adjacent mineral resources. *See* LAND RECLAMATION.

Thomas Gray

Bibliography. S. M. Cassidy (ed.), *Elements of Practical Coal Mining*, SME/AIME, New York, 1973; J. T. Crawford III and W. A. Hustrulid (eds.), *Open Pit Mine Planning and Design*, SME/AIME, New York, 1979; *Glossary of Surface Mining and Reclamation Terminology*, 2d ed., Bituminous Coal Research, Monroeville, PA, 1983; T. A. Gray and R. E. Gray, Mine closure, sealing and abandonment (Chap. 8.7), in H. L. Hartman et al. (eds.), *SME Mining Engineering Handbook*, 2d ed., Society for Mining Metallurgy & Exploration, 1992; B. A. Kennedy (ed.), *Surface Mining*, 2d ed., Society for Mining Metallurgy & Exploration, 1990; C. T. Shaw and V. Pavlovic, *Surface Mining and Quarrying: Mechanization, Technology, and Capacity*, Ellis Horwood, 1993; P. W. Thrush (ed.), *A Dictionary of Mining, Mineral, and Related Terms*, Intertech Publishing, 1990.

Surface physics

The study of the structure and dynamics of atoms and their associated electron clouds in the vicinity of a surface, usually at the boundary between a solid and a low-density gas. Surface physics deals with those regions of large and rapid variations of atomic and electron density that occur in the vicinity of an interface between the two "bulk" components of a two-phase system. In conventional usage, surface physics is distinguished from interface physics by the restriction of the scope of the former to interfaces between a solid (or liquid) and a low-density gas, often at ultrahigh-vacuum pressures $p = 10^{-10}$ torr (1.33×10^{-8} newton/m^2 or 10^{-13} atm). *See* SOLID-STATE PHYSICS.

Surface physics is concerned with two separate but complementary areas of investigation into the properties of such solid-"vacuum" interfaces. Interest centers on the experimental determination and theoretical prediction of surface composition and structure (that is, the masses, charges, and positions of surface species), of the dynamics of surface atoms (such as surface diffusion and vibrational motion), and of the energetics and dynamics of electrons in the vicinity of a surface (such as electron density profiles and localized electronic surface states). As a practical matter, however, the nature and dynamics of surface species are determined experimentally by scattering and emission measurements involving particles or electromagnetic fields (or both) external to the surface itself. Thus, a second major interest in surface physics is the study of the interaction of external entities (that is, atoms, ions, electrons, electromagnetic fields, or mechanical probes) with solids at their vacuum interfaces. It is this aspect of surface physics that most clearly distinguishes it from conventional solid-state physics, because quite different scattering, emission, and local probe experiments are utilized to examine surface properties as opposed to bulk properties.

Physical principles of measurements. Since the mid-1960s, surface physics has enjoyed a renaissance by virtue of the development of a host of techniques for characterizing the solid-vacuum interface. These techniques are based on one of three simple physical mechanisms for achieving surface sensitivity. The first, which is the basis for field emission, field ionization, and scanning tunneling microscopy (STM), is the achievement of surface sensitivity by utilizing electron tunneling through the potential-energy barrier at a surface. This concept provides the basis for the development of STM to directly examine the atomic structure of surfaces by measuring with atomic resolution the tunneling current at various positions along a surface. It also has been utilized for direct determinations of the energies of individual electronic orbitals of adsorbed complexes via the measurement of the energy distributions either of emitted electrons or of Auger electrons emitted in the process of neutralizing a slow (energy $E \sim 10$ eV) external ion. *See* FIELD-EMISSION MICROSCOPY; SCANNING TUNNELING MICROSCOPE; TUNNELING IN SOLIDS.

The second mechanism for achieving surface sensitivity is the examination of the elastic scattering or emission of particles which interact strongly with the constituents of matter, for example, "low-energy" ($E \lesssim 10^3$ eV) electrons, thermal atoms and molecules, or "slow" (300 eV $\lesssim E \lesssim 10^3$ eV) ions. Since such entities lose appreciable ($\Delta E \sim 10$ eV) energy in distances of the order of tenths of a nanometer, typical electron analyzers with resolutions of tenths of an electronvolt are readily capable of identifying scattering and emission processes which occur in the upper few atomic layers of a solid. This second mechanism is responsible for the surface sensitivity of photoemission, Auger electron, electron characteristic loss, low-energy electron diffraction (LEED), and ion scattering spectroscopy techniques. The strong particle-solid interaction criterion that renders these measurements surface-sensitive is precisely the opposite of that used in selecting bulk solid-state spectroscopies. In this case, weak particle-solid interactions (that is, penetrating radiation) are desired in order to sample the bulk of the specimen via, for example, x-rays, thermal neutrons, or fast ($E \gtrsim 10^4$ eV) electrons. These probes, however, can sometimes be used to study surface properties by virtue of special geometry, for example, the use of glancing-angle x-ray diffraction to determine surface atomic structure. *See* AUGER EFFECT; ELECTRON DIFFRACTION; ELECTRON SPECTROSCOPY; PHOTOEMISSION; X-RAY CRYSTALLOGRAPHY.

The third mechanism for achieving surface sensitivity is the direct measurement of the force on a probe in mechanical contact or near contact with the surface. At near contact, the van der Waals force can be measured directly by probes of suitable sensitivity. After contact is made, a variety of other forces dominate, for example, the capillary force for solid

surfaces covered with thin layers of adsorbed liquid (that is, most solid surfaces in air at atmospheric pressure). When this mechanism is utilized via measuring the deflection of a sharp tip mounted on a cantilever near a surface, the experiment is referred to as atomic force microscopy (AFM) and results in maps of the force across the surface. Under suitable circumstances, atomic resolution can be achieved by this method as well as by STM. Atomic force microscopy opens the arena of microscopic surface characterization of insulating samples as well as electrochemical and biochemical interfaces at atmospheric pressure. Thus, its development is a major driving force for techniques based on surface physics. *See* INTERMOLECULAR FORCES.

Surface preparation. An atomically flat surface, labeled by $M(hkl)$, may be visualized as being obtained by cutting an otherwise ideal, single-crystal solid M along a lattice plane specified by the Miller indices (hkl), and removing all atoms whose centers lie on one side of this plane. On such a surface the formation of a "selvedge" layer can also be envisaged. Such a layer might be created, for example, by the adsorption of atoms from a contiguous gas phase. It is characterized by the fact that its atomic geometry differs from that of the periodic bulk "substrate." From the perspective of atomic structure, this selvedge layer constitutes the "surface" of a solid. A good example is the thin film of condensed water that exists on most surfaces at atmospheric pressure. In principle, the thickness of the selvedge layer is a thermodynamic variable determined from the equations of state of the solid and the contiguous gas phase. In practice, almost all solid surfaces are far from equilibrium, containing extensive regions (micrometers thick) of surface material damaged by sample processing and handling. *See* CRYSTALLOGRAPHY.

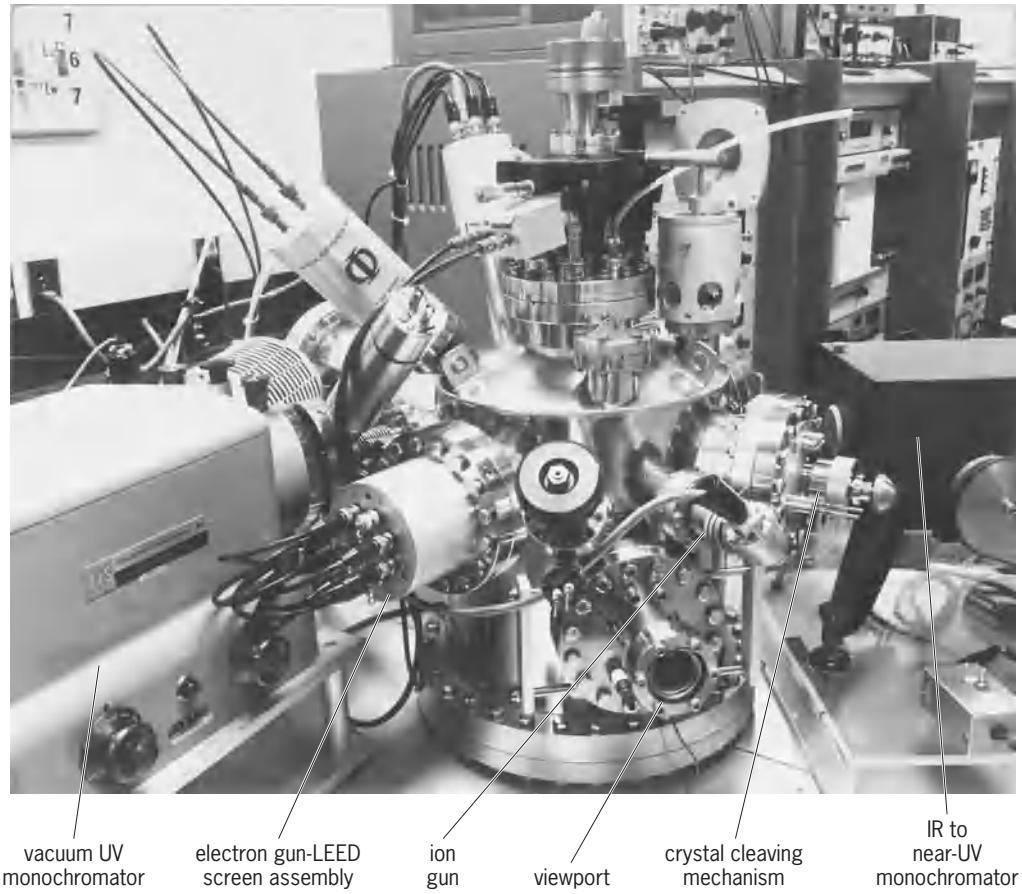
Another reason for the renaissance in surface physics is the capability to generate in a vacuum chamber special surfaces that approximate the ideal of being atomically flat. These surfaces may be prepared by cycles of fast-ion bombardment, thermal outgassing, and thermal annealing for bulk samples (for example, platelets with sizes of the order of 1 cm × 1 cm × 1 mm), molecular beam epitaxy of a thin surface layer on a suitably prepared substrate, or field evaporation of etched tips for field-ion microscopes. Alternatively, the sample may be cleaved in a vacuum chamber. In such a fashion, reasonable facsimiles of uncontaminated, atomically flat solid-vacuum interfaces of many metals and semiconductors have been prepared and subsequently characterized by various spectroscopic techniques. Such characterizations must be carried out in an ultrahigh vacuum ($p \sim 10^{-8}$ N/m^2) so that the surface composition and structure are not altered by gas adsorption during the course of the measurements. *See* EPITAXIAL STRUCTURES.

Experimental apparatus. Modern experimental surface physics is devoted to the determination of the chemical composition, atomic geometry, atomic dynamics, and electronic structure of surfaces. Since different measurements are required to assess each of these four aspects of a surface, the typical surface-characterization instrument consists of equipment for performing multiple measurements in a single ultrahigh-vacuum chamber or for moving a sample between multiple connected chambers without breaking vacuum. Two types of sample geometry are common. Platelet or wafer samples are studied using scattering and emission experiments. A typical modern apparatus (**Fig. 1**) contains an electron gun, an ion gun, an electron energy analyzer, a source of ultraviolet or x-ray electromagnetic radiation, a local probe microscope, and a sample holder permitting precise control of both its orientation and temperature. Occasionally, other features (such as a mass spectrometer) also are incorporated for special purposes. For specific applications in which less than a complete characterization of the surface is required, commercial instruments designed to embody only one or two measurements often are available. Such instruments commonly are utilized to determine the chemical composition of surfaces by, for example, ion scattering, secondary ion mass spectrometry, x-ray photoemission, or Auger electron emission. Similarly, local probe scanning tunneling microscopes and atomic force microscopes are available commercially. Obviously, the utility or such instruments is not limited to atomically flat or even crystalline surfaces, so that they find widespread applications in metallurgy and polymer science. Analogous measurements also are commonly utilized for the in situ characterization of materials made by vacuum deposition techniques, such as sputtering or molecular beam epitaxy. *See* MASS SPECTROMETRY; SPUTTERING.

The second common sample geometry is an etched tip, about 100 nm in radius. Such specimens are studied by field emission and ionization experiments. These experiments provide a direct magnified image of the surface structure in contrast to the statistical description of platelet or wafer surfaces afforded by instruments like that in Fig. 1 or the composite local images of such surfaces afforded by STM or AFM.

Data acquisition, analysis, and theory. Given the ability to perform surface-sensitive spectroscopic measurements, questions naturally arise concerning analysis of the raw spectra to extract parameters characterizing the structure of a given surface and the synthesis of such data to form a coherent picture of the behavior of electrons and atomic species at the surface. Thus, surface physics may be divided into three types of activity: the acquisition of surface-sensitive spectroscopic data, the analysis of these data using physical models of the appropriate scattering or emission spectroscopy, and the construction of theoretical models of surface structure and properties to be tested via critical comparison of their predictions with the results of such data analyses.

Ground- and excited-state properties. Theoretical models have been proposed for the description of two distinct types of surface properties. The stability of surface structures is examined by calculations of ground-state properties, such as surface energies



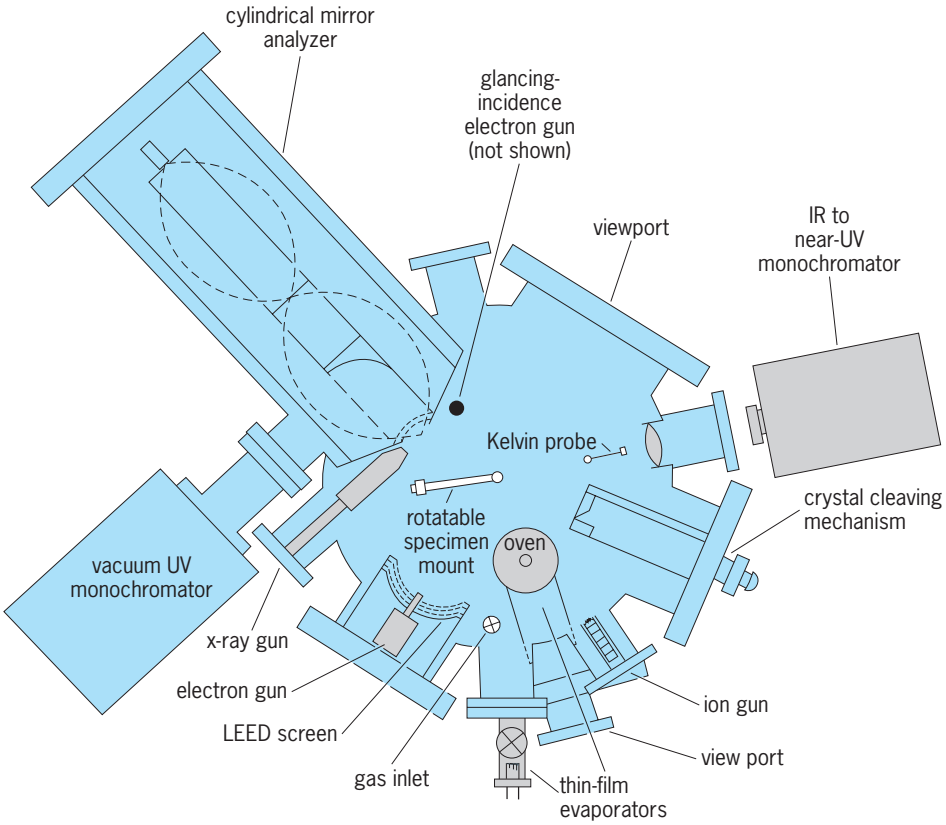


Fig. 1. Photograph and schematic diagram of a multiple-technique ultrahigh-vacuum surface characterization instrument for the study of semiconductor and insulator surfaces. (*Courtesy of L. J. Brillson, Xerox Corporation*)

or effective potential-energy diagrams for adsorbed species. These quantities are difficult to measure experimentally, although they are the most direct manifestations of the intrinsic behavior of an undisturbed surface. The interactions of external projectiles or fields with a solid create excited states of the electrons or atoms within the solid. Consequently, the associated scattering and emission spectra indicate the nature and energies of these excited states (called excitations) rather than of the ground state. Two kinds of excitations occur. Electronic excitations are generated when a disturbing force causes the electrons in the solid to alter their quantum states, whereas atomistic excitations are associated with the vibration or diffusion of atomic species (such as adsorbed atoms of molecules). It is important to distinguish between ground-state properties, electronic excitations, and atomic excitations because different models are used to describe each of these three types of phenomena. *See* EXCITED STATE; GROUND STATE.

Quantum theory of surfaces. The theory of the properties of solid surfaces does not differ in any fundamental way from the quantum theory of bulk solids. Specifically, the conventional quantum theory of interacting electron systems is thought to be applicable, although technical refinements are required because of the loss of translation symmetry and the presence of large electron density gradients normal to the surface. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM THEORY OF MATTER.

Macroscopic models. It is premature to speak of an embracing theory of surface phenomena. Rather, a diverse array of specific models has been proposed for the description of various properties. In the case of macroscopic models, the presence of a surface is treated as a boundary condition on an otherwise continuum theory of bulk behavior. Such models have found widespread use in semiconductor and insulator physics because the penetration depth of electrostatic fields associated with surface charges usually is large ($\lambda_e \sim 10^4$–10^5 nm) relative to the spatial extent of the charges themselves ($d \lesssim 1$ nm). Moreover, they continue to describe adequately the electrostatic fields involved in the operation of microelectronic semiconductor devices down to submicrometer dimensions. To describe the atomic and electronic properties associated with the upper few atomic layers at a surface, however, one must make use of a description of surfaces at the atomic or electronic level. This is becoming an issue for semiconductor device technology as the size of these devices approaches 0.1 μm and less. *See* SEMICONDUCTOR.

Microscopic models. Four major classes of microscopic models of surface properties have been explored. The simplest consists of models in which consideration of the electronic motion is suppressed entirely, and the solid is visualized as composed of atomic species interacting directly, for example, via two-body forces. While such models may suffice to describe the vibrational motion of atoms near a surface, they are inadequate to describe ground-state properties such as adsorbate potential-energy curves. Nevertheless, such models can be used to describe interactions between steps and kinks formed on a surface by growth or annealing. In this context they have found considerable favor to describe the evolution of the morphologies of surfaces during growth and processing.

The next more sophisticated models are empirical quantum-chemical models (such as tight-binding or empirical pseudopotential models, in solid-state terminology), in which electronic motions are considered explicitly but electron-electron interactions are incorporated either phenomenologically or not at all. Such models have proved useful in solid-state physics, although their value for surface physics is more limited because the large charge rearrangements (relative to the bulk) which occur at surfaces require an accurate, self-consistent treatment of both electron-electron and electron-ion interactions.

The simplest model in which electron-electron interactions are treated explicitly is the jellium model of metals, in which the positive charge associated with the ion cores immersed in the sea of conduction electrons is replaced by a uniform positive background charge terminating along a plane. This model, popular in the early 1970s, permits an informative but approximate treatment of electron-electron interactions at the expense of losing the effects of atomic lattice structure because of the uniform-positive-background hypothesis. Since 1990, computers have permitted the construction of realistic pseudopotential, local-density, and quantum-chemical models in which both the electron-electron and electron-ion interactions are treated in a self-consistent, if approximate, fashion. Such models are now routinely applied to predict ground-state electronic charge densities and surface excitation spectra (both electronic and atomic) at the low-index surfaces of essentially any metal, semiconductor, or insulator. The major tests of their adequacy arise from comparisons of their predictions with measured atomic geometries, work functions, photoemission spectra, and characteristic electron-loss spectra. Quite detailed calculations of the work functions, surface-state eigenvalue spectra, and surface geometries are now available for a variety of metal and semiconductor surfaces. The geometry of both metal and semiconductor surfaces can be predicted quantitatively by energy-minimization techniques applied to these models. Moreover, they are currently being extended to study atomic motions at surfaces associated with crystal growth and the dynamics of chemical reactions at the atomic level. *See* CRYSTAL GROWTH; QUANTUM CHEMISTRY.

Theoretical models for data analysis. Another group of theoretical models in surface physics consists of those utilized to analyze observed scattering and emission spectra in order to extract quantitative assessments of the atomic and electronic structure of surfaces. These models differ substantially from their bulk counterparts because of the necessity of strong particle-solid interactions to achieve surface-sensitive spectroscopies. Consequently, the fundamental assumption underlying the linear-response

theory of bulk solid-state spectroscopies—that is, the appropriate particle-solid interaction is weak and hence can be treated by low-order (usually first) perturbation theory—is invalid. This fact results in collision theories of surface-sensitive particle-solid scattering exhibiting a considerably more complicated analytical structure in order to accommodate the strong elastic as well as inelastic scattering of the particle by the various constituents of the solid.

Applications to LEED and photoemission. While the above considerations are quite general, the special case in which they have been developed in most detail is the coherent scattering (that is, diffraction) of low-energy electrons from the surfaces of crystalline solids. This is an important case because elastic low-energy electron diffraction (LEED) is the analog of x-ray diffraction for surfaces—that is, it is the major vehicle for the achievement of a quantitative surface crystallography. Since 1968 quite complete quantum field theory models of the LEED process have been developed, tested, and reduced to computational algorithms suitable for the routine analysis of LEED intensity data. From such analyses the surface atomic geometry of the low-index faces of a host of metals and semiconductors has been determined, as have the geometries of a wide variety of overlayer structures. Local surface structures of adsorbates have also been determined by applying these techniques to analysis of the diffraction of photoemitted electrons from the adsorbates, and to the diffuse scattering of incident electrons by disordered adsorbate overlayers. Atomic geometries of semiconductors generated in situ by molecular beam epitaxy are determined by the application of these techniques to analyze the diffracted intensities of glancing-incidence high-energy (keV) electrons via reflection high-energy electron diffraction (RHEED). Similar quantitative analyses of inelastic low-energy electron diffraction intensities have yielded the energy-momentum relations of collective surface electronic excitations (for example, surface plasmons). *See* PLASMON.

Data acquisition. It is the development of a host of novel surface-sensitive spectroscopic techniques, however, which has provided the foundation for the renaissance in surface physics. Having recognized that low-energy electrons, thermal atoms, and slow ions all constitute surface-sensitive incoming or exit entities in particle-solid collision experiments, one can envisage a wide variety of surface spectroscopies based on these plus quanta of electromagnetic radiation (photons) as possible incident or detected species. Most of these possibilities have been realized in some form. The selection of which technique to use in a particular application depends both upon what one wishes to learn about a surface and upon the relative convenience and destructiveness of the various measurements.

Typically, one wishes to determine the composition and atomic structure of a surface region, and often to ascertain its electronic structure as well. In the measurements of any of these quantities, important issues are the lateral and depth resolution of the possible techniques. In the case of scattering-based probes, the depth resolution is determined by the particle-solid force law of the incident and exit particles, high resolution being associated with stronger inelastic collision processes. The lateral spatial resolution depends on the ability to focus the incident beam. For typical focused-beam experiments, this is on the order of 1 cm^2 for photon beams, 10^{-8} to 10^{-12} cm^2 for electron beams, and 10^{-8} cm^2 for ion beams. Thus, scanning microscopies are both feasible and common with electron and ion beams. For photon beams, more sophistication is required. By using shaped fiber or semiconductor tips as light pipes and an appropriate local probe geometry, scanning near-field optical microscopy (SNOM) can yield useful images with lateral resolutions comparable to and below the wavelength of the incident light. Depth resolution is a single monolayer for thermal-atom and slow-ion scattering, and a few monolayers for slow-electron scattering. It can become 1000 atomic layers or more, however, for fast (MeV) ions and fast (10-keV) electrons. *See* OPTICAL MICROSCOPE.

The development of local probe STM and AFM methods has revolutionized surface physics by providing readily accessible maps of the geometry, from atomic to morphological distance scales (for example, steps or islands), of complex surfaces. The interpretation of the raw data is more complicated than commonly realized, because the depth resolution of these techniques is so high. For example, STM measures the electronic charge density in a region a fraction of a nanometer outside the geometrical surface defined by the position of the nuclei of the surface atoms. Similar caveats apply to SNOM, for which quantitative determination of the electromagnetic fields in the vicinity of the probe tip is required for precision data analysis. With the proper care in interpreting the data, however, these local probes offer the possibility of the application of surface physics techniques to characterize a wide range of surfaces, interfaces, and atomic processes (such as crystal growth or surface chemical reactions).

Surface composition. The elemental composition of surfaces is specified by measuring the masses or atomic numbers, or both, of resident species. Their masses may be ascertained either by the elastic backscattering of slow incident ions (ion scattering spectrometry, or ISS) or by using such ions to erode the surface, detecting the ejected surface species in a mass spectrometer (secondary ion mass spectrometry, or SIMS). *See* SECONDARY ION MASS SPECTROMETRY (SIMS).

The atomic numbers of surface species are determined by measuring the energy of tightly bound core electrons. A schematic diagram illustrating the nature and labeling of the various physical processes that can be utilized to accomplish this task is shown in **Fig. 2**. An electron, photon, or chemical species incident on a surface excites a low-energy core electron. The binding energy of this electron commonly is determined by measuring the energy loss of the incident electron (characteristic loss spectroscopy,

or CLS), the energy of the core electron ejected by an incident x-ray photon (x-ray photoelectron spectroscopy, or XPS, sometimes referred to as electron spectroscopy for chemical analysis, or ESCA), or the threshold energy of an incident particle necessary to generate a threshold in the secondary x-ray yield (soft x-ray appearance potential spectroscopy, or SXAPS). Alternatively, the binding energy of the core electron may be ascertained by secondary processes in which an initially empty core state (generated by a direct process) is filled by an electron in a higher-energy state. If the filling process is radiative recombination, then the energy of the emitted x-ray yields the binding energy (soft x-ray emission spectroscopy). If this process is radiationless, however, the energy of the electron excited by the Auger process indicates the binding energy of the initially empty core state (Auger electron spectroscopy, or AES).

These techniques operate on the dictionary premise; that is, calibration spectra are obtained on surfaces of independently known composition, with elemental analysis on unknown samples being performed by comparison of their spectra with the reference calibration spectra. Consequently, although the detailed interpretations of observed line shapes often has eluded surface physicists, the use of these spectroscopies for elemental analysis has proved both practical and eminently useful. Difficulties in interpretation usually preclude the use of these techniques for quantitative chemical analysis (for example, the determination of whether C and O are adsorbed on aluminum as CO, as CO_2 or C on Al_2O_3, and so on). Progress has been made, however, in developing this aspect of the core-electron spectroscopies. *See* ATOMIC SPECTROMETRY.

Surface atomic geometry. The atomic geometry of planar surfaces of crystalline solids usually is obtained by electron diffraction, although in certain cases slow-ion backscattering, megaelectronvolt ion channeling, photoelectron diffraction, surface-sensitive extended x-ray absorption fine structure, or valence-electron photoemission spectroscopy also may be employed. Two experimental configurations commonly are used, as indicated in **Fig. 3**. The RHEED configuration embodies glancing incidence electrons at kiloelectronvolt energies. It yields primarily the space-group symmetry of the surface and is quite sensitive to surface topography. The LEED experiment consists of measuring the backscattering intensities of electrons in the energy range 50 eV $\lesssim E \lesssim$ 500 eV. The configuration of diffracted beams reveals the space group symmetry of the surface structure, whereas analysis of their intensities permits determination of their atomic geometry.

Surface atomic geometry may also be determined by direct imaging. In the case of tip sample geometries, field ion microscopy permits the direct imaging of atoms on the tip surface. For platelet and wafer samples, the local probe STM and AFM methods afford scanning determinations of the local charge density and the force, respectively, just outside the surface. These are related to the atomic geometry of the surface species, but only indirectly. Nevertheless, maps of tip height (used in STM) or cantilever deflection (used in AFM) versus lateral position along the surface are commonly interpreted as reflecting the surface atomic geometry. While this is rarely adequate for precision determination of surface atomic positions, it is very useful for the observation of qualitative details of the surface atomic geometry, especially for surface defects.

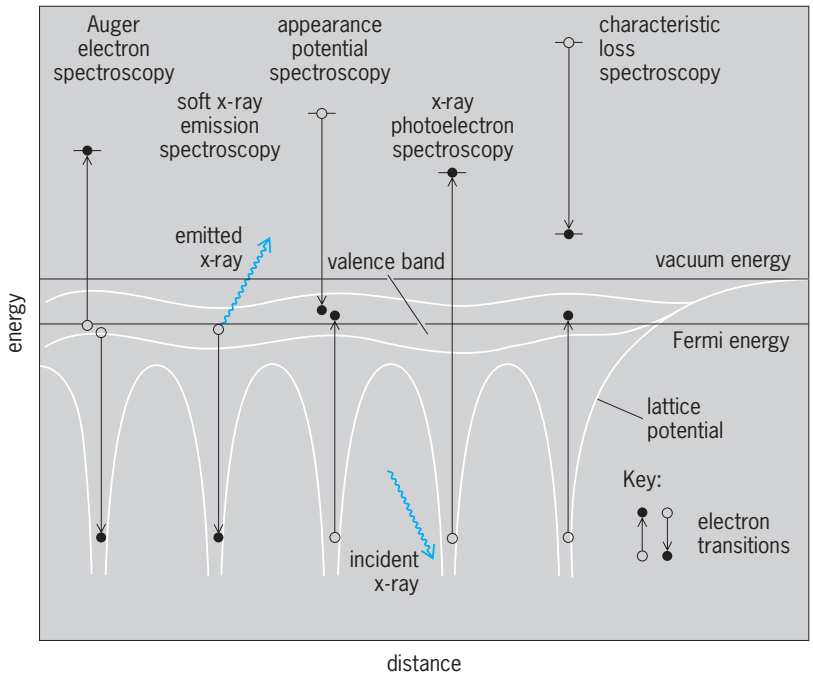


Fig. 2. Schematic diagram of the core-electron transitions utilized to ascertain the atomic number of surface species. (*After R. L. Park, Inner-shell spectroscopy, Phys. Today, 28(4):52–59, April 1975*)

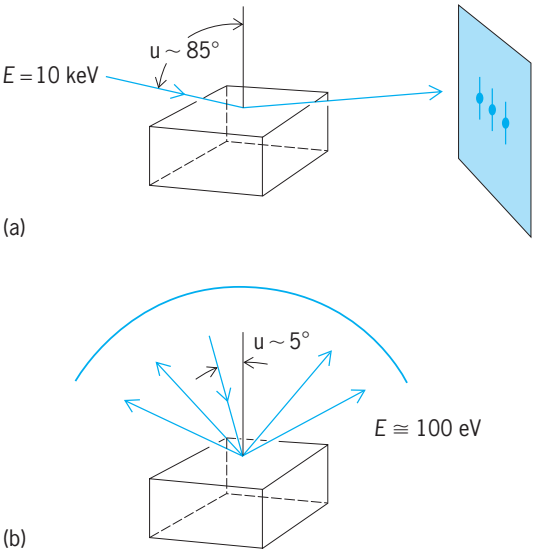


Fig. 3. Schematic diagram of the two electron diffraction techniques used to determine the precision atomic geometry of single-crystal surfaces. (*a*) Reflection high-energy electron diffraction (RHEED). (*b*) Elastic low-energy diffraction (LEED). (*After C. B. Duke, in I. Prigogine and S. A. Rice, eds., Aspects of the study of surfaces, Adv. Chem. Phys., 27(1):1–209, 1974*)

Surface atomic motion. The vibrational motion of surface species may be examined either by analysis of the temperature dependence of LEED intensities or by direct observation of small ($\Delta E \sim 0.01$ eV) electron energy losses caused by the excitation of a normal mode of vibration. The first approach provides the rms vibrational amplitudes of surface species, whereas the second yields the frequencies of localized surface normal modes of vibration. Both high-resolution electron energy-loss spectroscopy (HREELS: $\Delta E \sim 0.01$ eV) and infrared reflection and transmission spectroscopy have become popular techniques for the measurements of the vibrational spectra of adsorbed species. The complete energy momentum relations of surface vibrational excitations have been determined by inelastic helium (He) atom scattering. *See* LATTICE VIBRATIONS.

Surface electronic structure. The electronic structure of a solid-vacuum interface is studied by measuring the emission of valence electrons (induced by external fields, electrons, ions, or photons) or the inelastic scattering of an incident electron. A special situation arises when an emitted Auger electron is a valence electron. In this case, the initially empty core state is highly localized in space. Consequently, the emission line shape is a measure of the local electronic structure in the vicinity of this core state. The shifts in energy of core-level photoemission and Auger transitions (called chemical shifts) caused by the nearby electronic charge densities also yield an indication of the local electronic structure around a particular kind of surface atom.

In contrast to these emission processes involving localized core electrons, the photoemission, field emission, ion neutralization, and characteristic loss spectroscopies of valence electrons provide measures of their average behavior in the vicinity of a surface. Indeed, for precisely this reason such spectra from clean surfaces may be difficult to interpret because the distinction between bulk and surface features is often vague. Thus, their major use has occurred in the arena of chemisorption, in which case the changes in spectra upon adsorption can be monitored, and qualitative features of the electronic structure of the chemisorbed complexes can be inferred therefrom. A proper theoretical analysis of these valence electron emission processes is needed to convert the observed spectra into quantitative indicators of surface structure. Such analyses are not yet routine, although they have been given for a few exemplary materials such as clean nickel and gallium arsenide (GaAs) surfaces and a few chemisorption systems. The advent of synchrotron radiation has provided a major incentive for the development of computer programs to perform such analyses. *See* ADSORPTION; SYNCHROTRON RADIATION.

Several measurements probe the electric fields just outside the surface. Scanning tunneling microscopy measures the local electron density about 0.1 nm outside a surface. The effective potential seen by an atom scattering from a surface is also quite sensitive to the electronic structure in this region. Finally, work function measurements probe the dipole layer between the valence electrons and surface layer of ion cores in the solid. *See* WORK FUNCTION (ELECTRONICS).

Outlook. Since 1975, surface physics has advanced from a state in which the atomic composition, geometry, and atomic and electronic excitation spectra were essentially unknown, to one in which these quantities are routinely determined for the low-index faces of elements and binary alloys. Local probe microscopes have extended the reach of surface physics techniques from carefully prepared surfaces in ultrahigh vacuum to the domains of the biologist, chemist, and technologist. Surface physics is used to explore novel phenomena and characterize complex materials systems, especially various manifestations of atomic dynamics at surfaces and interfaces.

C. B. Duke

Bibliography. M. C. Desjonqueres and D. Spanjaard, *Concepts in Surface Physics*, 1993; C. B. Duke (ed.), *Surface Science: The First Thirty Years*, 1994; J. B. Hudson, *Surface Science: An Introduction*, 1991; J. Israelachvili, *Intermolecular and Surface Forces*, 1992; H. Luth, *Surfaces and Interfaces of Solids*, 1993; S. R. Morrison, *The Chemical Physics of Surfaces*, 2d ed., 1990; D. W. Pohl and D. Courjon (eds.), *Near Field Optics*, 1992; M. Prutton, *Introduction to Surface Physics*, 1994; W. N. Unertl (ed.), *Handbook of Surface Science I: Physical Structure*, 1996; M. A. van Hove et al. (eds.), *The Structure of Surfaces*, vols. 1–4, 1985–1994; J. M. Walls, *Surface Science Techniques*, 1994; D. P. Woodruff and T. A. Delchar, *Modern Techniques of Surface Science*, 2d ed., 1994.

Surface tension

The force acting in the surface of a liquid, tending to minimize the area of the surface. Surface forces, or more generally, interfacial forces, govern such phenomena as the wetting or nonwetting of solids by liquids, the capillary rise of liquids in fine tubes and wicks, and the curvature of free-liquid surfaces. The action of detergents and antifrothing agents, and the flotation separation of minerals depend upon the surface tensions of liquids.

Surface energy. In the body of a liquid, the time-average force exerted on any given molecule by its neighbors is zero. Even though such a molecule may undergo diffusive displacements because of random collisions with other molecules, there exist no directed forces upon it of long duration. It is equally likely to be momentarily displaced in one direction as in any other. In the surface of a liquid, the situation is quite different; beyond the free surface, there exist no molecules to conteract the forces of attraction exerted by molecules in the interior for molecules in the surface. In consequence, molecules in the surface of a liquid experience a net attraction toward the interior of a drop. These centrally directed forces cause the droplet to assume a spherical shape, thereby minimizing both the free energy and surface area.

From the macroscopic point of view, surface tension may be regarded either as a force exerted normally to a unit length in the surface, or as the work which must be expended upon the liquid to increase its area by unity. Accordingly, surface tension is expressed in SI units of newtons per meter (N/m) or joules per square meter (J/m^2). From the microscopic point of view, the surface tension (or its equivalent, surface energy) is the reversible isothermal work which must be done in bringing molecules from the interior of the liquid to the surface and creating 1 cm^2 of new surface thereby.

Most liquids have surface tensions of 20–40 millinewtons per meter at room temperature, but water has the exceptionally high value of 72.75 mN/m at 20°C (68°F). Condensed gases such as helium and nitrogen have quite low surface tensions [98 micronewtons per meter (−452°F) and 6.2 mN/m at 90.2 K (−297.2°F) respectively]. Liquid metals have large surface tensions by comparison: mercury, 0.47 N/m; and liquid copper at 1131°C (2068°F) has a surface tension of 1.103 N/m in hydrogen gas. Small but significant differences in the surface tensions of liquids depend upon the composition of the vapor phase.

In the wetting or nonwetting of solids by liquids, the criterion employed is the contact angle between the solid and the liquid (measured through the liquid) (**Fig. 1**). A liquid is said to wet a solid if the contact angle θ lies between 0 and 90°, and not to wet the solid if the contact angle lies between 90 and 180°. Three interfaces exist when a droplet of liquid contacts a solid, and three corresponding interfacial tensions exist: γ_{SL}, γ_{SV}, and γ_{LV}. The subscripts S, L, and V refer to solid, liquid, and vapor. At equilibrium, a balance of interfacial tensions exists at the line of common contact, which intersects the figures at point 0. For the case of a liquid which wets the solid ($\theta < 90°$), this equilibrium is expressed by Eq. (1).

$$\gamma_{SV} = \gamma_{SL} + \gamma_{LV} \cos\theta \tag{1}$$

Capillarity. Liquids which wet the walls of fine capillary tubes rise to a height which depends upon the tube radius, the surface tension, the liquid density, and the contact angle. In **Fig. 2**, a liquid of density ρ is shown as having risen to a height H in a capillary whose radius is R. A balance exists between the force exerted by gravity on the mass of liquid raised in the capillary and the opposing force caused by surface tension. The former is $\pi r^2 h \rho g$, whereas the latter is $2\pi r\gamma$, assuming the contact angle to be zero.

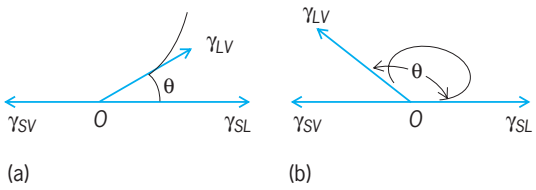


Fig. 1. Contact angle. (*a*) Liquid wets solid. (*b*) Liquid does not wet solid.

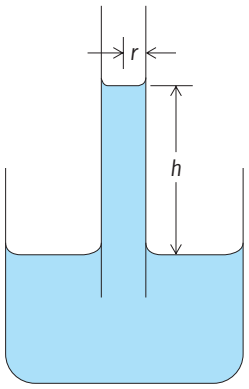


Fig. 2. Rise of liquid in capillary tube.

It is clear that $h = 2\gamma/r\rho g$, and that the capillary rise varies inversely with the tube radius and the liquid density. Liquids which do not wet the capillary walls are depressed in height according to the same equation.

The shape of the free surface of a liquid in a vessel is only an approximation to a plane. In narrow tubes the meniscus of a liquid is concave upward if the liquid wets the tube and, conversely, convex upward if it does not wet the tube. A pressure difference exists between the concave and convex sides of the surface, the excess pressure on the concave side over the convex side being given by Eq. (2), where r_1

$$p = \gamma \frac{1}{r_1} + \frac{1}{r_2} \tag{2}$$

and r_2 are the principal radii of curvature of the surface. The same equation applies for a bubble of gas within a liquid, with the consequence that the vapor pressure p is larger for small bubbles according to Eq. (3), where p_0 is the vapor pressure over a liquid

$$\ln\left(\frac{p}{p_0}\right) = \frac{2\gamma}{r\rho}\frac{M}{RT} \tag{3}$$

surface of infinite radius, R is the gas constant, and M is the molecular weight. *See* VAPOR PRESSURE.

Detergents, soaps, and flotation agents owe their usefulness to their ability to lower the surface tension of water, thereby stabilizing the formation of small bubbles of air. At the same time, the interfacial tension between solid particles and the liquid phase is lowered, so that the particles are more readily wetted and floated after attachment to air bubbles. *See* FLOTATION; INTERFACE OF PHASES; SURFACTANT.

Norman H. Nachtrieb

Bibliography. R. A. Alberty and R. J. Silbey, *Physical Chemistry*, 3d ed., 2000; K. L. Mittal, *Contact Angle, Wettability and Adhesion: In Honor of Professor Robert J. Good*, 1993; J. H. Noggle, *Physical Chemistry*, 3d ed., 1987.

Surface waves

Disturbances that are propagated at a gas-liquid interface and are dependent primarily upon the gravitational property of the liquid (surface tension and

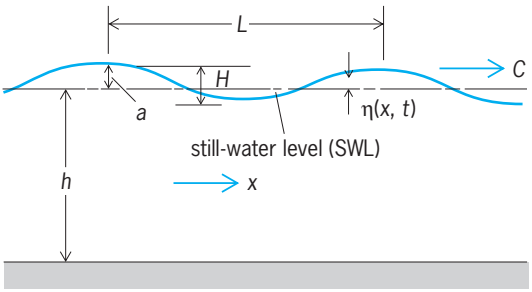


Fig. 1. Definition sketch for an oscillatory wave.

viscosity being of secondary importance). Wave motions that occur in confined fluids (either liquid or gaseous) are dependent primarily upon the elastic property of the medium. *See* WATER HAMMER; WAVE MOTION IN FLUIDS.

The fundamental concepts of gravity waves in liquids are presented in the one-dimensional form. For information on the generation of waves by wind in natural bodies of water and their transformation in coastal regions. *See* OCEAN WAVES.

Oscillatory waves. The term oscillatory implies a periodicity in the form of a disturbance moving past a fixed point. **Figure 1** is a definition sketch for an oscillatory wave propagating in a liquid of constant density ρ and depth h measured from the bottom to the still-water level (SWL). Wavelength L is the horizontal distance between successive crests of the wave. Wave height H is the vertical distance from crest to trough; amplitude a is the distance from the still-water level to the crest, and η is the elevation of the free surface with respect to the still-water level at any position x and instant of time t. In the linearized theory of small-amplitude waves ($H/L < 0.03$) the wave profile is sinusoidal and is givenby Eq. (1),

$$\eta = a \sin\left[2\pi\left(\frac{t}{T} - \frac{x}{L}\right)\right] \tag{1}$$

where T is the wave period. By definition, the celerity or speed of propagation $C = L/T$ and is given by Eq. (2) (in SI units or any other coherent system of

$$C = \sqrt{\left(\frac{\sigma}{\rho}\frac{2\pi}{L} + \frac{gL}{2\pi}\right)\tanh\left(2\pi\frac{h}{L}\right)} \tag{2}$$

units). The first term on the right expresses the influence of surface tension σ and need be considered only for waves of very small length (of the order of magnitude of 1 in. or 2.5 cm). In the remaining development, only gravity waves will be considered, as expressed by the second term of the celerity equation, where g is the acceleration of gravity (9.8 m/s^2 or 32.2 ft/s^2). *See* UNITS OF MEASUREMENT.

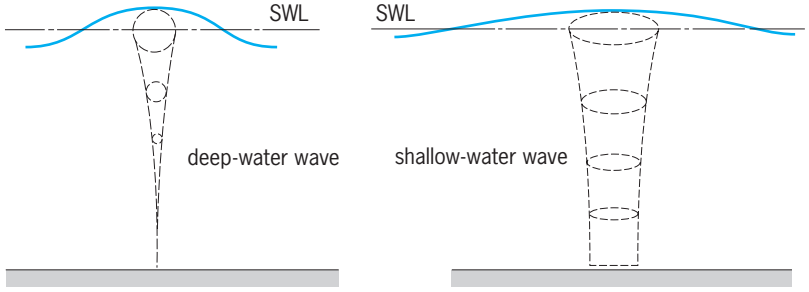


Fig. 2. Fluid particle orbital motions in deep- and shallow-water oscillatory waves. SWL = still-water level.

Oscillatory waves may be generated in a rectangular channel by a simple harmonic translation of a vertical wall forming one end of the flume. The wave amplitude will be determined by the displacement (stroke) of the wall, and the wavelength will be a function of the period of oscillation. The two major classes of oscillatory waves, deep-water and shallow-water waves, are determined by the magnitude of the ratio of liquid depth to wavelength h/L. An inspection of the celerity equation for gravity waves shows that as the depth becomes large in comparison with wavelength, Eqs. (3) hold. Hence, in a deep-water

$$\tanh\left(2\pi\frac{h}{L}\right) \to 1 \qquad C = \sqrt{\frac{gL}{2\pi}} \tag{3}$$

wave the celerity is a function only of the wavelength. This approximation is close if the depth is greater than one-half the wavelength. On the other hand, as the depth becomes small in comparison with wavelength, Eqs. (4) hold. Hence, the celerity

$$\tanh\left(2\pi\frac{h}{L}\right) \to 2\pi\frac{h}{L} \qquad C = \sqrt{gh} \tag{4}$$

depends only on the depth in a shallow-water wave. Somewhat arbitrarily, the limit $hL = 1/10$ is generally applied to this type of wave motion. In deep-water waves, individual fluid particles tend to move in circular orbits. The radius of the surface particle orbit is equal to the wave amplitude and the radius decreases exponentially with depth (**Fig. 2**). At a depth of one-half the wavelength, the orbital radius is about 1/20 of the amplitude. A zone of essentially zero fluid motion is rapidly approached and the character of the wave is therefore not affected by the total depth of the liquid.

In shallow water, no vertical particle motion can exist at the bottom; thus the wave characteristics are modified. The particle orbits are flat ellipses in which the minor axis is depressed to zero at the bottom (Fig. 2).

The energy of a wave consists of equal amounts of potential energy (due to particle position above or below the still water level) and kinetic energy (due to the motion of particles in their orbits). The rate of propagation of energy in the direction of wave travel is known as the group velocity to distinguish it from phase velocity C. In deep-water waves the group velocity is one-half the phase velocity; in shallow-water waves the two propagation velocities are equal.

Standing waves. A standing wave can be considered to be composed of two equal oscillatory wave trains traveling in opposite directions. The phase velocity of the resulting wave is zero; nevertheless the velocity of propagation of the component waves retains its usual meaning. In the notation of the previous section, the equation for the profile of a standing wave is obtained by adding the elevations of waves moving in the positive and negative x directions, given by Eqs. (5) and (6), respectively. Hence

Eq. (7) holds. If the length of the basin l in which a disturbance occurs is an integral number n of half wavelengths, a self-perpetuating (except for frictional dissipation) standing wave will result. Therefore, if $l = nL/2$, Eq. (8) is valid. For long waves, as with shallow water, in a basin of uniform depth, the period of oscillation T is defined by Eq. (9). Standing waves frequently occur in canal locks as a result of filling disturbances and in large lakes, bays, and estuaries as a result of wind or tidal action. *See* SEICHE.

$$\vec{\eta}_1 = a \sin\left[2\pi\left(\frac{t}{T} - \frac{x}{L}\right)\right] \tag{5}$$

$$\overleftarrow{\eta}_2 = a \sin\left[2\pi\left(\frac{t}{T} + \frac{x}{L}\right)\right] \tag{6}$$

$$\eta = \eta_1 + \eta_2 = H \sin\left(2\pi\frac{t}{T}\right)\cos\left(2\pi\frac{x}{L}\right) \tag{7}$$

$$\eta = H \sin\left(2\pi\frac{t}{T}\right)\cos\left(\frac{\pi n x}{l}\right) \tag{8}$$

$$T = \frac{2l}{n\sqrt{gh}} \tag{9}$$

Solitary waves. A solitary wave consists of a single crest above the original liquid surface which is neither preceded nor followed by another elevation or depression of the surface. Such a wave is generated by the translation of a vertical wall starting from an initial position at rest and coming to rest again some distance downstream. In practice, solitary waves are generated by a motion of barges in narrow waterways or by a sudden change in the rate of inflow into a river; they are therefore related to a form of flood wave. The amplitude of the wave is not necessarily small compared to the depth, and the wavelength is theoretically infinite because the elevation of the surface approaches the still water level asymptotically with distance as shown in **Fig. 3**. The profile of the solitary wave is given by Eq. (10) and the celerity by Eq. (11). When the solitary wave amplitude becomes approximately equal to the depth, the wave profile becomes unstable and a breaking wave results.

$$\eta = a\,\mathrm{sech}^2\left[\frac{x}{h}\sqrt{\frac{3}{4}\frac{a}{h}}\right] \tag{10}$$

$$C = \sqrt{g(h+a)} \tag{11}$$

Surges. A surge is generated by the forward motion of a vertical wall, at a constant speed, as shown schematically in **Fig. 4**. Surges in open channels are analogous to shock waves produced in a tube by the continuous motion of a piston. A zone of violent eddy motion occurs at the wavefront and the analysis of such motions must take into account the appreciable energy dissipation in this region. The velocity of propagation of a surge is given by Eq. (12). If a velocity V_1 equal and opposite to C is imposed on the fluid upstream of the disturbance, the absolute velocity of the surge front will become zero. In this form the surge is known as a hydraulic jump and it is often used as a means of dissipating flow energy at the bottom of dam spillways. *See* HYDRAULIC JUMP.

$$C = \sqrt{gh_1}\left[\frac{1}{2}\frac{h_2}{h_1}\left(\frac{h_2}{h_1} + 1\right)\right]^{1/2} \tag{12}$$

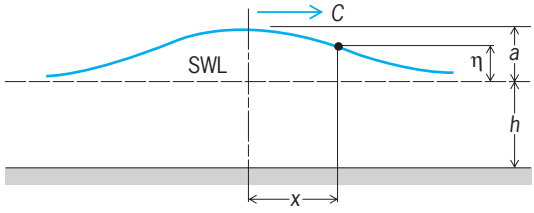


Fig. 3. Definition sketch for a solitary wave. SWL = still-water level.

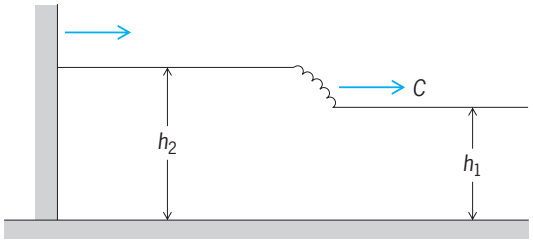


Fig. 4. Definition sketch for a surge wave.

Donald R. F. Harleman

Bibliography. R. G. Dean and R. A. Dalrymple, *Water Wave Mechanics for Engineers and Scientists*, 1991; Open University Course Team Staff (ed.), *Waves, Tides and Shallow-Water Processes*, 2d ed., 1999; J. J. Stoker, *Water Waves*, 1958, reprint 1992; A. Torum and O. T. Gudmestad (eds.), *Water Wave Kinematics*, 1990.

Surfactant

An amphiphilic (also called amphipathic) compound that adsorbs at interfaces to form oriented monolayers and shows surface activity. An amphiphilic compound is a molecule that has a hydrophilic (polar) head and a hydrophobic (nonpolar) tail. Common synonyms for the term "surfactant" include amphiphile, surface-active agent, and tenside. If a surfactant is placed into contact with both a polar medium, such as water, and a nonpolar medium, such as an oil, one part of its molecule has an affinity for the polar medium and one part that has an affinity for the nonpolar medium. An example is an organic molecule such as sodium dodecyl sulfate (SDS), which can be thought of as having a hydrocarbon (dodecyl) tail and a highly polar (sulfate) head group. If the molecule is placed into a system containing water and oil, the sulfate head group will have an affinity for the water, while the dodecyl tail will have an affinity for the oil. The energetically most favorable orientation for such molecules is at the interface between the polar and nonpolar media, so that each part of the molecule can reside in an environment for which it has the greatest affinity. In the case of SDS in a mixture of oil and water, the SDS molecules will preferentially adsorb at the water/oil interface, with the polar sulfate

group oriented into the water and the dodecyl tail group oriented into the oil. Three consequences of the amphiphilic nature of surfactants are their ability (1) to adsorb and form layers at interfaces, (2) to reduce the interfacial tension between fluids, and (3) to associate to form clusters, called micelles. *See* ADSORPTION; INTERFACE OF PHASES; MICELLE; SURFACE TENSION.

When surfactant molecules adsorb at an interface, they provide an expanding force that acts against the normal interfacial tension. At the interface between water and air, for example, there exists a natural interfacial tension that arises from chemical differences between the water and the air. The interfacial tension acts to cause the water/air interface to always strive toward a condition that minimizes the area of contact (interfacial area) between the two phases. This interfacial tension is what causes a waterdroplet to strive to achieve the shape of a sphere (the shape having the smallest interfacial area). When a small amount of a surfactant is added to the water/air system, surfactant molecules adsorb at the interface and reduce the interfacial tension. This makes it easier for the system to adopt shapes having a higher interfacial area. A waterdroplet in air that contains some surfactant can more easily distort and adopt shapes other than spherical. This makes it easier to form foams. The same thing happens when a small amount of surfactant is added to a mixture of water and oil. The reduced interfacial tension makes it easier for the system to adopt shapes having a higher interfacial area. This makes it easier to form emulsions either of oil-in-water or of water-in-oil. *See* EMULSION; FOAM.

Many hydrocarbon surfactants can lower the interfacial tension of air–water at 20°C (68°F) from 72.8 mN/m to about 28 mN/m. Polysiloxane surfactants can reduce it to about 20 mN/m, and perfluoroalkyl surfactants can reduce it still further to about 15 mN/m. Similarly, hydrocarbon surfactants can reduce the interfacial tension of water–mineral oil from about 40 mN/m to about 3 mN/m. In addition to lowering interfacial tension, adsorbed surfactant at interfaces can strongly influence other properties, such as interfacial elasticity and interfacial viscosity.

Lung surfactants, including phospholipids and proteins, are an example of natural surfactants present in the human body. The lung surfactants are necessary to maintain a low interfacial tension at the alveolar air–water interface and to help the alveolar spaces change size during the breathing cycle. These interfacial tensions also change rapidly in response to changes in alveolar radius, and can fall to values as low as 10 mN/m or less. A deficiency of lung surfactant causes the interfacial tension to be too high, which causes alveolar collapse at the end of expiration. The lung surfactants may also play a role in protecting the lung from injury and infection caused by inhalation of particles and microorganisms. *See* PHOSPHOLIPID; RESPIRATORY SYSTEM.

As surfactant is added to a multiphase system, the surfactant molecules will tend to adsorb at the interface(s), forming into an oriented monolayer at the interface and reducing interfacial tension. However, above a certain concentration, called the critical micelle concentration (cmc), the surfactant molecules will start to self-associate and form aggregates called micelles. The value of the cmc depends primarily on the nature of the surfactant. Other factors being equal, higher molar masses produce lower cmc values. Once micelles have been formed, almost all additional surfactant added to the system will become solubilized in the micelles. In aqueous solution the part of the surfactant molecules that have an affinity for the nonpolar medium (the tails) associates in the interior of the micelle, while that part that has an affinity for the water (the head groups) faces the aqueous medium. Micelles typically contain 20–100 surfactant molecules each. The insides of the micelles, being filled mostly with nonpolar tails, have properties similar to an oil phase.

Surfactant types. Depending on the nature of the polar (hydrophilic) part of the molecule, surfactants are classified as anionic (negatively charged), cationic (positively charged), nonionic (noncharged), or zwitterionic (able to contain either or both kinds of charge). A common anionic surfactant is sodium dodecyl sulfate, $CH_3(CH_2)_{11}SO_4^-Na^+$. A common cationic surfactant is cetyl trimethylammonium bromide, $CH_3(CH_2)_{15}N^+(CH_3)_3Br^-$, a common type of nonionic surfactant is polyoxyethylene alcohol, $C_nH_{2n+1}(OCH_2CH_2)_mOH$. An example of a zwitterionic surfactant is dodecyl betaine, $C_{12}H_{25}N^+(CH_3)_2CH_2COO^-$. A soap is a particular kind of anionic surfactant comprising any of the surface-active fatty-acid salts containing at least eight carbon atoms. Typically, the molar masses of surfactants range from a few hundreds to several thousands of grams per mole.

Although surfactants are not generally considered to be a serious threat to humans, they can be toxic to aquatic organisms. This can be a concern due to the large quantity of surfactants used. It has been estimated that the global use of synthetic surfactants is on the order of several million metric tons per year.

Applications. Of the numerous surfactant applications, the most common are emulsifying, foaming, suspending and floating, wetting, and detergency. Emulsion, foam, and suspension products span a diverse range of industries, including agriculture, food processing, cosmetics and personal care, pharmaceuticals, papermaking and deinking, water and wastewater treatment, environmental remediation, oil recovery, and minerals beneficiation.

Emulsifying and demulsifying. An emulsion is a system of small liquid droplets dispersed in a second, immiscible liquid. Usually one of the liquids is aqueous (that is, water containing dissolved substances), while the other is an oil, usually a hydrocarbon liquid. Surfactants are the most common agents used to stabilize emulsions. The surfactants adsorb at the interface between the two liquids and act to lower interfacial tension, which makes an emulsion easier to form. Surfactants may also act to increase surface elasticity, increase electric double-layer repulsion (in the case of ionic surfactants), and increase surface viscosity, among other properties. Surface elasticity can be thought of as elasticity operating only at a surface, in two dimensions. Increased elasticity

permits a surface to better withstand deformations without rupturing. The electric double layer includes the charged surface and the oppositely charged ions that are distributed in solution but close to the surface. Increased surface electric charge causes an increased repulsive force between emulsion droplets. In addition to influencing stability, some surfactants can have an influence on the sizes of the dispersed droplets and on the overall viscosity of the emulsion. *See* EMULSION.

Although surfactants have one part of their molecule that has an affinity for the water and one part that has an affinity for the oil, as mentioned above, most surfactants have a greater overall affinity for one or the other. Surfactants that have an overall greater affinity for water (hydrophilic surfactants) tend to stabilize emulsions of oil droplets dispersed in water (oil-in-water, or O/W). Surfactants that have an overall greater affinity for oil (oleophilic surfactants) tend to stabilize emulsions of water-droplets dispersed in oil (water-in-oil, or W/O). Skin-care creams are usually formulated as O/W emulsions to promote easy spreading and adsorption while allowing a single product to contain both water- and oil-soluble components. Both water- and oil-soluble surfactants are used to stabilize the emulsions.

Surfactants can also be used to break (demulsify) already-formed emulsions, which are undesirable. For example, several kinds of emulsions are produced at wellheads during oil production operations. These emulsions have to be broken and the water and solids separated out before the crude oil can be sent to a refinery. Specialized surfactant-containing demulsifier formulations are almost always needed in the treatment of these emulsions. When crude oil is spilled at sea, a slick is formed that can turn into a particularly tenacious W/O emulsion, called a mousse emulsion. Surfactants used to treat oil spills tend to be aimed at breaking these mousse emulsions and/or at causing O/W emulsions to be formed. The O/W emulsions are desirable because they help disperse oil into the water column and away from sensitive shorelines and birds.

Foaming. A foam is a system of small gas bubbles dispersed in a liquid. Usually the liquid is aqueous while the gas is just air. Surfactants are the most common agents used to stabilize foams. The surfactants adsorb at the interface between the gas and the liquid and act to lower interfacial tension, which makes a foam easier to form. Effective foaming agents also act to increase surface elasticity, increase electric double-layer repulsion (in the case of ionic surfactants), and increase surface viscosity, among other properties. Just as for emulsions, in addition to influencing stability, the nature of the surfactant can have an influence on the sizes of the dispersed gas bubbles and on the overall viscosity of the foam. In general, ionic surfactants tend to be the best foaming agents, as are those with long, unbranched alkyl tails. *See* FOAM.

There are many examples of foams. Several kinds of foams have even been developed for use as temporary blankets. These foam blankets have been used in fire fighting (to shut off oxygen flow to the fire), over sanitary landfill sites (to suppress emissions and odors), at mine sites (to suppress dusts), at blasting sites (to suppress blast noise and pressure waves), and on petroleum products (to suppress flammable vapors), among others.

Suspending and floating. A suspension is a system of small, solid particles dispersed in a liquid. The liquid is usually aqueous. When added to a suspension, surfactants will adsorb at the interface between the solid and the liquids. For solids dispersed in water, ionic surfactants will usually stabilize the suspension by creating or increasing electric double-layer repulsion between particles. *See* SUSPENSION.

A complex example of a food suspension is ice cream, a partially frozen mixture containing many components, including proteins, fat, water, and air. The particles are ice crystals and corn-syrup solids. When the mix is whipped and partially frozen, a complex internal structure is achieved that has fat globules both adsorbed onto air bubbles and aggregated to each other, in addition to the ice crystals. The surfactants used to stabilize ice cream include casein and other proteins. In addition to being a suspension, ice cream is a foam and an emulsion.

Adsorbed surfactant may also change the wetting preference of the solid particles. If adsorbed surfactant causes finely divided particles to strongly prefer to be wetted by (in contact with) oil than water, then such particles will tend to float on water. This principle is used in the treatment of mineral ores by flotation, in which a small amount of surfactant, added during the grinding and slurrying process, alters the wettability of the ore particles so that they become hydrophobic and attach to air bubbles and float to the surface, where they are recovered by skimming. Flotation processes are used to concentrate a wide range of metallic and nonmetallic minerals, including copper, lead, zinc, nickel, calcium fluoride, barium sulfate, potash, sulfur, coal, phosphates, alumina, silicates, and clays. Flotation is also used in the removal of bacteria and particulates from water and wastewater.

Wetting. Another application in wetting is when adsorbed surfactant is used to make cloth fibers hydrophobic, which in turn makes a fabric water-repellent. Long-chain cationic surfactants are commonly used for this purpose, since the fiber surfaces are usually anionic. With the cationic heads adsorbed to the fiber surfaces, the hydrocarbon tails orient away from the surface, toward the water. As long as the weave is reasonably tight so the pore sizes between fibers are fairly small, water will enter only between the fibers under high pressure. The result is a fabric that is water-repellent but which allows air to pass through: the fabric "breathes."

In contrast, if the orientation is with the polar head away from a surface, a surfactant will make the surface hydrophilic—preferentially wetted by water. Some common wetting agents include barely water-soluble anionic hydrocarbon surfactants and some siloxane surfactants. Applications in which surfactants are used to promote water-wetting of otherwise waxlike surfaces include (1) wetting of fabrics to ensure even distribution of a textile dyes, (2) dips for treating animal hides, (3) adjuvants,

ensuring rapid wetting and spreading of insecticides and herbicides, and (4) ensuring uniform coverage when applying paints and coatings.

Detergency. Detergency involves the action of surfactants to adsorb at interfaces, alter wetting and/or interfacial tensions, and thereby reduce the energy needed to cause the removal of dirt from solid surfaces. Soaps were the first detergents, but are sensitive to acidic conditions and to the presence of hardness in the water (calcium and magnesium ions), causing soap scum. Modern detergents, such as alkyl sulfates and alkyl-aryl sulfonates, are much more effective and have almost entirely replaced soaps. These detergents comprise synthetic surfactants formulated with a number of other additives. An effective detergent formulation has to be able to induce water-wetting of the dirt (solids and/or oils) and encourage the dirt to be displaced from the surfaces to be cleaned into the wash water. Ideally, a detergent will also solubilize or otherwise prevent redeposition of the dirt. Surfactants cannot do all these things by themselves in practical situations, so commercial detergent formulations also contain additives such as builders (to enhance the detergent action), anti-redeposition agents, brighteners, and cosurfactants. *See* DETERGENT; SOAP.

Familiar applications of detergency include dish washing, clothes washing, shampooing, bathing, and hand washing. It is sometimes desirable for detergent formulations to contain surfactants that can act as germicides as well. Some cationic surfactants, like quaternary ammonium surfactants, are toxic to bacteria, fungi, and algae. An example is *N*-alkyldimethylbenzylammonium chloride. Detergency is also used in industry, for example, to clean hard surfaces like glass and metal, and soft surfaces like textiles.

Detergency can also form the basis for an enhanced oil recovery (EOR) process. In oil-containing reservoirs, the oil is distributed among pores in the rock. After conventional processes, such as waterflooding, there remain in the rock considerable amounts of oil trapped by interfacial forces. Surfactant-based EOR processes are intended to overcome these forces by altering the wetting and/or interfacial tension properties so that the oil is released from the pores in the form of emulsion droplets that can be driven toward a producing well. In the same way, surfactants can be used to flush nonaqueous phase liquid contaminants, such as gasoline or other fuels, out of underground aquifers as part of an environmental soil remediation process. *See* PETROLEUM ENHANCED RECOVERY.

Laurier L. Schram

Bibliography. D. R. Karsa (ed.), *Industrial Applications of Surfactants*, Royal Society of Chemistry, London, 1987; D. Myers, *Surfactant Science and Technology*, VCH, 1988; M. J. Rosen, *Surfactants and Interfacial Phenomena*, 2d ed., Wiley, 1989; M. J. Rosen and M. Dahanayake, *Industrial Utilization of Surfactants*, AOCS Press, 2000; L. L. Schramm (ed.), *Surfactants: Fundamentals and Applications in the Petroleum Industry*, Cambridge University Press, 2000.

Surge arrester

A protective device which is connected between an electrical conductor and the ground with the intent of limiting the magnitude of transient overvoltages on equipment.

The valve arrester consists of disks of zinc oxide material that exhibit low resistance at high voltage and high resistance at low voltage. By selecting an appropriate configuration of disk material, the arrester will conduct a low current of a few milliamperes at normal system voltage. When insulation levels of equipment are coordinated with the surge arrester protective characteristics, and under lightning or switching surge overvoltage conditions, the surge current is limited by the electric circuit; and for the magnitudes of current that can be delivered to the arrester location, the resulting voltage will be limited to controlled values, and to safe levels as well.

A typical surge arrester consists of disks of zinc oxide material sized in cross-sectional area to provide desired energy discharge capability, and in axial length proportional to the voltage capability. The disks are then placed in porcelain enclosures to provide physical support and heat removal, and sealed for isolation from contamination in the electrical environment (see **illus.**).

System overvoltages may be of either external or internal origin, that is, lightning or switching. The arrester is not intended to determine the origin of the overvoltage and must attempt to limit the magnitude of all abnormal voltages above a specific level. Hence a lightning arrester is really a voltage-surge arrester. Important characteristics of arresters are steady-state voltage capability, discharge voltage

Surge arrester unit rated 40 kV, consisting of porcelain enclosure and multiple zinc oxide disks in series. (*General Electric Co.*)

across the arrester during passage of current, and maximum current discharge capability.

The discharge voltage characteristic determines the maximum transient voltage level permitted by the arrester. It is the measure by which the protective efficiency of arresters is determined. The discharge capability is the measure of the arrester endurance to severe lightning and switching surges.

Arresters are classified as station, intermediate, or distribution-type. Station-class arresters are used for the protection of apparatus in important substations where the highest level of protection is desired. Intermediate-class arresters are used for the protection of apparatus in small- and medium-size substations where economic considerations do not justify the use of station-type arresters. Distribution-class arresters are used to protect of pole-type distribution apparatus, principally distribution transformers. *See* LIGHTNING; LIGHTNING AND SURGE PROTECTION.

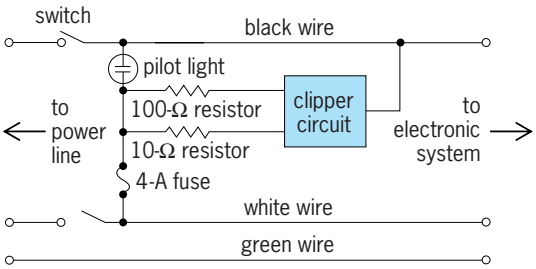
Glenn D. Breuer

Surge suppressor

A device that is designed to offer protection against voltage surges on the power line that supplies electrical energy to the sensitive components in electronic devices and systems. The device offers a limited type of protection to computers, television sets, high-fidelity equipment, and similar types of electronic systems.

A voltage surge is generally considered to be a transient wave of voltage on the power line. The amplitude of the surge may be several thousand volts, and the duration may be as short as 1 or 2 milliseconds or as long as about 100 ms. Typical effects can be damage to the electronics or loss of programs and data in computer memories. Many events can cause the surges, including lightning that strikes the power lines at a considerable distance from the home or office; necessary switching of transmission lines by the utilities; and rapid connections or disconnections of large loads, such as air conditioners and motors, from the power line, or even other appliances in the home. Lightning is perhaps the most common. *See* LIGHTNING.

Specifications. The suppressor acts to limit the peak voltage applied to the electronic device to a level that normally will not cause either damage to the device or software problems in the computers. The normal operating voltage and frequency of the power line in which the suppressor is to be used, the current rating, and the types of loads that are to be connected are usually specified, as are the types of protection that the suppressor will provide. Typical specifications include the maximum voltage that will be applied to the electronic system, the response time required (indicating how quickly the device responds), the maximum surge voltage against which the device offers protection, and the amount of energy that can be safely dissipated. For example, in countries where systems with an alternating-current voltage of 120 V and a frequency of 60 Hz are common, a typical set of specifications for a surge suppressor designed for the home or small office might include a peak clamping voltage of 250 V, a response time of 5 nanoseconds, a maximum transient protection level of 6000 V, and an energy dissipation of 40 joules. In countries having systems with an alternating-current voltage of 220 V, the clamping level would be higher, perhaps 500 V.



Typical circuit diagram for a surge suppressor; the colors are typically found on wires in North America.

Operation. The device may include a pilot light, a fuse, a clipping circuit, resistors, and a main switch (see **illus.**). The main switch serves to energize or deenergize the entire system. The pilot light is illuminated when the suppressor is operating correctly. The clipper circuit is the principal item, and the design of this portion is usually proprietary information. If a high-energy transient is absorbed by the clipper circuit, then it is likely that the fuse will be blown. Then the pilot light goes out, but the suppressor still supplies power to the electronic system. It does not, however, provide additional protection until the fuse is replaced. The unilluminated light is a signal to the operator that the suppressor has functioned as intended and that it must be checked. The minimum servicing is replacing the fuse. *See* FUSE (ELECTRICITY).

Limitations. A basic surge suppressor will not offer protection against massive, nearby lightning strokes. Lightning or surge arresters may provide such protection. It will not offer protection against sags or drops in line voltage, such as drops that cause lights to dim when, for example, an air conditioner switches on. Voltage regulators and constant voltage transformers serve this purpose. It will not reduce high-frequency electrical noise that may be on the power line. Filters are available to meet this need. *See* ELECTRIC FILTER; ELECTRONIC POWER SUPPLY; LIGHTNING AND SURGE PROTECTION; SURGE ARRESTER; TRANSFORMER; VOLTAGE REGULATOR.

Edwin C. Jones, Jr.

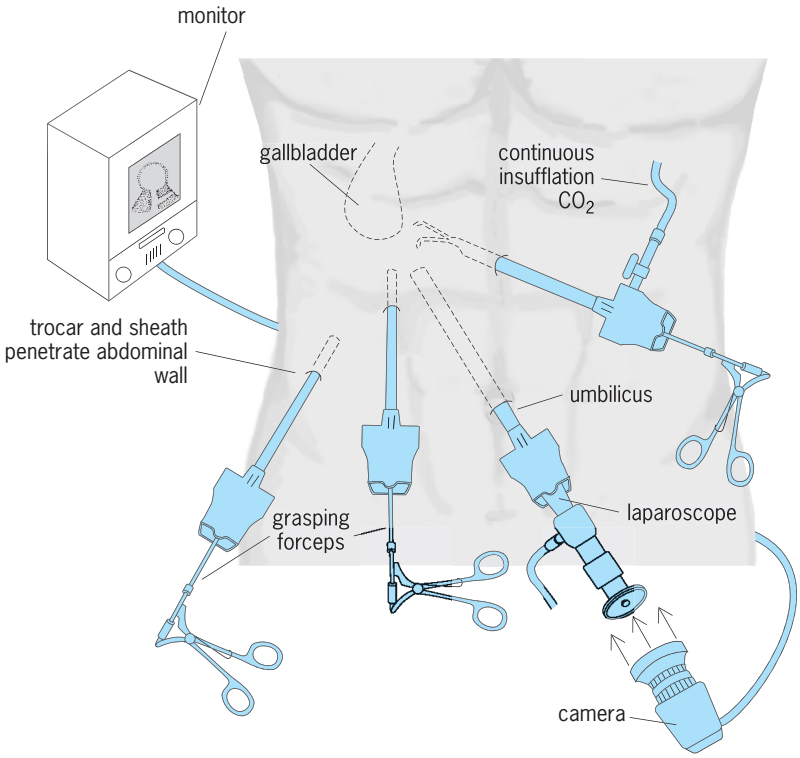
Bibliography. Institute of Electrical and Electronics Engineers, *Guidelines for Surge Voltages and Low-Voltage AC Power Circuits*, IEEE C62. 41, 1991; G. Mager, Saving your computer from surges, sags, or noise, *DEC Professional*, 3(7):26–38, December 1984; H. O. Nash, Jr., and F. M. Wells, Power systems disturbances and considerations for power conditioning, *IEEE Trans. Ind. Appl.*, IA-21(6):1472–1481, 1985; Underwriter's Laboratories, *Underwriters Standard 1449 for Surge Suppression*, rev. ed., 1995.

Surgery

General surgery includes broad operative expertise and experience with diseases of the gastrointestinal tract and abdomen; benign and malignant diseases of the breast; disorders of the endocrine system, including the thyroid, parathyroid, and adrenal glands; and vascular disease; as well as the comprehensive management of trauma. *See* BREAST DISORDERS.

Gastrointestinal surgery. This specialty encompasses the treatment of diseases of the esophagus, stomach, small intestine, large intestine, rectum, anus, pancreas, liver, and biliary tree. Traditionally, the organs of the abdomen (intraabdominal viscera) are exposed by making a major incision through the abdominal musculature to enter the peritoneal cavity, the direction and extent of the incision being individualized according to the condition being addressed. With the advent of video cameras, fiber-optic technology, and specialized instrumentation, some intraabdominal procedures are being performed through much smaller, less traumatic, and less uncomfortable incisions, resulting in lessened disability and minimal scarring. Videoscopic approaches have been successfully applied to cholecystectomy, surgery for peptic ulcer disease, surgery for gastroesophageal reflux, correction of inguinal hernia, colectomy for benign disease, and splenectomy for hematologic disease. *See* OPTICAL FIBERS.

In order to be able to visualize intraperitoneal structures with the video camera, carbon dioxide is pumped into the peritoneal cavity continuously to maintain constant distention. A series of small incisions are then made to introduce a laparoscope and video camera along with instrumentation such as retractors and dissectors. Coagulation of blood vessels by cauterization, that is, the use of a heated instrument, or laser is used regularly. Suturing also occurs through the laparoscope. The entire procedure is viewed by the surgeon on high-resolution television monitors which provide outstanding detail of intraabdominal structures. Videoscopic surgery is one of the most significant advances in general surgery since the mid-1980s (see **illus.**).



Position and function of laparoscopic instruments during laparoscopic cholecystectomy. (***After K. A. Zucker, ed., Surgical Laparoscopy, Quality Medical Publishing, 1991***)

Endoscopic procedures on the esophagus, stomach, duodenum, bile duct, and colon depend on fiber-optic technology to carry images from one end of a flexible telescope to the other, with a lens system at each end to enhance imaging. The ends of these endoscopes are flexible with control being maintained by wires in the wall of the instrument, permitting precise manipulation. Images are now usually viewed on digital television monitors; this is possible because a tiny television camera replaces the fiber-optic visualization system. A wide variety of instruments are available for passage through the endoscope to permit biopsy, excision, cauterization of bleeding points, and other major manipulations. *See* DIGESTIVE SYSTEM; MEDICAL IMAGING.

Vascular surgery. This specialty includes the treatment of diseases of the arteries, veins, and lymphatic vessels, exclusive of the heart, thoracic aorta (cardiothoracic surgery), and intracranial vessels (neurosurgery). A wide variety of blockages, aneurysms, and degenerative conditions of the vascular system are treated with a variety of techniques. Angiography is a frequently used procedure for planning an operative approach. Through direct puncture of an artery located in an extremity, catheters are passed centrally into the artery to be studied. This is followed by injection of contrast medium in order to achieve excellent visualization of the arterial system. While the procedure is performed by a radiologist, the vascular surgeon interprets and applies the information in order to develop a therapeutic plan. The vascular surgeon also uses a variety of noninvasive diagnostic procedures, including ultrasonic imaging of peripheral, carotid, and abdominal arteries and veins, as well as blood flow measurements with peripheral pressure determinations. With the improvement of noninvasive vascular techniques for the diagnosis and surveillance of vascular disease and their treatment, diagnoses are being made sooner, results of treatment are being followed systematically, failures or disease progression are being identified promptly, resulting in more timely intervention as indicated. *See* RADIOGRAPHY; ULTRASONICS.

Occlusion or blockage of the aorta or of the arteries to the legs is most often due to atherosclerosis, resulting in diminished flow with pain due to poor circulation (initially with exercise and then at rest) and possibly gangrene. Similar lesions occur in the carotid arteries, causing stroke; in the renal arteries, causing hypertension or renal failure; or in the intestinal arteries, causing chronic abdominal pain, weight loss, and gangrene. The vascular surgeon uses a variety of techniques to deal with these problems. Obstructions to the larger vessels in

the abdomen or upper leg are bypassed by using a tube graft made of synthetic material such as dacron or poly(tetrafluoroethylene). For bypass into the lower leg or to the foot, saphenous vein is preferred because of a higher long-term patency rate. For occlusive disease of the carotid arteries, endarterectomy is the preferred technique. This consists of actual removal of the atherosclerotic plaque from the opened artery, followed by restoration of blood flow to the brain. *See* ARTERIOSCLEROSIS.

Transplant surgery. Transplant surgery involves the surgical placement of an organ taken from a donor into a recipient with end stage failure of that organ. Donors are most often individuals who have been declared brain-dead; that is, there is irreversible brain damage incompatible with life. Following the consent of the donor or the family, normal organs can be removed for transplantation into a suitable recipient (cadaveric transplant). Less frequently, a healthy individual, often a close relative, serves as a donor (living related transplant). Diseases treated in this fashion include renal failure, liver failure, and insulin-dependent diabetes mellitus. Cardiac and pulmonary failure are also successfully treated by transplantation in selected cases.

Previously, early transplantation surgery was severely hindered by the recipient's aggressive immune response to the foreign organ which led to rejection of the organ and ultimate failure. The advent of a variety of effective and increasingly better-tolerated immunosuppressant medications has led to a blossoming of transplant surgery with generally excellent results in function and longevity of the transplanted organ. These agents include prednisone, azathioprine antithymocyte globulin, and especially cyclosporine. Unfortunately, these medications also impair resistance to infection, an ever-present potential problem for transplant recipients. The principal obstacle to widespread transplantation is a shortage of donors. *See* TRANSPLANTATION BIOLOGY.

Surgical oncology. This discipline deals with the treatment of the various forms of cancer amenable to surgical therapy. It encompasses all forms of gastrointestinal cancer, cancer of the breast, skin cancers such as melanoma, and many connective tissue malignancies (sarcomas) of the extremities. Traditionally, surgical oncologists removed large sections of tissues and regional lymph nodes along with the tumor. In addition, the surgical oncologist specializes in the application of multidisciplinary approaches to the management of malignancy, such as radiation therapy, chemotherapy, and nutritional support. In some instances, such as for many individuals with breast cancer, multimodality therapy has resulted in less aggressive, less destructive operations which, coupled with the use of chemotherapy and radiation therapy, offer the same or improved long-term survival benefit. More modern approaches include manipulation of the immune system and gene therapy. *See* CANCER (MEDICINE); CHEMOTHERAPY AND OTHER ANTINEOPLASTIC DRUGS; ONCOLOGY; TUMOR.

Trauma and critical care surgery. This discipline involves the postinjury and postoperative care of critically ill individuals. The goal is to prevent and treat multiorgan system failure and the other complications, as well as to address issues of nutrition, speech and communication, physical therapy, occupational therapy, wound care, and decubitus ulcer prevention, and to prescribe the appropriate rehabilitation plan for the individual.

Cardiothoracic surgery. Cardiothoracic surgery attempts to treat diseases of the heart and thorax, including the lungs, aorta, esophagus, thymus, pleura, diaphragm, and chest wall. The development of the heart-lung machine has allowed cardiac surgery to be a standard therapeutic option for many conditions, especially coronary atherosclerosis and valvular heart disease. This machine allows for cardiac and pulmonary function to be performed through the use of a pump and oxygenator, while the heart is stopped and operated upon. The most commonly performed procedure is coronary artery bypass grafting, in which saphenous veins from the leg or the internal mammary artery are used to bypass obstructed or narrowed coronary arteries, thereby increasing blood flow to the heart. This procedure is required for many despite the widespread use of balloon angioplasty to dilate narrowed segments of coronary arteries.

The application of videoscopic techniques to thoracic surgery using a miniature television camera, that is, the thoracoscope, and specialized instrumentation which can be introduced into the chest through small incisions has allowed surgeons to perform major thoracic procedures. Surgery of this type has been used for staging lung cancer, esophageal resection, limited pulmonary resection, resection of pulmonary cysts, and biopsy. The minimally invasive procedural approach markedly decreases morbidity, hospital stay, and postprocedure disability. *See* HEART DISORDERS.

Plastic surgery. The aim of plastic surgery is to repair, replace, and reconstruct defects of form and function of the skin and its underlying musculoskeletal system, with emphasis on the craniofacial structures, the oropharynx, the breast, the hand, and the external genitalia. It also includes esthetic surgery of structures with undesirable form. Plastic surgeons have extensive experience in the design and transfer of tissue flaps, in the transplantation of tissues, and in the replantation of various structures, along with skill in excisional surgery, the management of complex wounds, and the use of synthetic materials to augment the reconstructive process. The principles of plastic surgery can be applied to a wide variety of problems, such as congenital defects of the head and neck, including clefts of the lip and palate; neoplasms of the head and neck; facial trauma, including fractures; esthetic surgery of the head and neck, trunk, and extremities; and reconstructive surgery of the breast, hand, and extremities.

Pediatric surgery. This field involves surgical problems of the newborn, including premature infants, as well as children of all ages. Care of the pre- and postoperative child employs very different principles than those used for the adult. The pediatric surgeon treats congenital diseases in children, as well as a variety of illnesses and neoplasms that occur

in this population. There is special expertise in the management of congenital and acquired conditions of the gastrointestinal tract and abdominal cavity, the vascular system, the integument, the diaphragm and thorax exclulsive of the heart, the endocrine glands, the gonads and reproductive organs, and the head and neck, as well as in the comprehensive management of trauma.

Urology. Urology manages the treatment of the surgical diseases of the genitourinary tract. Infections, stone disease, organ dysfunction, and neoplasms of kidneys, ureters, bladder, prostate, urethra, and external genitalia all fall within the purview of the urologist. Techniques include the use of endoscopy, extracorporeal shock wave lithotripsy (ultrasonic fragmentation of kidney stones), videoscopic staging of neoplasms, and an emphasis on multidisciplinary management of neoplastic disease.

Orthopedic surgery. The surgical management of diseases of the musculoskeletal system, orthopedic surgery, includes congenital, neoplastic, traumatic, and degenerative diseases of long and short bones, the spine, the hand, the joints, and the associated musculature and tendons. The orthopedic surgeon is familiar with all aspects of adult orthopedics, pediatric orthopedics, fractures and trauma, surgery of the spine including disc surgery, hand and foot surgery, athletic injuries, and rehabilitation. There has been continued improvement in joint replacement surgery through the introduction of prostheses much more resistant to deterioration and significantly improved rehabilitation programs. Endoscopic instrumentation has been used for arthroscopic surgery, whereby miniaturized cameras and instruments are introduced into joints for diagnosis and therapy. Another advancement in orthopedics has been the development of distraction osteogenesis, that is, the formation of new bone to replace that which has been destroyed by infection to lengthen limbs in selected individuals. *See* PROSTHESIS.

Otolaryngology. Otolaryngology specializes in the treatment of diseases of the head and neck, the ears, nose, and vocal cords, the upper alimentary tract, and the upper respiratory tract. It also deals with many communicative disorders, especially those affecting hearing and voice production. The otolaryngologist is experienced in temporal bone surgery; paranasal sinus and nasal septum surgery; maxillofacial, plastic, and reconstructive surgery of the head and neck; surgery of the salivary glands; head and neck oncologic surgery; peroral endoscopy, both diagnostic and therapeutic; surgery of lymphatic tissues of the pharynx; and recognition and medical management of diseases and abnormal function of the ears, upper and lower respiratory tract, and upper alimentary tract. Treatments include videoscopic approaches to skull base tumors and microsurgery of the inner ear. Multimodality management of many head and neck cancers has also resulted in improved survival.

Neurological surgery. This discipline deals with the diagnosis, evaluation, and treatment of disorders of the central, peripheral, and autonomic nervous systems, including their supporting structures and vascular supply. Neurological surgery includes the operative and nonoperative management, diagnosis, evaluation, treatment, critical care, and rehabilitation of individuals with disorders of the nervous system and cerebrovascular disease, including aneurysms, trauma, and a wide range of tumors.

Minimally invasive approaches are being applied to a greater extent. Advances have been made in surgery for epilepsy, which has been unresponsive to medical treatment, through the precise excision of abnormal foci of electrical discharge in the brain. The gamma knife has been introduced to treat arterial-venous malformations and tumors located in parts of the brain inaccessible to direct surgery. The gamma knife consists of 202 individually focused beams of radiation which can be directed at such lesions with a very high degree of precision, sparing the surrounding normal brain almost completely. *See* MEDICINE.

Jeremy Steinbaum; Anthony L. Imbembo

Bibliography. R. W. Bailey et al., Laparoscopic cholecystectomy: Experience with 375 consecutive patients, *Ann. Surg.*, 214:531–541, 1991; M. J. Krasna and M. J. Mack, *Atlas of Thoracoscopic Surgery*, 1994; D. Sabiston (ed.), *Textbook of Surgery: The Biological Basis of Modern Surgical Practice*, 15th ed., 1997; S. I. Schwartz, G. T. Shires, and F. C. Spencer (eds.), *Principles of Surgery*, 6th ed., 1993; K. A. Zucker (ed.), *Surgical Laparoscopy Update*, 1993.

Surveillance radar

Any radar equipment used by civil or military authorities for locating aircraft, ships, or ground vehicles; most commonly, in civil air-traffic management, a radar used by controllers to indicate the position of aircraft aloft or on the airport surface. Two types of air-traffic control surveillance radar are employed internationally. Ground-based conventional, or primary, radar operates by transmitting microwave pulses and detecting the resulting energy reflected from the aircraft body (airframe). Secondary surveillance radar employs cooperative radio receiver-transmitter units (transponders) on the aircraft.

Primary radar. Primary radar was developed as an outgrowth of military surveillance efforts against hostile aircraft and ships in World War II. It can locate and track most aircraft without the need for special airborne equipment. Primary radar requires high-powered transmitters, capable of hundreds of kilowatts of peak power. It is sometimes limited in sensitivity by reflections (clutter) from terrain, ground objects, birds, and precipitation. However, clutter can be reduced by Doppler processing that discriminates between targets of different velocities and by adaptive clutter maps. *See* DOPPLER RADAR.

Secondary radar. Secondary radar, which evolved from military friend-or-foe identification techniques, provides several advantages. Sensitivity is improved

Characteristics of air-traffic surveillance radars in the United States

Characteristic	ASR-9	ARSR-4
Frequency, MHz	2800	1300
Primary reflector size, ft (m)	20 × 12 (6 × 3.7)	42 × 32 (13 × 10)
Secondary reflector size, ft (m)	28 × 4 (8.5 × 1.2)	42 × 32 (13 × 10)
Rotational rate, revolutions/min	12	5
Typical range, mi (km)	60 (100)	250 (400)

by the use of a transponder on the aircraft because the ground and airborne transmitters need only overcome the one-way path loss between the radar and the aircraft. (Radar transmitted power diminishes as the square of the range to the aircraft. In primary radar, the diminished power is reflected by the aircraft and is further diminished on the return trip to the radar for a total diminishment by the fourth power. In secondary radar, the power incident on the aircraft triggers a transponder, which then emits a fresh signal for the return trip, so that the power is only diminished as the square rather than the fourth power of the aircraft range.) Clutter is eliminated by the use of different frequencies for ground and airborne transmissions, 1030 MHz for interrogation and 1090 MHz for reply. The transponder antenna also makes sensitivity independent of the aircraft's size (radar cross section). In addition to sensitivity advantages, the transponder allows ground radar and aircraft to exchange coded messages.

The secondary surveillance radar system in current international use provides 4096 pilot-selected codes for limited aircraft identification and for altitude reporting in 100-ft (30-m) increments. An advanced international secondary radar standard operates on the same frequencies and provides 16 million codes to allow unique identifications for every aircraft in the world. This system, designated Mode S for selective address, also allows two-way exchange of data messages.

Mode S has been designed to be compatible with current ground and airborne equipment. Mode S ground sensors develop surveillance data for both current and Mode S airborne transponders using monopulse azimuth estimation techniques. The monopulse approach requires only four to six interrogations for each current transponder and one interrogation for each Mode S transponder; current ground interrogators use 20–30 interrogations. The time saved is used to selectively interrogate Mode S transponders and to conduct Mode S data-link transactions. *See* MONOPULSE RADAR.

Mode S transponders are designed to respond as a current transponder when interrogated by a conventional ground interrogator and to respond with Mode S replies when interrogated by either a ground Mode S sensor or an aircraft equipped with a Traffic Alert and Collision Avoidance System (TCAS). Thus there is complete compatibility between current ground and air equipment and Mode S. *See* AIRCRAFT COLLISION AVOIDANCE SYSTEM.

Monopulse azimuth estimation operates by simultaneously measuring the transponder reply amplitudes on sum and difference patterns of the ground antenna to determine the azimuth of the reply. The reply azimuth estimates are used to provide more accurate azimuth data (1 milliradian root-mean-square error) and assist in degarbling replies from current transponders to improve the reliability of the 4096-code and altitude data presented to the air-traffic controller.

Secondary radar surveillance is limited in that it operates only with transponder-equipped aircraft. All aircraft flying in United States controlled airspace above 10,000 ft (3048 m) and near major airports are equipped with transponders. However, 18% of general-aviation aircraft in the United States are not transponder-equipped. Another limitation of secondary radar results from shielding of the airborne antenna by the airframe in certain flight attitudes. Shielding sometimes causes temporary loss of coverage during takeoff or in steep banks. This problem can be overcome by using two aircraft antennas, with automatic antenna selection based on a comparison of the power received at the two antennas. All aircraft of large airlines in the United States carry Mode S transponders with diversity antennas.

Use. Although secondary radar generally provides more reliable and complete surveillance information than does primary radar, air-traffic authorities use both. In the United States, two major examples are the Federal Aviation Administration's Airport Surveillance Radar (ASR) and Air Route Surveillance Radar (ARSR), the latter shared at some locations with the U.S. Air Force (see **table**). Both normally operate unattended, their outputs fed by a landline or microwave link to their control positions. Primary radar provides backup coverage for nonequipped aircraft and for aircraft with airframe shielding. Primary radar can also detect potentially hazardous precipitation cells and potentially hostile aircraft or inadvertent intruders. Secondary radar is sometimes used without a primary radar backup for long-range, high-altitude surveillance where all aircraft are transponder-equipped, where airframe shielding is rare, and where the sensitivity disadvantage of primary radar requires costly high-power transmitters. The Federal Aviation Administration has authorized use of certain secondary radars having faster update rates and monopulse accuracy for monitoring aircraft conducting multiple parallel approaches in bad weather. *See* AIR-TRAFFIC CONTROL; RADAR.

Raymond R. LaFrey

Bibliography. J. L. Baker et al., Mode S system design and architecture, *Proc. IEEE*, 77(11):1684–1694, 1989; V. A. Orlando, The Mode S beacon radar system, *Lincoln Lab. J.*, 2(3):345–362, 1989; M. C. Stevens, *Secondary Surveillance Radar*, 1988.

Surveying

The measurement of dimensional relationships among points, lines, and physical features on or near the Earth's surface. Basically, surveying determines horizontal distances, elevation differences, directions, and angles. These basic determinations are applied further to the computation of areas and volumes and to the establishment of locations with respect to some coordinate system.

Surveying is typically used to locate and measure property lines; to lay out buildings, bridges, channels, highways, sewers, and pipelines for construction; to locate stations for launching and tracking satellites; and to obtain topographic information for mapping and charting.

Horizontal distances are usually assumed to be parallel to a common plane. Each measurement has both length and direction. Length is expressed in feet or in meters. Direction is expressed as a bearing of the azimuthal angle relationship to a reference meridian, which is the north-south direction. It can be the true meridian, a grid meridian, or some other assumed meridian. The degree-minute-second system of angular expression is standard in the United States.

Reference, or control, is a concept that applies to the positions of lines as well as to their directions. In its simplest form, the position control is an identifiable or understood point of origin for the lines of a survey. Conveniently, most coordinate systems have the origin placed west and south of the area to be surveyed so that all coordinates are positive and in the north-east quadrant.

Coordinate systems may be assumed, or established coordinate systems may be used. The most widely used systems in the United States are the various state plane coordinate systems established by the U.S. Coast and Geodetic Survey (now the National Geodetic Survey), which approximate in each case a small portion of the Earth's spherical form.

Vertical measurement adds the third dimension to an object's position. This dimension is expressed as the distance above some reference surface, usually mean sea level, called a datum. Mean sea level is determined by averaging high and low tides during a lunar month, but for certainty this must be carried out for as long as 19+ years.

Precision. Surveying devices and procedures possess individual limitations for accuracy of measurement. The term precision expresses the notion of degree of accuracy.

The choice of methods depends on the precision required for the result. Surveys of major installations or primary control surveys demand higher orders of precision than, for example, a preliminary highway profile or some lesser layout.

Instrumental surveys are designated first-order, second-order, and third-order. First-order is maximum precision, the others gradually less. Established criteria exist to differentiate the levels of precision.

Horizontal control. The main framework, or control, of a survey is laid out by traverse, triangulation, or trilateration. Some success has been achieved in locating control points from Doppler measurements of passing satellites, from aerial phototriangulation, from satellites photographed against a star background, and from inertial guidance systems.

In traverse, adopted for most ordinary surveying, a line or series of lines is established by directly measuring lengths and angles. In triangulation, used mainly for large areas, angles are again directly measured, but distances are computed trigonometrically. This necessitates triangular patterns of lines connecting intervisible points and starting from a baseline of known length. New baselines are measured at intervals. Trigonometric methods are also used in trilateration, but lengths, rather than angles, are measured. The development of electronic distance measurement (EDM) instruments has brought trilateration into significant use.

Traversing. Traverse lines are usually laid out as a closed polygon. This makes it possible to check the accuracy of the field measurements by calculating how nearly the figure closes mathematically. Commonly there are unavoidable random or accidental errors in angular and in distance measurements that tend to coalesce and militate against exact mathematical closure. Angular error is checked by geometric theory: the sum of all interior angles in any polygon equals the number of angles, less two, times 180°. The distance misclosure is determined by resolving the field-measured distance for each line into its north-south component (its latitude) and its east-west component (its departure). For the figure to close, components bearing north must equal in total length those bearing south, and components bearing east must equal in total length those bearing west. The net misclosure in the sum of latitudes and the net misclosure in the sum of departures are the two components of the linear error of closure of the traverse. If the total error of closure is within the prescribed error limit for the order of precision chosen for the survey, the total error is apportioned among the several angles and lengths of the traverse.

Triangulation. This begins with the selection of points whose connecting sight lines form one triangle or a series of triangles. In a series, each triangle has at least one side common to each adjacent triangle. An initial side length, the baseline, is measured, as are all the angles. By successive application of the law of sines, each successive side length is computed. The directions (azimuths) of the sides are also carried forward, so that the relative positions of all points can be determined.

Trilateration. In trilateration the figures are selected as for triangulation. An initial direction is determined and all side lengths are measured. Interior angles are computed by oblique-triangle formulas to obtain geometric checks on distances and to establish triangle-side directions. Ensuing latitude and departure computations yield the relative positions of the angle points of the figures.

Distance measurement. Traverse distances are usually measured with a tape or by EDM, but also may sometimes be measured by stadia, subtense, or trig-traverse.

Taping. The surveyor's tape, a steel ribbon graduated throughout its length, is usually 100 ft (30.48 m) long. However, 200-ft (60.90-m) and 300-ft (91.44-m) tapes are common.

Whether on sloping or level ground, it is horizontal distances that must be measured. Horizontal components of hillside distances are measured by raising the downhill end of the tape to the level of the uphill end. On steep ground this technique is used with shorter sections of the tape. The raised end is positioned over the ground point with the aid of a plumb bob.

Where slope distances are taped along the ground, the slope angle can be measured with the clinometer. The desired horizontal distance can then be computed. Precision in tape measuring is increased by refinements such as standardization, sag correction, tension control, and thermal-expansion correction.

Electronic distance measurement. In EDM the time a signal requires to travel from an emitter to a receiver or reflector and back to the sender is converted to a distance readout. It utilizes the precisely known speed of radio waves or of light, as well as a crystal-controlled chopped or sinusoidal variation in the emitted signal to measure distance in terms of number of wavelengths and a fractional part thereof. Microwave EDM requires a master instrument and a slave instrument, but has a capability of measuring up to 20 mi (32 km) or more. Shorter distances, from a few feet to about 2 mi (3.2 km), are best measured with ordinary light, laser light, or infrared EDM equipment. These require only a single instrument for emitting the signal and a simple retro-reflector or a reflector prism (corner cube) for returning the signal.

The great advantage of electronic distance measuring is its unprecedented precision, speed, and convenience. Further, if mounted directly onto a theodolite, and especially if incorporated into it and electronically coupled to it, the EDM with an internal computer can in seconds measure distance (even slope distance) and direction, then compute the coordinates of the sighted point with all the accuracy required for high-order surveying. The development of EDM is a great breakthrough, changing the nature and methods of a great part of modern surveying. Accuracies of 1 part in 10^6 are now assuredly within reach (**Fig. 1**).

Stadia. The stadia technique also requires no tape, although it is rather an imprecise distance-measuring method. A graduated stadia rod is held upright on a point and sighted through a transit telescope set up over another point. The distance between the two points is determined from the length of rod intercepted between two horizontal wires in the telescope.

Subtense. In the subtense technique the transit angle subtended by a horizontal bar of fixed length (usually 2.00000 m, approximately equivalent to 6.56168 ft) enables computation of the transit-to-bar distance (**Fig. 2**). With the 2-m (6.6-ft) bar and a theodolite to measure the angle (α) to 1 second, precise distance (D) can be calculated up to 200 m (660 ft) by the formula $D = 1.00000 \text{ cotan } {}^1/_2\alpha$ m. The subtense method can replace taping across a busy highway, canal, or ravine, although it is slower than EDM. For short to medium distances it is more precise than stadia.

Fig. 1. Instrument combines EDM and transit. (*Keuffel & Esser Co.*)

Trig-traverse. In trig-traverse the subtense bar is replaced by a measured baseline extending at a right angle from the survey line whose distance is desired. The baseline is at least long enough relative to the desired distance to assure the order of precision desired. The distance calculated in either subtense of trig-traverse is automatically the horizontal distance and so needs no correction because of difference of

Fig. 2. Subtense bar. (*Lockwood, Kessler, and Bartlett Inc.*)

elevation between theodolite and baseline or subtense bar.

Angular measurement. The most common instrument for measuring angles is the transit or theodolite. It is essentially a telescope that can be rotated a measurable amount about a vertical axis and a horizontal axis. Carefully graduated metal or glass circles concentric with each axis are used to measure the angles. Glass circles permit much greater magnification and more precise readings because concentrated light is passed through the portions of the circle being read. Further, superimposing the two opposite sides of the circle eliminates eccentricity error. The transit is centered over a point with the aid of either a plumb bob suspended by a string from the vertical axis or (on some theodolites) an optical plummet, which enables the operator to sight along the instrument's vertical axis to the ground through a right-angle prism.

Horizontal angles. To measure a horizontal angle between two intersecting lines, the transit is set up over the intersection. The telescope is sighted along one of the lines, the graduated circle is clamped against rotation, and the telescope is rotated to sight along the other line. The angle is indicated on the horizontal circle by another concentric circular plate, inscribed with an index and a vernier, that rotates with the telescope. Glass-circle transits and theodolites use optical micrometers instead of verniers, giving a much more precise reading. In either case, the angle can be read directly if the initial reading has been set at zero. Otherwise, the angle is found as the difference between the initial and final readings.

To lay off a predetermined angle from some reference line, the initial sight is taken along the line, the telescope is rotated through the angle desired, and a stake or other marker is set on the new line.

The special case of laying off a 180° angle is simply the extension of a straight line. It is done by backsighting along the reference line and rotating the telescope about its horizontal or elevation axis (transiting) for sighting ahead.

Vertical angles. These are measured by rotating the telescope about its elevation axis and reading the angle on the vertical circle. Vertical angles are measured from a hortizontal reference usually, though sometimes from the zenith or from the nadir.

Angular precision. Transit precision is denoted by the smallest angular distinction or resolution of which the instrument is capable. A 1-minute transit is so called because its circle is graduated to half-degrees, with the vernier measuring a thirtieth of the graduation, to enable an angular reading to 1 minute. Other transits read to 30 or 20 seconds. Theodolites with glass circles usually give readings to 0.1 minute (6 seconds); higher-order theodolites read to 1 second, even to a few hundredths of a second.

Elevation differences. Elevations may be measured trigonometrically in conjunction with reduction of slope measurements to horizontal distances, but the resulting elevation differences are of low precision.

Differential leveling. Most third-order and all second- and first-order measurements are made of differential leveling, wherein a horizontal line of sight of known elevation is sighted on a graduated rod held vertically on the point being checked. The transit telescope, leveled, may establish the sight line, but more often a specialized leveling instrument is used. For approximate results a hand level may be used.

In differential leveling, the rod is held on a bench mark, a point of known or assumed elevation. The level is set up and sighted for a reading on the rod. This reading, called the backsight or plus sight, is added to the bench-mark elevation to establish the height of the instrument. With the level remaining where it is, the rod is moved to a forward point (turning point). The reading to that point, called the foresight or minus sight, is subtracted from the height of the instrument to yield the elevation of the turning point. Distances to foresight and backsight, in pairs, are kept about equal so as to balance small instrumental errors and effects of Earth curvature and refraction of the line of sight. These are shown in exaggerated form in **Fig. 3**.

The foregoing process is duplicated successively as necessary to obtain the desired new elevation or elevations. Verification of the work is obtained by closing back on the original point by a rerun or by closing on another point of known elevation. Total error, if within allowable limits, is distributed along the level survey.

Reciprocal leveling, a variant of straight differential leveling, is used where the level cannot be set up for a long distance, as in crossing a ravine or a river. Here a short backsight and a long foresight (across the inaccessible distance) to a turning point on the opposite side are taken. Then the level is moved to the opposite side so that a long backsight is made to the original point of known elevation, and a short foresight is taken to the turning point. Since any errors of curvature and refraction in the long forward direction are canceled by those of the long backward sight, the average of the two elevation differences so obtained will be the actual difference.

Trigonometric leveling. Elevation difference can be determined also by trigonometric means, as in measuring the height of a mast, or even the elevation difference between two survey points. If the horizontal distance and the vertical angle are measured, the height of the sighted object above the

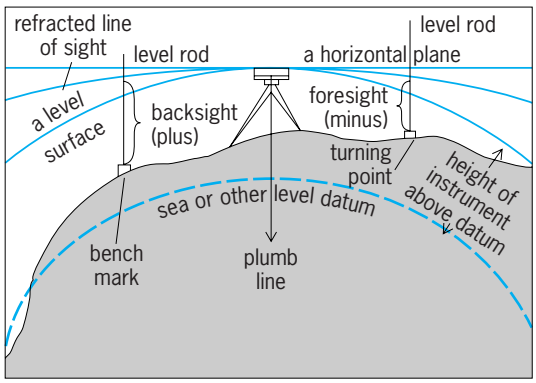


Fig. 3. Theory of differential leveling.

instrument can be calculated by right-triangle methods. In making long sights in this circumstance, a correction may be required for curvature and atmospheric refraction. An altitude observation (elevation angle) on a star or on the Sun will require refraction and parallax corrections.

Barometric leveling. Approximate elevation differences may be measured with the aid of a barometer. This method is particularly useful for reconnaissance of a substantial area.

Airborne profiling. A radar altimeter can give quite accurate elevation of an aircraft above the surface (of the ocean, for instance) by bouncing a signal from the surface back to the aircraft. Using it to give the altitude by bouncing a signal from any surface whose elevation is known can give the exact altitude as well. Thus by continuously recording a terrain profile on an overflight, the correct ground elevations can be obtained if a sufficient number of known ground elevations are scanned to provide control information. The technique is known as airborne profiling and renders fairly accurate results, adequate for a number of less demanding purposes.

Astronomical observations. To determine meridian direction and geographic latitude, observations are made by a theodolite or transit on Polaris, the Sun, or other stars. Direction of the meridian (geographic north-south line) is needed for direction control purposes; latitude is needed where maps and other sources are insufficient.

Meridian determination. The simplest meridian determination is made by sighting Polaris at its elongation, as the star is rounding the easterly or westerly extremity of its apparent orbit. At these times Polaris appears to be moving up or down the transit's cross hair, and there is ample time for assuring an accurate sighting. An angular correction is applied to the direction of sighting, which is referenced to a line on the ground. The correction value is found in an ephemeris. *See* EPHEMERIS.

The ephemeris also gives the noon declination of the Sun at Greenwich throughout the year. Declination is the angle of inclination of the Sun with respect to the equatorial plane. For direct solar observation the direction from the observer to the Sun is found from the equation below, where z is the

$$\cos z = \frac{\sin D}{\cos h \cos L} - \tan h \ \tan L$$

azimuth angle east or west from the meridian to the Sun; D is the declination at the instant of observation; h is the vertical angle to the Sun, corrected for refraction and parallax; and L is the observer's latitude.

A nonastronomic device for meridian determination is the portable gyroscope azimuth instrument, a newer development based on missile and aircraft inertial guidance systems.

Latitude determination. Latitude of a position may be determined directly at night by vertical-angle sighting on Polaris at upper or lower culmination (the northerly or southerly extremity of the star's apparent orbit) and application of the suitable vertical-angle correction from the ephemeris. In another method, the vertical angle to the Sun is observed at noon, and the Sun's declination is subtracted algebraically to yield a net angle, which is subtracted from 90° to give the latitude.

Longitude and time observations are made occassionally by noting the passage of the Sun across an established meridian. If Greenwich time is known, longitude can be computed.

Geoid. Astronomic observations from any station yield geographic latitude and longitude for that station; a great number of these stations connected by a network of triangles form a control network. An ellipsoid chosen most nearly to fit the geoidal shape of the Earth affords a base to calculate the interrelationship of the astronomic stations. Lack of conformity between measured distances and calculated distances between stations indicates disparity between astronomical and geoidal (geodetic) latitudes and longitudes, which in turn is traceable to deflection of the astronomic vertical from the ellipsoid vertical at the station or stations. These deflections of the vertical are due to undulations of the geoidal surface and its lack of conformity to the ellipsoid. Adjustments over a wide area (a continent) are usually made. The initial choice of a best ellipsoid renders the conformity closest, so that geodetic (ellipsoidal) computations will best fit the true geoid.

Several ellipsoids have been in use for different continents, but massive accumulations of measurements of gravity and deflections of the vertical through the space satellite program furnish the basis for an ellipsoid that more closely fits the geoid worldwide. It is the World Geodetic System Ellipsoid, the first of its kind. It permits the major geodetic networks of the world to be unified and the coordinates of points anywhere on Earth to be compatible. *See* GEODESY.

Types of surveys. These include geodetic control surveys, route surveys, construction surveys, and cadastral (property) surveys. *See* TOPOGRAPHIC SURVEYING AND MAPPING.

Geodetic control surveys. Control surveys provide the reference framework for lesser surveys. A traverse, with elevations of its points, may be the control for mapping a limited area. The broadest control surveys are the National Geodetic Survey (NGS) in the United States and by corresponding agencies in other countries, wherein the horizontal and vertical positions of points are established with first- or second-order accuracy.

In geodetic surveys, Earth curvature is taken into account. Coordinate positions are established in terms of latitude and longitude. Geodetic coordinates are convertible into state plane coordinates.

The traditional method of extending horizontal geodetic control is triangulation. Trilateration and traverse are supplementary techniques.

In geodetic surveys, refinements are added to the simple chain-of-triangles scheme. The most notable refinement is use of the quadrilateral as a geometric unit (**Fig. 4**). Diagonal directions as well as the sides are observed, so that the quadrilateral includes two

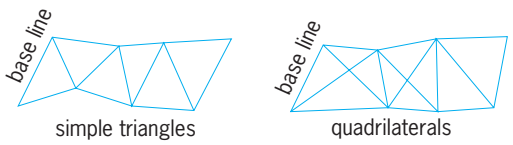


Fig. 4. Triangulation figures, used in geodetic surveys.

pairs of overlapping triangles, one being redundant although valuable as a check. In this way additional angular checks are available, and four separate calculations can be made for the entering side length of the next quadrilateral in the chain. The shapes of the triangles are important, small angles being avoided to preserve high strength of figure. A strength of figure analysis for the four different computational routes reveals the one that will best carry the length forward.

For precision and economy, geodetic triangulation stations are located as far apart as possible, consistent with the requirement of intervisibility. On the other hand, a precise control network for a builtup area may require numerous close-spaced auxiliary stations within quadrilateral figures.

Angles are measured by a precise theodolite, all angles around the horizon being observed several separate times for determination of the statistically best value of each angle. Long sightings are usually made on lighted station signals at night.

Because of the Earth's spheroidal shape, the interior angles of large triangles add up to more than 180°. A 75-mi^2 (200-km^2) triangle, for example, has 1 second more than the 180° total of interior angles. The extra second is called spherical excess.

Large triangulation chains of a primary network will have widely spaced ground stations of calculated position, which are of value as control to only a limited ground area. Thus, secondary triangulation networks are run to connect points on the primary. Then, to densify the control and make it available virtually everywhere, tertiary networks of triangulation or traverse are employed. Precise control for serving a city requires a very heavy saturation of auxiliary stations within the smallest triangulation figures.

Over-ocean geodetic control extension, formerly infeasible, now is accomplished by systems of simultaneous electronic distance measurements from the ground principles. Doppler observations of passing satellites have proved an astounding advance, enabling the location of ground stations to within less than a meter virtually anywhere on the Earth's surface. *See* SATELLITE NAVIGATION SYSTEMS.

In another method, simultaneously photographing a satellite against a star background from three camera stations (one, say, on a island) affords measurements to calculate the island station's position, again conquering the overwater problem. Offshore drilling platforms are located by one or several of these now feasible methods.

Each precisely determined distance between geodetic control stations is reduced to its sea-level projection, that is, the distance between the stations if projected vertically to sea level.

Geodetic data convert to state plane coordinate data with the aid of tables published by the NGS. Each plane coordinate system is expressed as a projection of the curved surface onto a plane intersecting either two standard parallels (latitudes) or two meridians. In both cases the meridian and parallel projections are at right angles to each other. At the standard meridians or parallels, scale is exact; elsewhere a known variable scale factor must be applied.

Advantages of state plane coordinate surveys include the ability to initiate a long route survey several places at once, and to interrelate survey points anywhere in the system. Proposed land data systems or a national cadastre would depend on the coordinate systems.

The vertical control system of the NGS consists of a first-order level network with supplementary second-order lines. In most parts of the United States, bench-mark elevations are available within a few miles of any point.

The geodetic leveling procedure is a refinement of that previously described, using more precise instruments and methods. The level instrument is built to rigid specifications. Rod graduations are on an Invar strip (Invar varies only negligibly under temperature change). Special care is taken to equalize foresight and backsight distances, to assure stable turning points, and to protect the instrument from minor stresses.

Route surveys. Surveys for the design and construction of linear works, such as roads, canals, pipelines, or railways, are called route surveys. They begin with reconnaissance and continue through preliminary, location, and construction surveys.

Reconnaissance for a new highway, for example, may be accomplished by study of existing maps together with a visual appraisal of field conditions, or even quick low-order horizontal and vertical field measurements. Controlling points, such as favorable ridge and river crossings, are found, and a preliminary line is selected. This line traditionally is laid out by transit and tape or EDM, being a linear traverse. It is profiled, that is, levels are run along the traverse line to find elevations. Transverse profiles (cross sections) are made at needed intervals. Structures and natural objects that would affect the final location are fixed by side shots. A side shot may be a direction and distance from a transit point, the intersection of two directions or two distances, or a perpendicular offset from a traverse line. Transverse profiles are made by hand level, tape, and level rod.

The result of the preliminary survey is a strip topographic map of sufficient precision to permit preliminary design of the final location, including approximate determination of earthwork quantities. More and more preliminary maps and designs are executed from topographic maps constructed by airborne photographic (photogrammetric) methods, with need for only limited ground surveying prior to construction layout.

The location survey line is conducted with at least third-order precision. Traverse procedures are followed, but curves are laid out and stationed. The

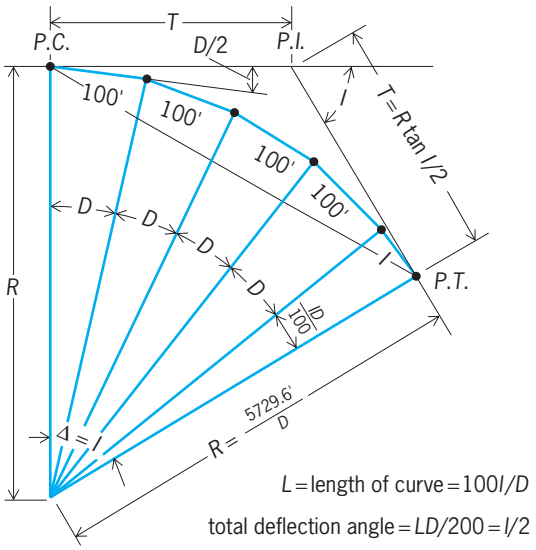


Fig. 5. Circular curve theory.

result is a staked centerline for the route to be constructed. Profile leveling and cross-sectioning for earthwork quantities also are part of the location survey.

Contract plans normally are based on the location survey. The construction stakeout is usually conducted later, but sometimes is carried out at the same time as the location survey. A prerequisite is final grade selection so that cuts and fills can be marked on centerline stakes and slope stakes can be set. The slope stake indicates the lateral limits of cut or fill at a given cross section. It is marked with the distance, plus or minus, from existing ground level to the proposed centerline grade. Being at the edge of earthwork, the slope stake supposedly is available for reference throughout construction.

A horizontal curve provides for change in direction for traffic. The directions, called tangents, connect at an angle point, called the point of intersection (P.I.). The intersection angle I is the deflection angle, right or left, from the forward projection of the tangent into the angle point (**Fig. 5**).

Any circular curve or arc, connecting two intersecting tangents, has a central angle, equal to the angle I formed by perpendiculars (radii) to the two tangents, one at the point of curvature (P.C.) and the other at the point of tangency (P.T.). The greater the radius, the more gradual the curve. The sharpness of the curve also is expressed as degree of curvature D, the central-angle increment that subtends 100 ft (30.48 m) of arc (highway definition) or 100 ft of chord (railroad definition). A 1° curve has a radius of about 5729.6 ft (1746.4 m) under either definition.

Circular curves are laid out by the following procedure, assuming D, and consequently radius R, have been chosen:

1. The tangent distance is measured back from the P.I. to establish the P.C.

2. The transit is set up at P.C., and a series of deflection angles, each equal to $D/2$, is turned in the direction of curvature to establish a point on the curve for each 100 ft (30.48 m) of chord or arc. For chords or arcs shorter than 100 ft the turned deflections angle is proportionally smaller.

With the advent of EDM equipment, alternatives have evolved for laying out points on a horizontal circular curve. The transit with an adjunct EDM device can be mounted at the P.C. or the P.T. (or indeed any control point of the survey), and distances and directions can be used directly to set curve stakes. The need for convenient and easy taping is no longer a constraint as long as sight distances are clear.

Vertical curves, connecting different grades, usually are parabolic. The parabola has an inherent transition, and it is easy to lay out with tape, level, and level rod after computation of ordinates for vertical-curve grade points along the centerline.

The traditional route survey procedures described above are economical for narrow routes in moderate terrain. Major dual highways with broad rights of way are located more efficiently on large-scale aerial topographic maps. On such maps, "paper" location surveys can be made, complete with earthwork quantity takeoff precise enough for contract bidding. Furthermore, digitized terrain data, directly available from some photogrammetric procedures, permit rapid electronic computation of quantities. With this approach it may be feasible to deter all field staking until construction time, the staking points being selected in advance by scaling or computation of map-coordinate relationships with points established during the small amount of ground survey required to control the photogrammetric survey. *See* PHOTOGRAMMETRY.

Construction surveys. Surveys for construction layout establish systems of reference points that are not likely to be disturbed by the work. Slope stakes are earthwork references. Buildings, bridge abutments, sewers, and many other structures traditionally are controlled by batter boards, horizontal boards fastened to two uprights. The top of a batter board is set at the elevation of the line to be established, and the horizontal position of the line is indicated by a mark or nail. Across the site another batter board is set up for the given line, and a string or wire stretched between is the line. The line may be a building face at first-floor level, or it may be a reference line. In trench work the line between batter boards may run at a fixed distance above the invert centerline.

In lieu of strings or wires, lines of sight may be used. A transit is set up on lines outside the work area, and points of the line are sighted as required. Further, a low-power laser, available as an attachment to the transit or as a separate special purpose instrument, can project a visible reference line or plane from an unattended position. A means of locating critical construction points, such as those for anchor bolts, is to compute their positions in a coordinate grid, then compute directions and distances from a reference point for their location by transit sighting and tape measurement.

Elevation controls are provided by bench marks near the construction area. The foregoing

construction techniques are adaptable into industrial plants for building large mock-ups and jigs as well as for the alignment of parts. Transits and levels on stable mounts may be used; however, related specialized equipment called optical tooling instruments are more readily applied. These include a transit or level telescope with an optical micrometer on the objective, to move the line of sight to a locus parallel with the initial sighting and flat self-reflecting mirrors for use as targets. Angular sight lines are established by a mirror mounted vertically on a transit base.

Underground surveys. Mine and tunnel surveys impose a few modifications on normal surveying techniques: Repeated independent measurements are made because normal checks (such as closed traverses) are not available; cross hairs illuminated because work is performed in relative darkness; vertical tape measurements and trigonometric levels, instead of differential levels, frequently must be relied upon; and in adits and tunnels survey points are placed overhead, rather than underneath, to save them from disturbance by traffic. On the mining transit, an auxiliary telescope outside the trunnion bracket facilitates steep sightings.

Traditionally the most exciting underground survey process has been the transferring of a direction from the surface. In shallow shafts, steep (but not vertical) sights may transfer the direction. Another technique is to hang two weighted wires down the shaft, observe the direction between them on the surface, and use this direction as a control below. The relative shortness of distance between wires makes the transfer geometrically weak, but procedural care enables satisfactory results. An alternative procedure applies inertial navigation principles to underground surveying. A north-seeking gyro mounted on a transit gives azimuth accuracy to about 6 seconds from a plumbed position at shaft bottom and carries tunnel lines forward without the angle-error accumulation in ordinary transit traverse lines (**Fig. 6**).

Hydrographic surveys. Data for navigation charts and underwater construction are provided by hydrographic surveys. The horizontal locations of depth measurements must be referenced to recognizable controls. Where the shoreline is visible, it is mapped and a system of triangulation stations is established on shore. Transits at two triangulation stations can be used to observe directions to the sounding vessel whenever it signals that a depth measurement is made. A check angle may be obtained by sextant observation of shore points from the deck of the vessel, or a third transit on shore may be used to provide a check intersection. Angular observations from a shore point, coupled with EDM lengths to the sounding vessel reflector or transponder, will also fix the sounding location. In another procedure, a sending microwave unit on the vessel can find distances from slave stations set on shore points, to be recorded (even plotted) instantly along with the depth sounded.

Cadastral surveys. To establish property boundary lines, cadastral surveys are made. Descriptions based on horizontal surveys are essential parts of any document denoting ownership or conveyance of land. The basis rule of property lines and corners is that they shall remain in their original positions as established on the ground. This basis rule is important because most land surveys are resurveys. They may follow the original description, but this description is merely an aid to the discovery of the originally established lines and corners. Substantial discrepancies are frequently found in original descriptions because low-order surveying devices such as the compass or the link chain were once used.

Fig. 6. North-seeking gyro eases underground survey tasks. (*Wild Heerbrugg Instruments, Inc.*)

Surveys in the original 13 states, plus Tennessee, Kentucky, and parts of others, are conducted on the metes and bounds principle. In the so-called public land states and in Texas, the basic subdivisions are rectangular.

If the boundaries to be described border an irregular line, such as a winding stream, a good mathematical description may be impossible. Such a line can, however, be located by a series of closely spaced perpendicular offsets from an auxiliary straight line.

In the rectangular system, the land parcels of a region are described by their relationship to an initial point. In public-land states the initial point is the intersection of a meridian (principal meridian) and a latitude (baseline), as in **Fig. 7**. Townships, normally 36 mi^2 (93.24 km^2), are designated by their position with respect to the initial point—the number of tiers north or south of a given baseline and the number of ranges east or west of the corresponding principal meridian.

Within the township each square mile, or section, has a number, from 1 to 36. Sections are subdivided into quarter sections (160 acres or 0.6475 km^2), and they may be subdivided further. North-south section lines are meridians originating at 1-mi (1.609-km) intervals along the baseline and along standard

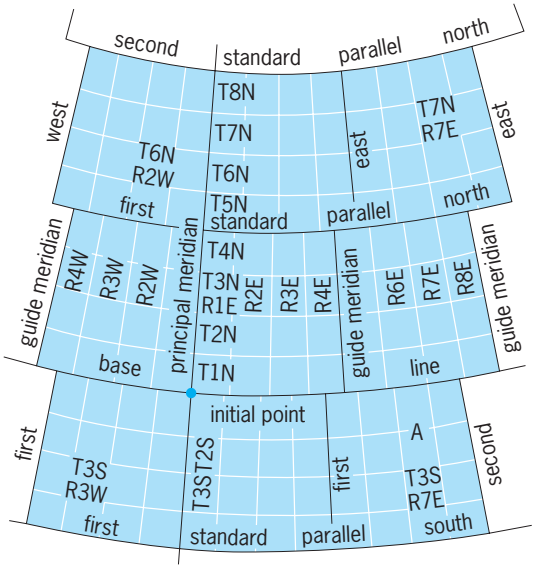


Fig. 7. Rectangular land layout. (*After P. Kissam, Surveying for Civil Engineers, McGraw-Hill, 1956*)

parallels of latitude (correction lines) generally spaced 24 mi (38.62 km) apart. The respacing of meridional lines every 24 mi reduces the effect of meridian convergence on the size of sections.

Resurveys become difficult where the corners are obliterated or lost. An obliterated corner is one for which visible evidence of the previous surveyor's work has disappeared, but whose original position can be established from other physical evidence and testimony. A corner is deemed lost when no sufficient evidence of its position can be found. Restoration requires a faithful rerun of the recorded original survey lines from adjacent points, distance discrepancies being adjusted proportionately. *See* CIVIL ENGINEERING.

B. Austin Barry

Bibliography. B. A. Barry, *Construction Measurements*, 2d ed., 1988; C. M. Brown, *Evidence and Procedures for Boundary Location*, 3d ed., 1993; H. Kahmen and W. Faig, *Surveying*, 1988; J. C. McCormac, *Surveying*, 3d ed., 1994; F. H. Moffitt and H. Bouchard, *Surveying*, 10th ed., 1997; P. R. Wolf and R. C. Brinker, *Elementary Surveying*, 9th ed., 1997.

Surveying instruments

Instruments used in surveying operations to measure vertical angles, horizontal angles, and distance. The devices used for these measurements were originally mechanical only, but advances in the technology led to the development of mechanical-optical devices, optical-electronic devices, and finally, electronic-only devices.

Level. Four types of levels are available: optical, automatic, electronic, and laser.

Optical level. An optical level is used to project a line of sight that is at a 90° angle to the direction of gravity. Both types, dumpy and tilting, use a precision leveling vial to orient to gravity (**Fig. 1**). The dumpy type was used primarily in the United States, while the tilting type was of European origin and used in the remainder of the world. The dumpy level has the leveling vial fixed to the telescope, which is fixed at 90° to a rotatable vertical spindle. Leveling screws, attached to the spindle, are used to center the leveling vial. The telescope can be pointed in any direction, and the line of sight through the telescope will then project at 90° to the direction of gravity. Tilting levels are similar to dumpy levels, with the only exception being that the leveling vial is connected to the telescope with a fine-movement (tilting) screw. Leveling screws are used to roughly center the vial, and the tilting screw is used for the final centering of the vial for each observation. *See* LEVEL MEASUREMENT.

Automatic level. Automatic levels (**Fig. 2**) use a pendulum device, in place of the precision vial, for relating to gravity. The pendulum mechanism is called a compensator. The pendulum has a prism or mirror, as part of the telescope, which is precisely positioned by gravity. The pendulum is attached to the telescope by using precision bearings or wires (metallic or nonmetallic). Leveling screws are used to roughly center a circular vial, and the optics on the pendulum then correct the line of sight through the telescope. Automatic levels are easy and fast to use, resulting in their domination of the optical-level market segment.

Fig. 1. Engineer's dumpy level. (*W. and L. E. Gurley Co.*)

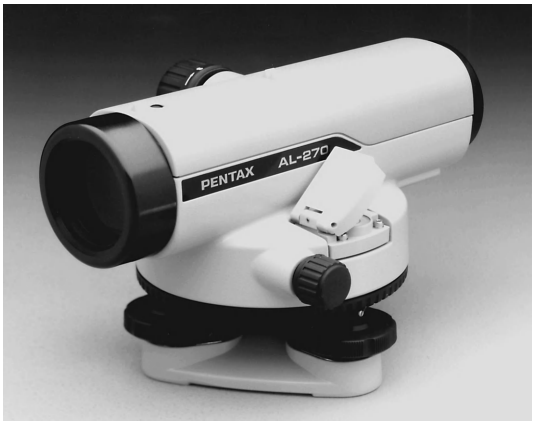


Fig. 2. Automatic level. (*Pentax Corp.*)

Electronic level. This type of instrument has a compensator similar to that on an automatic level, but the graduated leveling staff is not observed and read by the operator. The operator has only to point the instrument at a bar-code-type staff, which then can be read by the level itself. The electronic determination of the data is a further advantage because the data can be transferred to a data collector or stored on-board in a memory module or card. The electronic level eliminates human reading error and increases the speed at which leveling work can be performed. The only significant disadvantage is the high cost as compared to the optical automatic level.

Laser level. Although this type of instrument is categorized as laser, these levels actually employ three different types of light sources: tube laser, infrared diode, and laser diode. The instrument uses a rotating head to project the laser beam in a level 360° plane. The advantages are twofold: no operator is required once the instrument is set up; and different people in various locations can work by using a single light source. The disadvantages are that accuracy is less than that provided by other types of levels and that the cost is significantly higher. This type of instrument is used primarily for construction surveying.

Tube lasers have a visible beam with high power requirements. The power source is usually a 12-V car battery or an ac-to-dc converter. Infrared diode units have a nonvisible beam that can operate on flashlight batteries. The nonvisible beam requires that an electronic detector be used to "see" it. This instrument has a lower cost than tube lasers, and is physically smaller. The laser diode level has all the advantages of the other two types: low power requirement, small size, and visible beam.

There are three different mechanisms used to level-up laser levels. Leveling vials were used initially, and they are still used on the most inexpensive units. Moderate- to mid-priced units employ compensators. For the higher-priced units, electronic vials with servomotors mechanically level the mechanism inside the case. Servo types are also available in models that can tilt the rotating beam to a particular percent of grade. Dual-grade instruments, the most expensive type, can project two different grades at the same time, for example, (+)3% in the north direction and (−)5% in the east direction. *See* SERVOMECHANISM.

Transit. The primary purpose of a transit (**Fig. 3**) is to measure horizontal and vertical angles. Circles, one vertical and one horizontal, are used for these measurements. The circles are made of metal or glass and have precision graduations engraved or etched on the surface. A vernier is commonly used to improve the accuracy of the circle reading. *See* VERNIER.

The vertical circle, fixed to a horizontal axis that is part of and at 90° to a telescope, lets the telescope rotate in a vertical plane. The alidade, or the framework that supports the telescope axis, has a vertical axis (spindle) attached to its base, which allows the alidade to rotate with respect to the horizontal circle (Fig. 3). Level vials, attached to the alidade, are used to make the spindle vertical (in line with gravity), so that measurement of horizontal angles can be measured. A level vial is also attached to the telescope and provides the gravity index to measure vertical angles.

Theodolite. The theodolite serves the same purpose as the transit, and they have many similar features. The major differences are that the measuring circles are constructed only of glass and are observed through magnifying optics to increase the accuracy of angular readings. Theodolites are generally smaller than transits. Some models have a compensator, which allows quicker setups while maintaining vertical angle accuracy. Theodolites also have a separate, low-power telescope to view the setup point on the ground. This telescope replaces the use of a plumb bob to position the instrument over a point. Optical theodolites largely replaced transits but then were replaced by electronic theodolites.

The electronic theodolite uses electronic reading circles in place of the optically read ones. The circle readings are observed on a display located on the alidade. Some versions of this instrument can transmit the electronic readings via a serial port to an external data memory device or to an electronic distance meter attached to the theodolite.

Electronic distance meter. Electronic distance meters use either light waves or radio waves as their

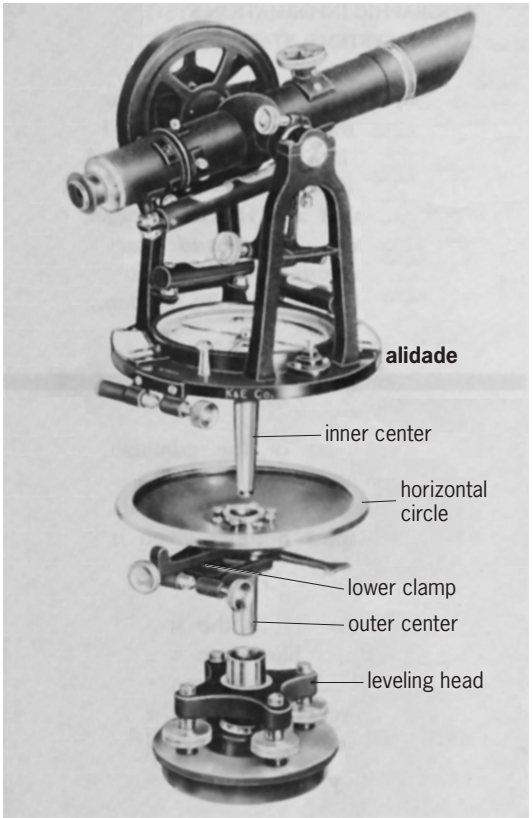


Fig. 3. Disassembled transit, showing its components. (*Keuffel and Esser Co.*)

measuring device. Radio-wave instruments dominated this category until the late 1960s, when lower-cost light-wave instruments became available.

Radio-wave instruments require similar units placed at each end of the line to be measured, are capable of longer measurements (up to 90 mi or 140 km), and do not require a direct line of sight between the units; however, they are more sensitive to atmospheric conditions.

Light-wave types use visible, laser, or infrared light. Only one unit is required, with the other end of the measured line occupied by a special type of prism reflector. Measurements are more precise, distance capability is shorter (up to 40 mi or 60 km for laser and 6 mi or 10 km for infrared), and a direct line of sight is required between the unit and the reflectors.

These instruments provide more stable results. Accuracy has improved to the degree that a basic accuracy of ±2 mm is commonly achieved. These instruments typically use several frequencies and phase-comparison techniques to determine distance, not time measurement.

Total stations. These units consist of combinations of devices, which can be optical, electronic, data-electronic, or motorized electronic.

Optical. The combination of a theodolite and an infrared electronic distance meter into one instrument is commonly referred to as an optical total station. The configuration can be one integrated unit, or it can be modular in design. Integrated types typically have longer-range distance capabilities matched with higher angle accuracy. Modular types can have any combination of short- to long-range distance measuring matched with low-to-high angle accuracy.

Electronic. An electronic total station (**Fig. 4**) comprises an electronic theodolite in combination with an infrared electronic distance meter. This type of instrument has become the standard surveying instrument, having replaced all other angle and distance instruments for most applications. Compensators are common on these instruments, and many models have dual-axis types that correct for misleveling in both directions. Some units have been developed that have many useful programs built in. The category of electronic total station includes models that serve virtually all survey needs, and less expensive models have been designed specifically for use in construction. These instruments are very efficient when used alone, and the addition of a data collector, replacing hand-written notes, is the next step in a fully electronic system. The data collector reduces mistakes in the field and can perform some useful calculation functions quickly and easily. The final component is the office computer to which the data collector can transfer the information. Through the use of specialized surveying software, maps are generated by a plotter or printer (Fig. 4).

Data electronic. Miniaturization of electronic components made it possible to combine an electronic total station with the data collector. Small removable modules or cards are used to store the data, which can subsequently be transferred to the office computer. These models provide a one-piece instrument, eliminating cables and facilitating transport. Some models have full alphanumeric keyboards so that descriptions and notes can be added to the measured data. Special units have been designed that use custom programs.

Fig. 4. Electronic total station. (*Pentax Corp.*)

Motorized electronic. This type of electronic total station has two characteristics that distinguish it from the standard type. The manual alignment mechanisms are replaced by motors, and the optical telescope is replaced by an electronic type. The instrument can align itself to the prism reflector, thus eliminating the instrument operator. Data recording can be initiated remotely by the reflector operator, making the process a one-person operation. This reduction in personnel is the primary advantage of this system.

Global Positioning System. The U.S. Department of Defense installed a satellite system for navigation and for establishing the position of planes, ships, vehicles, and so forth. This system uses special receivers and sophisticated software to calculate the longitude and latitude of the receiver. It was discovered early in the program that the distance between two nonmoving receivers could be determined very accurately and that the distance between receivers could be many miles apart. This technology has become the standard for highly accurate control surveys, but it is not in general use because of the expense of the precision receivers, the time required for each setup, and the sophistication of the process.

Mapping of city features such as systems for supplying water, sewage collection, and electricity does not require the precision of normal surveying. Low-cost, low-precision receivers are used for this purpose, and cities are able to have cost-effective multilayer databases that contain all of the city structure located below and above the ground. *See* GEOGRAPHIC INFORMATION SYSTEMS; SATELLITE NAVIGATION SYSTEMS; SURVEYING. Kerry W. Kemper

Bibliography. J. M. Anderson and E. M. Mikhail, *Surveying: Theory and Practice*, 7th ed., 1997; C. D. Burnside, *Electromagnetic Distance Measurement*, 3d ed., 1991; M. A. R. Cooper, *Modern Theodolites and Levels*, 2d ed., 1982; F. Deumlich, *Surveying Instruments*, 1982; B. Hofmann-Wellenhof and H. Lichtenegger, *Global Positioning System: Theory and Practice*, 5th ed., 2001; A. Leick, *GPS Satellite Surveying*, 3d ed., 2003; J. M. Rueger, *Electronic Distance Measurement*, 4th ed., 1996.

Susceptance

The imaginary part of the admittance of an alternating-current circuit.

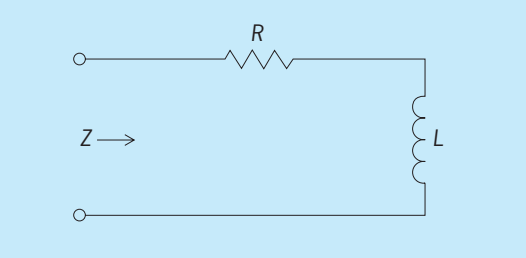
The admittance, Y, of an alternating current circuit is a complex number given by Eq. (1). The imaginary part, B, is the susceptance. The units of susceptance like those of admittance are called siemens or mhos. Susceptance may be either positive or negative. For example, the admittance of a capacitor C at frequency ω is given by Eq. (2), and so B is positive. For an inductor L, the admittance is given by Eq. (3), and so B is negative.

$$Y = G + jB \tag{1}$$

$$Y = jC\omega = jB \tag{2}$$

$$Y = -\frac{j}{L\omega} = jB \tag{3}$$

In general, the susceptance of a circuit may depend on the resistors as well as the capacitors and inductors. For example, the circuit in the **illustration** has impedance given by Eq. (4) and admittance given by Eq. (5), so that the susceptance, given by Eq. (6), depends on the resistor R as well as the inductor L.



Circuit with a resistor and inductor in series.

$$Z = R + jL\omega \tag{4}$$

$$Y = \frac{1}{R + jL\omega} \tag{5}$$

$$B = \frac{-L\omega}{R^2 + L^2\omega^2} \tag{6}$$

See ADMITTANCE; ALTERNATING-CURRENT CIRCUIT THEORY; ELECTRICAL IMPEDANCE. J. O. Scanlan

Suspension

A system of small solid particles dispersed in a liquid. Suspensions are a type of colloidal dispersion. By classical definition, suspension particles have diameters (or, if not spherical, the largest dimension) between 1 nanometer and 1 micrometer. However, for practical purposes the principles of colloid science can be usefully applied to suspensions for which the particles are smaller, down to about 0.1 nm (nanoparticle suspensions), and also to those for which the largest particle dimensions are tens or even hundreds of micrometers. The particles in a suspension are large enough that they do not behave like the atoms and molecules of classical chemistry. For example, suspensions do not generally behave like true solutions, and may have undetectable freezing-point depressions. On the other hand, the particles are small enough that they do not behave like the macroscopic particles of classical physics. For example, sedimentation of a suspension's particles may occur extremely slowly in apparent violation of Stokes' law. *See* COLLOID; PARTICULATES; SEDIMENTATION (INDUSTRY).

In most suspensions, the solid particles are dispersed in an aqueous liquid phase (that is, water containing dissolved substances), which may be denoted solids-in-water or S/W. In some suspensions, however, the liquid phase is an oil (usually a hydrocarbon liquid), and the suspension is denoted solids-in-oil or S/O. The **table** shows some examples.

Preparation. Suspensions can be formed by one of two basic methods. In the nucleation process, either a gas is made to condense to form solid particles or particles are precipitated from a solution. Once the fine particles have been created, they can grow to larger sizes as additional material condenses or precipitates onto the surfaces of existing particles (Ostwald ripening). The second basic suspension preparation method involves breaking-up large particles into progressively smaller ones (comminution). This process requires shearing forces, which can be provided by various devices, including cutting, rolling, crushing, or grinding mills. When the solid particles are already quite small, other devices such as propeller-style mixers or ultrasound generators can be used.

If a small amount of powdered fine solid is added to a cup of tap water and vigorously agitated in a blender, the fine solids will become quite well dispersed in the water and a suspension is made. This suspension may or may not be very stable, however.

Examples of suspension occurrence

Application area	Examples	Suspension types*
Environment	Sediment particles in oceans, lakes, rivers	S/W
	Muds, quicksand	S/W
	Rock fragments in lava	S/L
	Solid silicates in magma	S/L
	Solid particles in raindrops	S/W
	Biocolloids (viruses, bacteria) in rivers, ponds	S/W
Mineral processing	Mineral process slurries (pulps)	S/W
	Mineral flotation froths	S/W
	Oil flotation and oil-assisted froths	S/W, S/O
	Mineral tailings slurries and ponds	S/W
Petroleum production	Migrating fine solids in an oil reservoir	S/W
	Drill-cutting slurries	S/W, S/O
	Dispersed asphaltenes in crude oils	S/O
	Oil-well-produced solids	S/W, S/O
	Oilfield tailings ponds, sludges	S/W, S/O
Manufacturing and materials	Cellulose fiber pulp slurries	S/W
	Deinking pulp slurries	S/W
	Pigment-containing paints	S/L
	India ink	S/W
Food products	Ice crystals in frozen ice cream	S/O
	Vegetable shortening	S/O
	Jellies	S/W
	Chocolate drinks	S/O
	Water ice (a semifrozen drink)	S/W
Agriculture	Pesticide, herbicide, and fungicide	S/W
Biology and medicine	Polymer-encapsulated drugs	S/W
	Biodegradable drug suspensions	S/W
	Diagnostic suspensions	S/W
Personal care	Exfoliating scrubs	S/W
	Facial masks	S/W
	Lipsticks and lip balms/glosses	S/O

*Solid, liquid, water, and oil.

In the absence of turbulence and if the particles are not electrically charged, they will fall (settle) out if their density is greater than that of the liquid. It is principally the small size that keeps suspended sediment particles in a river from rapidly settling out. A fine sand particle of 100-μm diameter will settle 1 m (3 ft) in water in about 2 minutes, whereas a 10-nm colloidal particle will take about 20 years (in the absence of turbulence).

In addition, the particles may not remain separated from each other. When Brownian motion, stirring, or settling cause particles to encounter each other, they may stick together, possibly touching at one or more points. This is called aggregation. In suspensions, there are two principal kinds of aggregation: coagulation and flocculation. Coagulation refers to the formation of compact aggregates. Flocculation refers to the formation of a loose network of particles. Coagulated particles tend to settle quickly and form compact, dense sediments, while flocculated particles tend to settle slowly and form loose, high-volume sediments. Which kind of behavior occurs depends mainly on the properties of the particles' surfaces. *See* BROWNIAN MOVEMENT.

In order for dispersed particles to be reasonably stable and not aggregate or settle quickly, suspension particles need to be stabilized. The two most common means of stabilizing suspended particles are by creating electric charges or adsorbing polymers on the particles' surfaces. Some particles gain a surface electric charge due to reactions that occur when they are dispersed into water. Another way to create particle surface charges is by adsorption of chemicals such as ionic surfactants or polymers. Electrically charged particles repel each other, which can dramatically slow down the rates of both aggregation and sedimentation, making a suspension reasonably stable. *See* ADSORPTION; POLYMER; SURFACTANT.

The stability of a suspension can also be enhanced by adsorbing certain kinds of polymers onto particle surfaces, without involving electric charges. In this case, the adsorbed polymer molecules extend outward from the particle surfaces into solution, making it difficult for particles to approach each other, and therefore providing a barrier to aggregation (steric stabilization). Properly formulated suspensions can remain stable for months or even years.

Properties. Although suspensions can be prepared that will remain stable for weeks, months, or years, no suspensions are truly stable. If the particles are dispersed in a liquid of lower density, they will eventually fall (sediment) due to gravity. Similarly, if the particles are dispersed in a liquid of greater density,

the particles will eventually rise (cream) to the surface.

By dispersing fine particles in a liquid, a very small amount of solid material can be made to expose a very large amount of surface area. This can create a very large number of sites at which chemical reactions can occur. As a result, nanoparticle suspensions have potential applications in high-surface-area catalysis, nanosized pigments and/or fillers in paints and other kinds of coatings, and ceramics. *See* NANOPARTICLES.

In a very dilute suspension of small particles, the individual particles will not encounter each other very often. However, this changes as the volume occupied by the particles increases above a few percent. As particles begin to encounter each other more and more frequently, there is a crowding effect, and it becomes more difficult to make the suspension flow. That is, the suspension's viscosity increases. If the particle surfaces are electrically charged and/or have adsorbed surfactant or polymer molecules, suspension viscosity increases even more. This can have a great impact on suspension stability since suspended particles will sediment out more slowly as the viscosity increases. *See* VISCOSITY.

Another consequence of dispersed particles encountering each other in suspension is that some of the particles will begin to aggregate. Clay mineral particles are of special interest because they can have very different particle dimensions. Their ratio of particle diameter to thickness can range from 10:1 to 250:1, making them appear to resemble sheets of paper more than spheres or polyhedra. For such particles, flocculation produces aggregates of particles that are oriented edge-to-edge and/or edge-to-face, in a "house of cards" fashion. Such particle structuring can increase to the point where essentially all particles in the suspension are networked and the suspension gels. The particular ways in which particles interact with each other in concentrated suspensions can cause some unusual properties. Some suspensions, like ketchup, flow more easily the faster they flow (pseudoplastic). Other suspensions, like the silicone-based polymer Silly Putty®, actually flow less easily the faster they are made to flow (dilatant). Still other suspensions, like grease and toothpaste, do not flow at all until a sufficiently high shearing force is applied (yield stress). *See* CLAY MINERALS; GEL; RHEOLOGY.

Characterization. Suspensions may be characterized in a number of ways.

1. The nature of the suspension (that is, whether the particles are dispersed in water, an aqueous solution, or an oil) can often be quickly determined by electrical conductivity.

2. Particle size and particle size distribution can be determined by techniques such as optical or electron microscopy and light scattering. Large-sized particles ($>5\ \mu$m) can be evaluated using sieves.

3. The nature of the solids themselves is usually determined by separating them out, followed by mineralogical analysis using a technique such as x-ray diffraction. A variety of spectroscopic techniques can be used to study the nature and properties of the particle surfaces.

4. Stability of a suspension against sedimentation is usually determined using standardized bottle or column tests, in which suspension samples are allowed to stand for a specified time with or without centrifuging and then are examined. *See* LIGHT-SCATTERING TECHNIQUES; X-RAY DIFFRACTION.

Occurrence. Suspensions of colloid-sized particles are important because they feature prominently in both desirable and undesirable contexts and in a wide variety of practical disciplines, products, and industrial processes (see table). The problems associated with suspensions are usually interdisciplinary in nature, and a broad scientific base is required to understand them completely.

Suspensions commonly occur in the environment. Examples range from the suspended sediments in rivers, lakes, and oceans to bogs and quicksand. Suspensions are commonly used in many industries, such as pulp and papermaking. They are created at an early stage of processes used to separate valuable minerals by froth flotation. Suspensions are also quite common in the petroleum industry, and may occur in reservoirs, drilling fluids, production fluids, process plant streams, and tailings ponds. In these kinds of applications, it is probably the norm rather than the exception for suspensions to contain not just solid particles and water but also emulsified oil and even dispersed gas bubbles.

Some familiar suspensions include those occurring in foods like batters, puddings, and sauces, in pharmaceuticals like cough syrups and laxatives, and in household and industrial products like inks, paints, and "liquid" waxes. Examples in personal care products include lipstick and lip balms, which are usually concentrated suspensions of solid oils in a liquid oil or in a mixture of liquid oils. The properties of such lipsticks can be adjusted by dispersing additional kinds of particles into them, such as adding solid titanium dioxide or zinc oxide to provide sunscreening, or by adding bismuth oxychloride (synthetic pearl) to create a frosted effect, or by adding mica or silica particles to create shimmer or glitter effects.

Suspension destabilization. If a suspension is stabilized by having electric charges at particle surfaces, a simple way to destabilize the suspension is to reduce the electric repulsion forces caused by those charges. This can be done in one of two ways. First, simply adding a soluble salt will produce a screening effect that reduces the strength and penetrating power of the electric repulsion forces, allowing the particles to aggregate and settle. Although almost any soluble salt will do, the most effective charge screening will be produced by salts having multicharged (multivalent) ions that are of opposite charge to that of the particle surfaces. For example, suppose a suspension of negatively charged particles can be made to aggregate by the addition of 1 mole/liter sodium chloride (NaCl). The same state of aggregation could be produced by the addition of only

about 0.018 mole/L calcium chloride ($CaCl_2$) or an even smaller amount, about 0.002 mole/L aluminum chloride ($AlCl_3$). Second, the particle surface charges can be directly counteracted by adding species of opposite charge that can adsorb directly onto the particle surfaces. For negatively charged particle surfaces, such species could be heavy-metal ions (such as copper ions) or positively charged surfactants or polymers.

Just as the stability of a suspension can be enhanced by adding polymers that adsorb onto particle surfaces (causing steric stabilization; see above), suspension stability can also be reduced by the adding large polymers that can act as bridging agents by adsorbing onto more than one particle (bridging flocculation). The bridging action causes the formation of large clusters of particles, called flocs, which can rapidly settle out and then be filtered out easily. Selecting an appropriate polymer can involve considerable testing due to the large number of factors than can influence suspension stability. Bridging flocculation also requires addition of just the right amount of polymer, since too little polymer will be ineffective and too great an addition can cause the polymers to adsorb onto individual particles instead of bridging across several particles (this can even make the suspension more stable than it was before treatment).

Once a suspension is destabilized, the particles can usually be easily separated by sedimentation, centrifugation, or filtration. A combination of simple salt addition (such as alum) for charge screening and polymer addition for bridging, followed by settling and filtration, forms the basis of many municipal water-treatment plant processes. *See* SURFACE AND INTERFACIAL CHEMISTRY; WATER SUPPLY ENGINEERING.

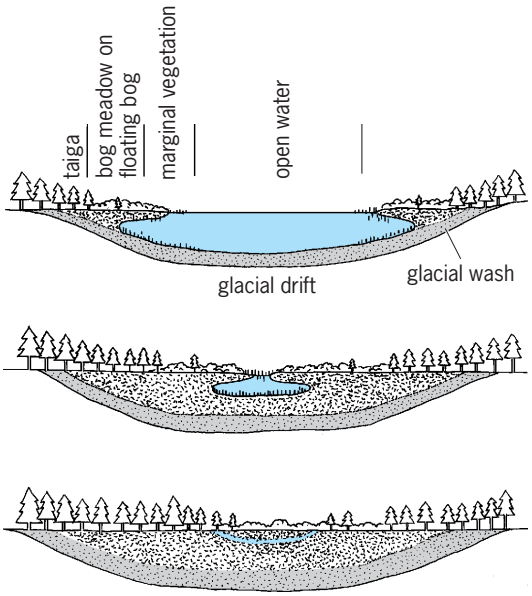
Laurier L. Schramm

Bibliography. R. K. Iler, *The Chemistry of Silica: Solubility, Polymerization, Colloid and Surface Properties and Biochemistry*, Wiley, 1979; L. L. Schramm, *Emulsions, Foams, and Suspensions: Fundamentals and Applications*, Wiley-VCH, Weinheim, 2005; L. L. Schramm (ed.), *Suspensions: Fundamentals and Applications in the Petroleum Industry*, American Chemical Society, 1996; H. van Olphen, *An Introduction to Clay Colloid Chemistry*, 2d ed., Wiley-Interscience, 1977; S. Yariv and H. Cross, *Geochemistry of Colloid Systems for Earth Scientists*, Springer-Verlag, Berlin, 1979.

Swamp, marsh, and bog

Wet flatlands, where mesophytic vegetation is areally more important than open water, are commonly developed in filled lakes, glacial pits and potholes (see **illus.**), or poorly drained coastal plains or floodplains. Swamp is a term usually applied to a wetland where trees and shrubs are an important part of the vegetative association, and bog implies lack of solid foundation. Some bogs consist of a thick zone of vegetation floating on water.

Unique plant associations characterize wetlands in various climates and exhibit marked zonation characteristics around the edge in response to different thicknesses of the saturated zone above the firm base of soil material. Coastal marshes covered with vegetation adapted to saline water are common on all continents. Presumably many of these had their origin in recent inundation due to post-Pleistocene rise in sea level.



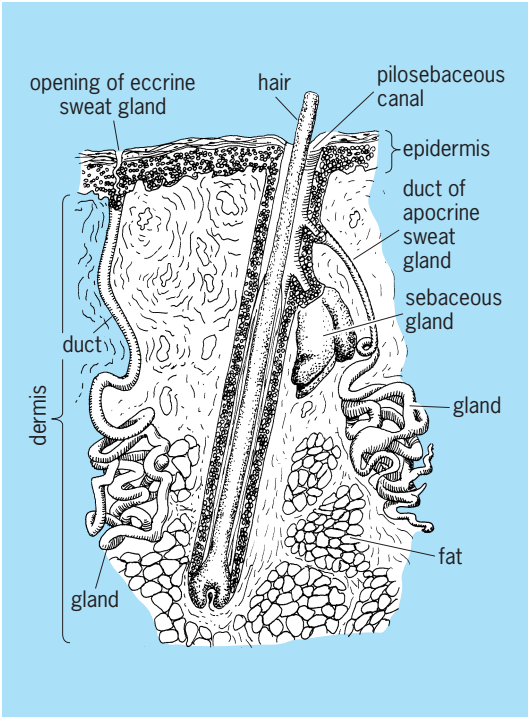
Cross-sectional diagram representing the progressive filling by vegetation of a pit lake in recently glaciated terrain.

The total area covered by these physiographic features is not accurately known, but particularly in glaciated regions many hundreds of square miles are covered by marsh. *See* GLACIATED TERRAIN; MANGROVE.

Luna B. Leopold

Sweat gland

A coiled, tubular gland found in mammals. There are two kinds, merocrine (or eccrine) and apocrine. The latter are generally associated with hair follicles (see **illus.**). Merocrine glands are distributed extensively over the body in the human, whereas the apocrine variety is restricted to the scalp, nipples, axilla, external auditory meatus, external genitals, and perianal areas. Apocrine sweat glands are more numerous in mammals, with the exception of the chimpanzee and human, in which the merocrine variety predominates. The mammary glands probably represent modified apocrine sweat glands which grow inward and increase in complexity. In association with adipose tissue, they eventually form pendant structures, the mammae, which project outward from the general contour of the skin's surface. The secretion process is apocrine with a considerable portion of the cell being discharged. The discharged portion of such a gland cell disintegrates to free fat droplets and albuminous substances. A mammary gland is complex and represents an association of lobes. Each



Human skin showing structure of both eccrine and apocrine sweat glands.

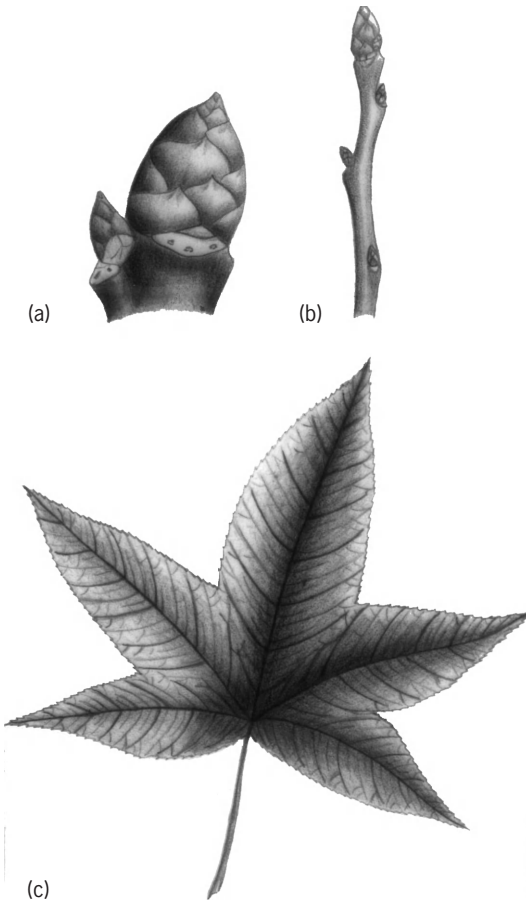
lobe contains a compound alveolar (acinous) gland with a separate lactiferous duct which opens on the nipple in the human. The glands of Moll associated with the eyelashes are relatively large modified apocrine glands as are the ceruminous or wax glands in the external auditory meatus. The anal sacs of the skunk presumably are apocrine glands modified by the addition of muscle fibers from the levator ani muscle which enables the pungent contents to be ejected with force. *See* EPITHELIUM; GLAND; LACTATION; MAMMARY GLAND.

Olin E. Nelsen

Sweetgum

The tree *Liquidambar styraciflua*, also called redgum, a deciduous tree of the southeastern United States. It is found northward as far as southwestern Connecticut, and also grows in Central America. The tree is commonly 80 to 120 ft (24 to 37 m) in height and $1^1/_2$ to 3 ft (0.45 to 0.9 m) in diameter, but individual trees may exceed these dimensions. Sweetgum is readily distinguished by its five-lobed, or star-shaped, leaves (see **illus.**) and by the corky wings or ridges usually developed on the twigs. The erect trunk is a dark gray, but the branches are lighter in color. In winter the persistent, spiny seedballs are an excellent diagnostic feature. The strong, close-grained wood is light brown tinged with red, and has a satiny luster and attractive grain. Because of its tendency to warp, it was long considered to be of inferior quality, but technical processing has largely overcome these difficulties. It is sometimes marketed as satin or Circassian walnut, or as hazelwood. *See* HAMAMELIDALES.

The annual commercial production of sweetgum comes mainly from the South Atlantic and Gulf Coastal Plains and from the lower Mississippi Valley. It is used for furniture, interior trim, railroad ties, cigar boxes, crates, flooring, barrels, woodenware,



Sweetgum (*Liquidambar styraciflua*): (*a*) terminal bud, (*b*) twig, and (*c*) the distinctive five-lobed leaf.

and wood pulp, and it is one of the most important materials for plywood manufacture. Sweetgum is one of the most desirable ornamental trees, chiefly because of its brilliant autumn coloration. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Swim bladder

A gas-filled sac found in the body cavities of most bony fishes (Osteichthyes). The swim bladder has various functions in different fishes, acting as a float which gives the fish buoyancy, as a lung, as a hearing aid, and as a sound-producing organ. In many fishes it serves two or three of these functions, and in the African and Asiatic knifefishes (Notopteridae) it may serve all four. The swim bladder contains the same gases that make up air, but often in different proportions. *See* OSTEICHTHYES.

Respiratory function. The most ancient bony fish probably had a single lung which gave rise in the course of evolution to the swim bladders of modern fishes. The lung is retained as such in some primitive modern bony fishes, such as the bowfin (*Amia*), the garpike (*Lepisosteus*), and the lungfishes. The African and South American lungfishes (*Protopterus* and *Lepidosiren*) have paired lungs. Fish with lungs use them to breathe air in essentially the same way as land-living vertebrates, though they use their gills to obtain oxygen from water as well. Most teleost swim bladders have lost the respiratory function, but a few retain it or have regained it. *See* RESPIRATORY SYSTEM.

Buoyancy. The gas in the swim bladder naturally makes a fish more buoyant, whether or not its main function is to give buoyancy. Most fish without swim bladders are a good deal denser than water, but many with swim bladders are almost exactly the same density as the water they live in. Most fish need swim bladders occupying about 7% of the whole volume of the body to make them the same density as freshwater. Rather smaller swim bladders suffice in seawater, because it is denser.

The main advantage of being the same density as the water is that it saves energy. It enables the fish to hover almost motionless in midwater, whereas a fish which is denser than water cannot rest in midwater but must either keep swimming or allow itself to sink. Fish which are denser than water also need more energy to swim at a given speed than similar fish of the same density as water. This is because they need hydrodynamic lift to support themselves in the water (they obtain this lift by holding their fins extended like airplane wings or by similar means). This increases the resistance to their progress through the water, because hydrodynamic lift cannot be obtained without induced drag.

It has been calculated that a fish which had no swim bladder and which swam most of the time would use up to 20% more energy than a similar fish with similar habits with a swim bladder which made it the same density as the water. This puts the fish without a swim bladder at a severe disadvantage, and it is easy to see why most teleosts which spend much time swimming have swim bladders. It is hard to see why some mackerels and tunas do not have swim bladders, although they swim all the time.

Many teleosts which live and feed on and near the bottom have rudimentary swim bladders or no swim bladders at all. Flatfishes (Pleuronectiformes), for instance, have no swim bladder. A fish with the same density as water has no weight in water, so there is no friction to keep it in position when it rests on the bottom. This may outweigh the advantages of having a swim bladder.

Secretion and resorption of gases. Pressure under water increases with depth, so a fish must increase the quantity of gas in its swim bladder as it swims deeper if it is to keep the volume of the swim bladder and the density of the body constant. Similarly, a fish must remove gas from its swim bladder when it swims nearer the surface if its density is not to change. The more primitive teleosts such as the trout (*Salmo*, **Fig. 1**) retain the tube connecting the swim bladder to the gut which served as a windpipe for the lung of their ancestors. They can spit gas out of their swim bladders whenever they have to but obviously can take in air only through this tube when they are at the surface. They also can add gas slowly to their swim bladders by releasing dissolved gases from their blood. More advanced teleosts, such as the perch (*Perca*), have lost the tube and can only change the quantity of gas in their swim bladders by secreting gases from the blood or resorbing them into the blood. These are slow processes.

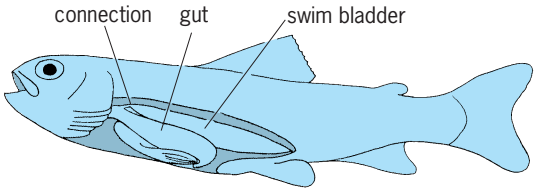


Fig. 1. Young trout (*Salmo trutta*), showing the swim bladder. The trout is one of the primitive teleosts which have the swim bladder connected to the gut.

Most fish which have been investigated take 4 to 48 h to refill their swim bladders by secretion after the gas has been removed experimentally. Secretion and resorption can therefore only keep pace with very slow changes of depth. The oceanic lantern fishes (Myctophidae), which swim hundreds of meters up toward the surface at dusk and down again at dawn, cannot secrete and resorb gases fast enough to keep the volumes of their swim bladders constant. Herrings (*Clupea*) and other fishes which make more modest daily changes of depth probably cannot compensate even for them.

Even if a fish makes no adjustments for depth changes, it still needs to be able to add gas to its swim bladder. This is because gases dissolved in the water tend to be in equilibrium with the atmosphere, so that their total partial pressure is about 1 atm (10^2 kilopascals). Gases in the swim bladder are always above atmospheric pressure, so long as the fish is submerged, because of the head of water above the fish. Therefore, gas tends to diffuse from the swim bladder to the water, mainly via the blood. The swim bladder would get gradually smaller if the lost gases were not replaced.

Gas gland. Many teleosts have a conspicuous red spot on the inside of the swim bladder wall from which gas is secreted. It is known as the gas gland. The bundle of blood vessels running to it is known as the rete mirabile. In the rete, capillaries carrying blood to the gland and capillaries carrying blood from it run parallel, mixed together (**Fig. 2**).

When the blood reaches the gas gland, lactic acid is secreted into it. This raises the partial pressure of the dissolved gases, especially of the oxygen. Therefore, when the blood returns to the rete, some of the oxygen and other gases diffuse across to the arterial capillaries instead of being carried away to the veins.

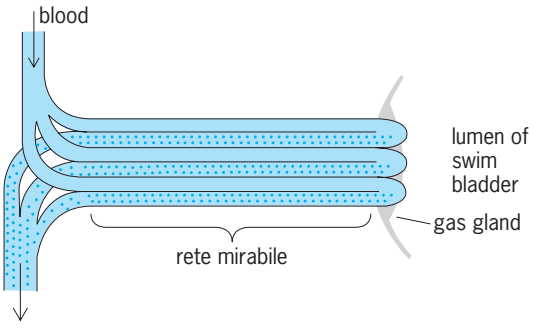


Fig. 2. Diagram showing the blood supply to the gas gland. Blood traveling away from the gland is stippled. There are far more capillaries in the rete than are shown in this diagram.

Thus high concentrations of gases get built up at the gas gland end of the rete. When the partial pressure of a gas in the blood comes to exceed its partial pressure in the swim bladder, it diffuses into the swim bladder. This theory of the mechanism of secretion was developed by Werner Kuhn and colleagues at the University of Basel. It has been largely confirmed by experiments on eels (*Anguilla*), though some doubt remains. Kuhn's calculations show that the theory provides a plausible explanation even for the secretion of gases at great depths in the sea, where the partial pressures of gases in the blood must be raised to enormous values before they will escape into the swim bladder.

Oval. Many teleosts have a specialized organ for resorbing gases from the swim bladder, as well as a gas gland for secreting them. The commonest type of resorbent organ is the oval. This is a pocket in the swim bladder wall, richly supplied with blood vessels. It can be closed by a ring of muscle around its opening. Resorption is a passive process because the pressure of the gases in the swim bladder always exceeds the total partial pressure of the dissolved gases in the blood. So long as the oval is open, gases from the swim bladder will diffuse into the blood in its blood vessels.

Sound production and reception. A swim bladder can vibrate, alternately swelling and contracting. Any disturbance will set it vibrating at its resonant frequency, just as a blow sets a gong vibrating at its resonant frequency. It can therefore be used to produce sounds. Various teleosts, including such different species as the toadfish (*Opsanus*) and the piranha (*Serrasalmus*), have special muscles which set their swim bladders vibrating and produce grunting or hooting noises. These sounds are probably used in courtship and in aggressive behavior. *See* UNDERWATER SOUND.

Sound vibrations are amplified near a swim bladder because gas is more compressible than water and responds with bigger volume changes to fluctuations of pressure. In some teleosts, such as herrings, an extension of the swim bladder reaches forward to the ears, giving them the full benefit of the amplifying effect of the swim bladder. In the minnows, catfishes, and their relatives, the same effect is achieved with tiny bones, the Weberian ossicles, which connect the swim bladder to the ear. The degree of amplification must depend on how near the frequency of the sound is to the resonant frequency of the swim bladder, but it has been calculated that sound at the resonant frequency may be amplified as much as a hundred times. It has been shown by experiment that catfish (*Ictalurus*) can hear much better than fish without Weberian ossicles and that their good hearing is spoiled if the swim bladder is deflated. *See* PHONORECEPTION.

R. McNeill Alexander

Bibliography. R. McN. Alexander, *Function Design in Fishes*, 3d ed., 1974.

Swine production

The science of breeding, raising, and marketing swine. This agricultural enterprise is usually located in proximity to sources of high-energy feedstuffs; in the United States, the geographical distribution of swine production is closely related to maize and sorghum production. Swine, as nonruminants, utilize large quantities of concentrate feeds; it is estimated that 30% of the maize grain produced is fed to swine. Swine can utilize only limited quantities of roughage in their diet.

The United States produces about one-fourth of the world's hogs. The swine production industry has changed from the status of a supplement to farm income to that of intensified units in which swine production is a major enterprise. The total number of swine producers decreased by 80% from 1950 to 1980, but the volume per unit is increasing sufficiently to maintain the total volume of production. More than 90% of all hogs are from independent producers. The others are produced by the intensified methods of integrated enterprises in confinement facilities which require highly trained biological managers capable of applying the principles of breeding, nutrition, physiology, environmental control, and economics.

Swine products. The primary products of swine are pork, lard, hides, and innumerable pharmaceutical by-products. Pork and lard supply about 15% of the total calories consumed as food in the United States. Pork is more successfully cured and stored than many other meats, and it is estimated that about 60% of the swine carcass is cured by various methods.

Breeding. There are at least 85 recognized swine breeds in the world, and perhaps more than 220 "varieties" recognized as breeds in certain geographical locations. Breeders of purebred, hybrid, and crossbred swine provide the seed stock for commercial hog production, and are entrusted with the responsibility of changing the carcass so that it better meets the preferences of consumers. The pure breeds of major importance in the United States include Berkshire, Chester White, Duroc, Hampshire, Landrace, Poland China, spotted Poland China, and Yorkshire. An example of a meat-type market-weight pig is shown in the **illustration**. In Great Britain and Europe, the Landrace and Yorkshire breeds are most prevalent.

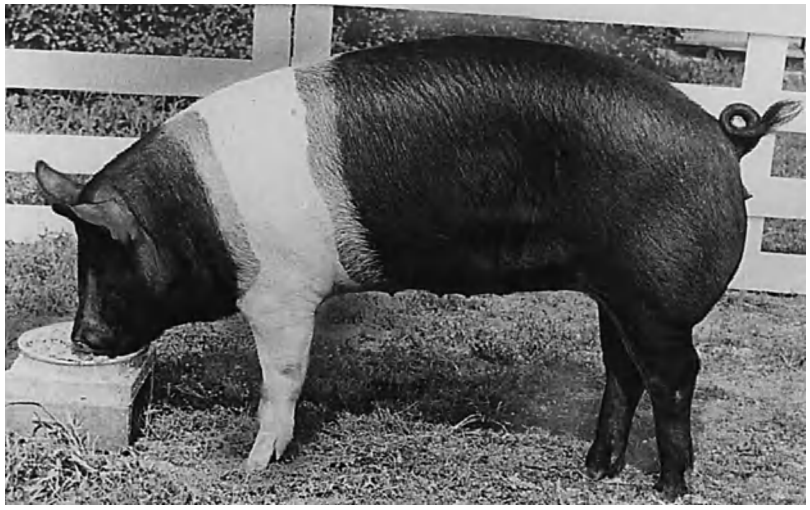
Most commercial producers of swine practice a system of crossbreeding, mating individuals from two, three, or four different breeds, since crossbreeding increases the vigor and feeding efficiency of the offspring. Crossbred dams mated to a boar of a third breed farrow and wean larger and heavier litters than purebred dams. In practical swine operations, rotation breeding is accomplished by using male "seed" stock from different breeds and retaining female stock sired by the boar. *See* BREEDING (ANIMAL).

Reproduction. American and European breeds reach sexual maturity between the ages of 4 and 7 months, though males usually exhibit sexual maturity at a slightly earlier age than females. Boars are not usually used for breeding until they are at least $7\frac{1}{2}$ months of age. Females are usually 7–8 months of age and weigh at least 220 lb (100 kg). Some native breeds and varieties, such as those in China, may reach sexual maturity by $3\frac{1}{2}$ months of age.

Estrus, or the time of sexual excitement in the sow, lasts 48–72 h and occurs approximately every 21 days with ovulation occurring in the latter part of the period. Female swine shed an average of 15–18 ova at each estrous period, although only 60–70% survive the gestation period of 114 days. With gilts (swine that have yet to produce a litter) the number of ova ovulated may be increased by feeding them a high-energy ration for 2–3 weeks before breeding. Once a gilt is bred, however, embryonic mortality is reduced by limiting energy intake during gestation. Maternal weight gain of 70 lb (32 kg) for gilts and 44 lb (20 kg) for sows (post–first litter females) during gestation is sufficient to produce a healthy litter of pigs and to form a nutrient reserve for lactation.

Litter size is affected by many factors, and averages about 9.8 pigs per sow in the United States breeds. Certain breeds in China and other Asian countries may average 14–16. The number of pigs per litter at birth usually increases up to the third litter, or when the sow is $2\frac{1}{2}$–3 years of age.

The majority of the sows in the United States are bred by natural service. Use of artificial insemination is increasing, however, and allows wider distribution of genetic material to the industry. Embryo transfer is especially effective for introducing new genetic material into a herd and eliminates the risks of possible disease introduction when intact animals are brought into a herd.



Typical example of a modern meat-type hog.

Nutrition. Cost of feed represents 50–75% of the total cost of production in commercial swine units. Diets used in feeding programs in modern swine production units are scientifically formulated to ensure nutrition that is adequate for maximum rate and efficiency of weight gain and for quality of product. In highly mechanized units, mixing and distribution to feeders of diets are computer-controlled.

The most critical nutritional period is during the first few weeks of the pig's life. Needs during the other periods of the life cycle—gestation, lactation, growing, and finishing—are easily met by fortifying natural feedstuffs. Water, energy (derived from carbohydrates, fats, and proteins), amino acids (protein), minerals, and vitamins are essential.

Water serves as a medium for digestion, absorption, transportation, and excretion of other nutrients. Dietary requirements decrease from about 12% of the liveweight for young swine to 5% for mature swine (about 2 kg of water per kilogram of feed consumed).

Most of the energy need is supplied by dietary carbohydrate. Fat usually plays a minor role, but supplemental fat is frequently in diets for very young pigs and for gestating (day 100 to parturition) and lactating sows. In general, carbohydrates such as starch, dextrins, disaccharides (except lactose), and monosaccharides are equally metabolizable by a weanling pig. Lactose is an excellent energy source

TABLE 1. Energy needs of swine*

Live weight, lb (kg)	Daily gain, lb (kg)	Daily feed, lb (kg)	Daily metabolizable energy, kcal (kJ)
2–11 (1–5)	0.44 (0.20)	0.55 (0.25)	900 (3800)
11–22 (5–10)	0.66 (0.30)	1.10 (0.50)	1700 (7100)
22–44 (10–20)	1.10 (0.50)	2.20 (1.00)	3160 (13,200)
44–77 (20–35)	1.32 (0.60)	3.30 (1.50)	4740 (19,800)
77–132 (35–60)	1.54 (0.70)	4.40 (2.00)	6320 (25,400)
132–220 (60–100)	1.76 (0.80)	6.60 (3.00)	9480 (39,700)

*Daily needs will vary due to environmental and management effects.

TABLE 2. Essential amino acids for growing-finishing swine

Amino acid	Content, %	
	Maize	Soybean meal*
Arginine†	0.52	3.40
Histidine	0.19	1.10
Isoleucine	0.37	2.50
Leucine	1.00	3.40
Lysine	0.22	2.90
Methionine‡	0.17	0.64
Phenylalanine§	0.44	2.20
Threonine	0.34	1.70
Tryptophan	0.09	0.70
Valine	0.42	2.40

*Remainder after the oil removed from soybeans by the solvent extraction method.
†Is not essential for satisfactory reproduction.
‡Cystine can supply up to 56% of the pig's methionine requirement.
§Tyrosine can satisfy 50% of the requirement for phenylalanine.

for the baby pig, but may cause diarrhea in older pigs because of a developmental shift with increasing age in intestinal enzyme activities. Energy requirements are usually expressed as a daily need for total feed or digestible nutrients. Feed intake increases as nutrient requirements for maintenance and weight gain increase (**Table 1**).

Amino acids required for maintenance, growth, gestation, and lactation, are chemically linked as proteins in feedstuffs. An essential amino acid is one that swine cannot synthesize at a sufficiently rapid rate to permit normal growth and activity. Ten of the amino acids in feedstuffs normally fed are classified as essential for growing pigs (**Table 2**). Lysine and tryptophan are of most practical importance since most cereal grains (maize, sorghum, wheat) have low levels. Soybean meal is the most common dietary supplement used to correct the amino acid deficiencies in cereal grains.

Weanling pigs have special dietary requirements (**Table 3**). Cereal grains and soybean meal lack adequate quantities of vitamins and minerals; hence, practical diets normally are supplemented accordingly.

Calcium, phosphorus, sodium, and chloride are normally added as ground limestone, dicalcium phosphate or defluorinated rock phosphate, and trace mineralized salt. The salt serves as a carrier for trace minerals, including iron, copper, manganese, selenium, iodine, and zinc. The trace minerals may not always be dietary essentials since they may be obtained from extradietary sources. Supplemental vitamins are most economically provided as synthetically produced vitamin supplements (**Table 4**).

Growth promotants (antibiotics and other antimicrobials) are regularly used in pig diets. Usually, the younger the pig, the greater the growth rate response. The amounts used in diets for finishing pigs and gestating and lactating sows should reflect the general health status of the herd. Supplements containing antimicrobials (such as chlortetracycline, oxytetracycline, penicillin, tylosin) are usually added to provide 10–20 ppm in the diet; copper sulfate is also often used in the diet as an antimicrobial. The specific action of these feed additives has not been completely delineated, but in general, herd health has been improved and the pigs perform more nearly to their genetic potential.

Most pig diets are mixtures (Table 4), and usually are self-fed to growing-finishing pigs. This ensures a "balanced" intake of nutrients by the pig. Gestating females are limit-fed to minimize excess body weight gain. Lactating sows are usually fed liberal amounts to provide adequate nutrition for full milk production. *See* ANIMAL FEEDS.

Management. The aim of swine management is to minimize adverse environmental factors and provided maximum opportunity for survival and growth. Mortality of piglets may reach 25% or higher within the first three days after birth, usually due to crushing by the sow, enteric infections, or starvation. Modern climatically regulated farrowing quarters frequently have farrowing crates to help protect the newborn.

A newborn piglet is incapable of maintaining body temperature, and is therefore dependent on supplemental heat from the environment. Other biological management practices for the newborn piglet include clipping needle teeth, applying antiseptic to the navel, and providing iron to compensate for its severe deficiency in milk.

Since the pig has very limited ability to dissipate body heat by sweating, finishing pigs or breeding stock must be protected during periods of high temperature. Shades and water mists have been used in the past, but modern confinement buildings for

TABLE 3. Mineral and vitamin needs for weanling pigs (33–77 lb or 15–35 kg)*

Mineral	Amount	Vitamin	Amount
Calcium, %	0.60	Vitamin A, IU/lb (IU/kg)	1500 (3300)
Phosphorus, %	0.50	Vitamin D, IU/lb (IU/kg)	150 (330)
Sodium chloride, %	0.35	Vitamin E, IU/lb (IU/kg)	5 (11)
Copper, ppm	4	Vitamin K, ppm	2.2
Iron, ppm	80	Riboflavin, ppm	2.2
Iodine, ppm	0.14	Niacin, ppm	17.8
Manganese, ppm	3	Pantothenic acid, ppm	11
Zinc, ppm	60	Choline, ppm	890
Selenium, ppm	0.15	Vitamin B_{12}, ppb	15

*IU = International Unit; ppm = parts per million; ppb = parts per billion.

TABLE 4. Typical ration for growing pigs

Ingredient	Amount, %
Yellow corn	76.45
Soybean oil meal	21.00
Dicalcium phosphate	1.25
Ground limestone	0.75
Trace mineralized salt	0.35
Vitamin supplement*	0.10
Antibiotic supplement	0.10

*Supplies per kilogram of ration: 3333 IU vitamin A; 333 IU vitamin D; 22 IU vitamin E; 1.1 mg riboflavin; 5.5 mg pantothenic acid; 16.5 mg niacin; 110 mg choline; 17.8 microgram vitamin B_{12}.

finishing hogs and breeding stock minimize environmental extremes. The reduction of temperature extremes encourages more rapid and more efficient gains, particularly in finishing swine.

Diseases. Swine are particularly susceptible to diseases and parasites. Although cholera was once a major problem, it has been eliminated from the United States and some other countries. Of major concern in recent years is African swine fever, which is similar in effect to cholera, but there is as yet no adequately effective vaccine or curative. Erysipelas, dysentery, leptospirosis, atrophic rhinitis, pseudorabies, and transmissible gastroenteritis are some of the other important diseases. Several diseases can be prevented by use of appropriate vaccines or other medication, but specific control has not been realized for many diseases.

Swine are especially subject to infestations with ascarids, which are parasites primarily of the intestinal tract that impair metabolism and growth. Following ingestion of the embryonated ascarid eggs by the pig, the larvae burrow through the intestinal wall and migrate through the body causing considerable damage to the liver and lungs. Passing from the lungs to the mouth, the larvae are ingested and develop into mature ascarids in the intestine. Such parasites are particularly prominent in tropical areas.

The use of confinement facilities, particularly those with slotted floors, reduces the spread of disease and parasites and allows rapid removal of excreta from the pens, thereby reducing opportunities for contact with the many vectors of the diseases. Excellent sanitation is part of any sound biological management system.

Marketing. In the United States, commercially produced pigs should attain a market weight of 198–220 lb (90–100 kg) at $5^1/_2$–6 months of age. In European countries where restricted-feed programs are used, pigs of the same size will be considerably older. About 87% of the market pigs in the United States are barrows (male pigs that were castrated at a young age) and gilts (a female that has not borne young), and the remainder are sows, boars, and stags (males castrated after becoming sexually mature). The market hogs move from farm to the slaughter plants through terminal public markets, country buying stations, and auction markets or by direct sale to the packer.

Market grades of slaughter barrows and gilts have been developed according to a scheme which places a premium on the four major lean cuts: ham, loin, picnic, and Boston butt. In addition, market grades penalize excess fat in the carcass. Barrow and gilt carcasses are graded as U.S. No. 1, U.S. No. 2, U.S. No. 3, U.S. No. 4, and utility. The utility grade is characterized by a low degree of finish. In 1980, 96% were graded 1 and 2 while in 1968 it was 50%. Thus, the swine producers, by careful selection, breeding, feeding, and management, have changed to a type of pig that provides products more acceptable to the consumers.

Aldon H. Jensen

Bibliography. S. Baxter, *Intensive Pig Production*, 1984; J. L. Howard, *Current Veterinary Therapy: Food Animal Practice*, 1981; W. G. Pond and J. H. Maner, *Swine Production and Nutrition*, 1984; R. N. Van Arsdall and K. E. Nelson, *U.S. Hog Industry*, USDA Agr. Econ. Rep. 511, 1984.

Switched capacitor circuit

A module consisting of a capacitor with two metal oxide semiconductor (MOS) switches connected as shown in **Fig. 1***a*. These elements in the module are easily realized as an integrated circuit on a silicon chip by using MOS technology. The switched capacitor module is approximately equivalent to a resistor, as shown in Fig. 1*b*. The fact that resistors are relatively difficult to implement gives the switched capacitor a great advantage in integrated-circuit applications requiring resistors. Some of the advantages are that the cost is significantly reduced,

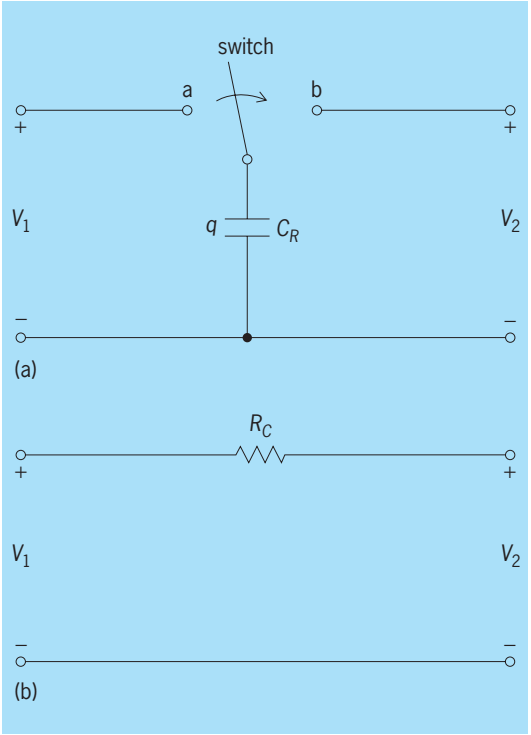


Fig. 1. Switched capacitor. (*a*) Basic circuit. (*b*) Equivalent resistive circuit.

the chip area needed is reduced, and precision is increased. Although the switched capacitor can be used for any analog circuit realization such as analog-to-digital or digital-to-analog converters, the most notable application has been to voice-frequency filtering. *See* ANALOG-TO-DIGITAL CONVERTER; DIGITAL-TO-ANALOG CONVERTER.

Although a switch has long been used as an element in circuits and systems, it was not until the late 1970s that its potential and practicality in integrated-circuit design was realized.

Integrator circuits. A conventional *RC* (resistance-capacitance) integrator circuit is shown in **Fig. 2***a*. The output voltage is given by Eq. (1), where $1/s$ in-

$$V_{\text{out}} = \frac{1}{R_1 C_2}\frac{1}{s}V_{\text{in}} \tag{1}$$

dicates the operation of integration (s is the differential operator), showing that integrator performance depends on R_1. In MOS integrated circuits, the value of R_1 cannot be controlled better than 20% by using standard fabrication techniques. In addition, considerable chip space is required to realize resistors in the megohm range. In contrast, the switched capacitor realizations shown in Fig. 2*b* and *c* depend on the ratio of capacitors which can be controlled with great accuracy. For example, the output voltage of the differential integrator shown in Fig. 2*c* is given by Eq. (2), where f_c is the clock frequency. With C_1 and

$$V_{\text{out}} = \frac{C_1 f_c}{C_2 s}(V_2 - V_1) \tag{2}$$

C_2 in the range of 1 picofarad and f_c at 100 kHz, this circuit has an equivalent resistance of 10 megohms, and the gain of the circuit is about 10^4. The silicon chip area required to implement the capacitors is about 0.01 mm^2. If a resistor is used in place of the switched capacitors, an area at least 100 times larger would be required.

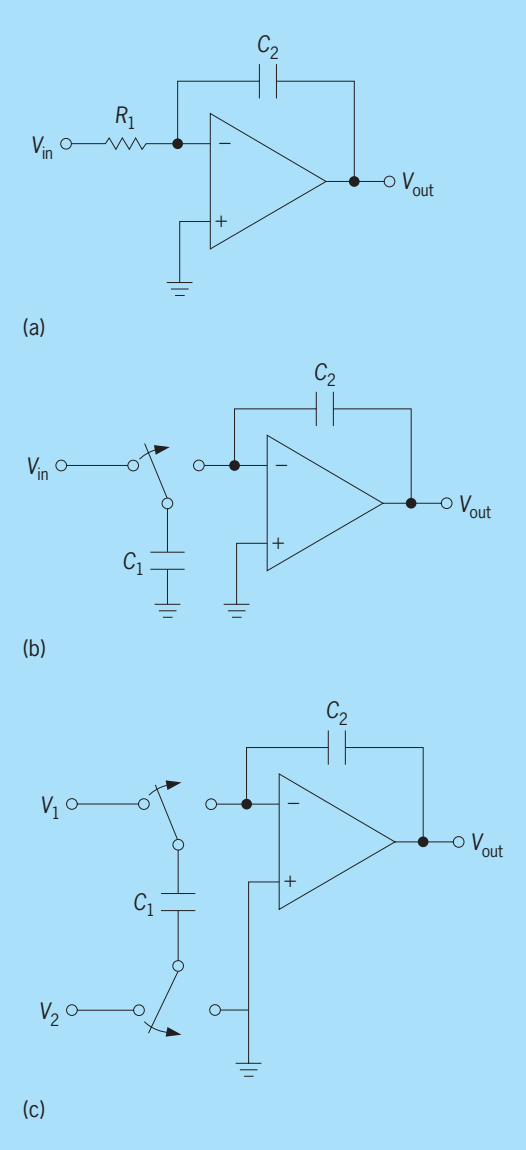


Fig. 2. Integrator circuits. (*a*) Conventional *RC* integrator circuit. (*b*) Single-input switched-capacitor integrator. (*c*) Switched-capacitor differential integrator circuit.

Equivalent resistance. Returning to the switched-capacitor circuit of Fig. 1*a*, the operation may be visualized as follows. With the switch in position *a*, the capacitor C_R is charged to the voltage V_1. The switch is then thrown to position *b*, and the capacitor discharged at voltage V_2. The amount of charge transferred is then $q = C(V_2 - V_1)$. If the switch is thrown back and forth at a clock frequency f_c, the average current will be $C(V_2 - V_1) f_c$. The size of an equivalent resistor to give the same value of current is given by Eq. (3). From this equation, it is seen that

$$R_C = \frac{1}{Cf_c} \tag{3}$$

with $C = 1$ pF and $f_c = 100$ kHz, the value of 10 megohms used previously is obtained.

The accuracy of the equivalence between the switched capacitor and the resistor depends on the relative size of the clock frequency f_c and the frequencies in the signal being processed. If the switching frequency is much larger than the signal frequencies of interest, the equivalence is excellent and the time sampling of the signal can be ignored in a first-order analysis, such that the switched capacitor is a direct replacement for a conventional resistor. If this is not the case, then sampled-data techniques in terms of a z-transform variable must be used for accuracy.

Analog operations. The switch that has been used in describing the switched capacitor is actually realized by an MOS transistor to which a pulse of voltage at the clock frequency is applied to produce the off and on conditions of the switch. This periodically operating switch is used for a number of analog operations, such as addition, subtraction, inversion, and integration. These operations are essential in the construction of analog filters, as well as in other applications of switched capacitors. These operations may be explained in terms of the circuits of Fig. 2. In Fig. 2*a* and *b* the analog operation of integration is accomplished. In addition, these circuits are of the inverting type, meaning that a sign reversal is

accomplished in addition to integration. The sign reversal of a voltage can be accomplished directly by using switches and a capacitor, as seen in Fig. 2*c*. Assume that V_1 is grounded or $V_1 = 0$. The operation of the switches is such that the voltage applied to the MOS operational amplifier is the negative of V_2. With the switch operating from left to right, V_2 with respect to ground is reversed. With V_1 not grounded, the circuit of Fig. 2*c* is a differential integrator, meaning that the output voltage is a function of the voltage difference, $V_2 - V_1$. In conventional active-filter design, these analog operations are accomplished by means of additional stages incorporating operational amplifiers. In switched-capacitor design, these analog operations are implemented with switches. *See* AMPLIFIER; ANALOG COMPUTER; TRANSISTOR.

Filter design. Although there are many strategies for filter design, the discussion will be restricted to the case of filters based on the passive *LC* (inductance-capacitance) ladder with resistive terminations at both ends. Extensive tables are available giving element values to achieve various forms of frequency response, such as Butterworth, Chebyshev, and Cauer (elliptic). All tables are given in terms of a normalized termination of 1 ohm, and a normalized frequency of $\omega_o = 1$ rad/s and for the low-pass case. It is standard procedure to make use of frequency transformations to realize high-pass, band-pass, band-elimination, and similar kinds of responses, and to use frequency and magnitude scaling to give practical element values. The passive ladder structure with double terminations is chosen because it has low sensitivity of changes in transmission with changes in element sizes.

Starting with the low-sensitivity, low-pass ladder structure, a frequency transformation is first accomplished. From these steps, a structural simulation is then carried out by replacing the actual filter by its signal flow graph representation. The flow graph is chosen so that most of the operations required are integration. The elements in the flow graph are then simulated by circuits like those shown in Fig. 2*c*.

An example of filter design is shown in **Fig. 3**. The ladder network shown in Fig. 3*a* is known as the low-pass prototype. In the usual case, $R_1 = R_2 = 1$ ohm. The elements C_1 and L_2 are determined from tables, depending on the form of frequency response required. Going from Fig. 3*a* to *b* accomplishes a low-pass to band-pass transformation in which all element values are determined from those in Fig. 3*a*, and the specification of the center frequency and bandwidth of the band-pass case. In Fig. 3*c* the filter of Fig. 3*b* is represented by its flow graph, in which the lines and arcs with arrows indicate the structure of the circuit of Fig. 3*b*, and the associated symbols represent the impedance or admittance. To this structural simulation of the ladder filter, an element simulation is next applied. All elements of a form such as $1/RC_A s$ are realized by using the integrator of Fig. 2*c* with differences of voltages accomplished by the switched capacitors. The final result, shown in Fig. 3*d*, is then implemented as an integrated circuit containing only switched capacitors, ordinary capacitors, and operational amplifiers. The chip area required to realize a filter of modest order might be 100 mils (2.5 mm) on each side. *See* ELECTRIC FILTER; INTEGRATED CIRCUITS.

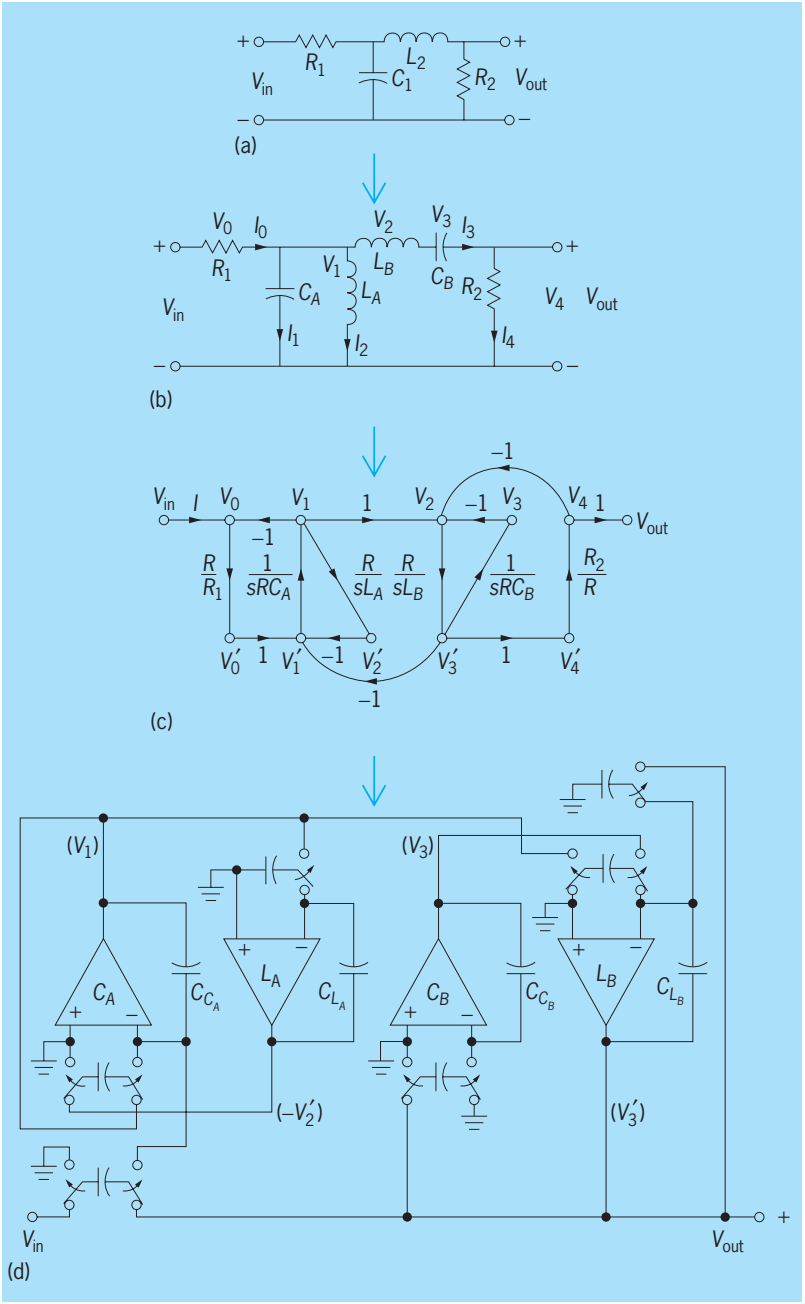


Fig. 3. Steps in the realization of a switched-capacitor filter. (*a*) Low-pass prototype filter. (*b*) Corresponding band-pass filter. (*c*) Signal flow graph representation of the circuit of *b*. (*d*) Final switched-capacitor band-pass filter.

M. E. Van Valkenburg

Bibliography. P. E. Allen and E. S. Sinencio, *Switched Capacitor Circuits*, 1984; H. Baher, *Selective Linear-Phase Switched-Capacitor and Digital Filters*, 1993; M. Ghausi and K. Laker, *Modern Filter Design: Active RC and Switched Capacitor*, 1981; G. M. Moschytz (ed.), *MOS Switched-Capacitor Filters: Analysis and Design*, 1984; R. Ubehauen and A. Cichocki, *MOS Switched-Capacitor and Continuous Time Integrated Circuits and Systems*, 1989.

Switching circuit

A constituent electric circuit of a switching or digital data-processing system which receives, stores, or manipulates information in coded form to accomplish the specified objectives of the system. Examples include digital computers, dial telephone systems, and automatic accounting and inventory systems. *See* DIGITAL COMPUTER; SWITCHING SYSTEMS (COMMUNICATIONS); SWITCHING THEORY.

Physically, switching circuits consist of conducting paths interconnecting discrete-valued electrical devices. The most generally used switching circuit devices are two-valued or binary, such as switches and relays in which manual or electromagnetic actuation opens and closes electric contacts; vacuum and gas-filled electronic tubes, and semiconductor rectifiers and transistors, which do or do not conduct current; and magnetic structures, which can be saturated in either of two directions.

The electrical conditions controlling these switching circuit devices are also generally two-valued or binary, such as open versus closed path, full voltage versus no voltage, large current versus small current, and high resistance versus low resistance. Such two-valued electrical conditions, as applied to the input of a switching circuit, represent either (1) a combination of events or situations which exist or do not exist; (2) a sequence of events or situations which occur in a certain order; or (3) both combinations and sequences of events or situations. The switching circuit responds to such inputs by delivering at its output, also in two-valued terms, new information which is functionally related to the input information.

Electronic switching circuits are characterized by these functional relationships between input and output information, as well as by such attributes as response time, power consumption, packaging densities, and other performance parameters.

Functional characteristics. Functional characteristics of switching circuits are defined by the logical operation and memory capabilities of the discrete devices from which they are assembled, as well as by the means used to interconnect the devices.

For example, switching circuits embody such logical relationships as output X is to exist only if input A and B occur simultaneously; and output Y is to exist if either input A or input B occurs. The factor of memory, in turn, enables a switching circuit to hold or retain a given state after the condition that produced the state has passed.

Even in large and complex switching systems, the majority of circuit requirements can be met by a relatively small number of types of circuits, each of which performs one or a limited number of somewhat distinct functions. These functional circuits, some examples of which are described below, are the basic building blocks of a switching system.

Basic combinational circuits. A combinational switching circuit is one in which a particular set of input conditions always establishes the same output, irrespective of the history of the circuit. An example involving a simple combinational circuit is the controlling of the entrance-hall light of a residence by three up-down wall switches located in three different rooms. Analysis of this problem shows that the circuit must meet the following simple requirements. If any one or all three wall switches are down, the hall lamp must light; if any one or all three switches are up, the lamp must be dark. An obvious (but not the most efficient) circuit meeting these requirements is shown in **Fig. 1**. In this problem the circuit inputs are, of course, the manual switch settings, and the circuit output is the control of the light.

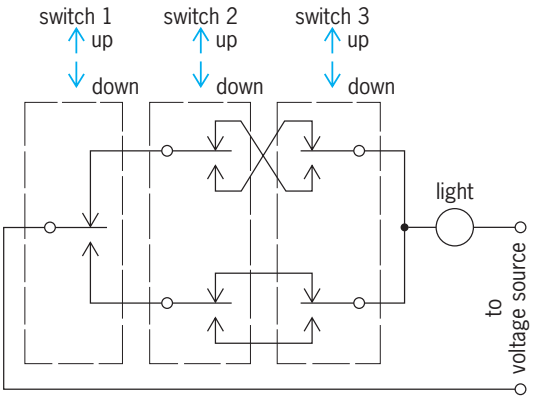


Fig. 1. Elementary combinational switching circuit.

In electronic switching circuits, so-called gates are used to perform logical functions equivalent to these series-parallel networks of switch contacts. In this sense, an electronic gate is an elementary combinational circuit. Gates do not function by physically inserting or removing metallic conduction paths between contacts of manually operated switches or remotely controlled relays. Instead, they function by control of voltage or current levels at their output.

The most commonly encountered gates are the AND and the OR gates. The AND gate produces an output only if all its inputs are concurrently present; an OR gate produces an output if any one or any combination of its inputs is present. **Figure 2** shows both an AND gate and an OR gate, using rectifier or diode elements. *See* LOGIC CIRCUITS.

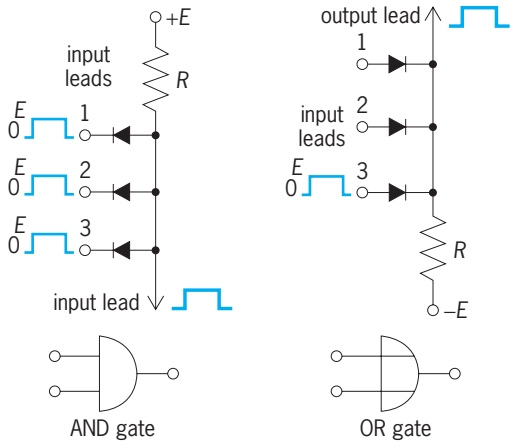


Fig. 2. Typical switching gates using crystal diodes.

In the AND gate the rectifiers are so oriented that current from a positive voltage source E passes through the relatively large resistance R and then through the low forward resistance of any one of the rectifiers to ground in the circuits controlling the gate. Thus, in the inactive state of the gate, the output lead is at or near ground potential. If all three input leads of this gate concurrently receive a positive voltage pulse of magnitude E, the rectifiers approach open circuit, and the output lead will be raised from near ground to a positive potential for the duration of the input pulse. In other words, input leads 1 and 2 and 3 must all receive the positive pulse to obtain the positive output voltage.

In the OR gate the rectifiers are reversed so that current flows from ground in the input circuits through the low forward resistance of any rectifier and then through the relatively large resistance R to the negative voltage source $-E$. Thus, in the inactive state of the gate the output lead is at or near ground potential. If, however, a relatively high positive voltage pulse is applied to input leads 1 or 2 or 3, the remaining two diodes are cut off and the output is raised to a positive potential for the duration of the input pulse.

Gates may, of course, be constructed with other electronic devices, such as tubes, transistors, and magnetic cores.

Basic sequential circuits. A sequential switching circuit is one whose output depends not only upon the present state of its input, but also on what its input conditions have been in the past. Sequential circuits, therefore, require memory elements.

By way of illustration, consider the following simple sequential circuit problem. When a telephone customer lifts the handset, a lamp is to light in front of a switchboard operator. When an operator answers, the light should go out to avoid other operators also answering. After satisfying the customer's request for a connection, the operator withdraws. The light, however, should not relight now, even though the conditions existing at this time are seemingly identical with those at the start; that is, the customer has the handset lifted and no operator is on the line. A sequential relay circuit meeting these simple requirements is shown in **Fig. 3**. In this circuit, when the hand is lifted, the handset off-hook switch connects

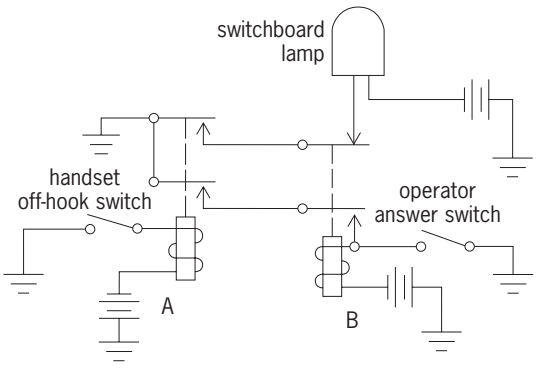


Fig. 3. Elementary sequential relay switching circuit.

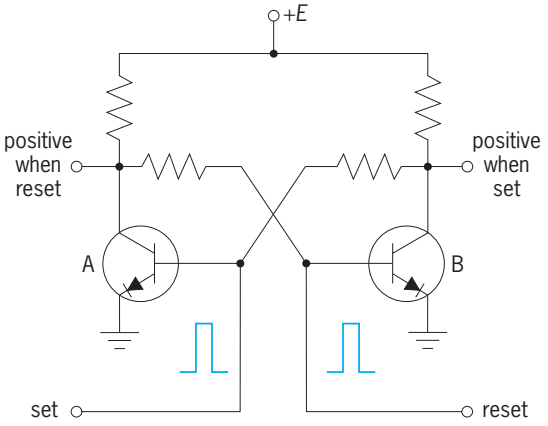


Fig. 4. Transistor switching memory element (flip-flop).

a ground input to relay A which operates and lights the switchboard lamp. When the operator answers, another ground input operates relay B and this relay puts out the light. A holding circuit on relay B keeps relay B operated until the handset off-hook switch is again opened and relay A is deenergized. Relay B "remembers" that the operator has answered and prevents the relighting of the lamp when the operator withdraws. It is, therefore, the memory element of the circuit.

A typical electronic memory element used in sequential circuits is a simple circuit called a flip-flop. A flip-flop consists of two amplifiers connected so that the output of one amplifier is the input of the other. A voltage pulse will set the flip-flop into one of two states, and that state remains until another voltage pulse resets the flip-flop or returns it to its original state. It can therefore be used to remember that an event has taken place.

Figure 4 is an *npn*-transistor flip-flop. When set, transistor A is conducting and transistor B is cut off. When reset, transistor B is conducting and transistor A is cut off. A positive output voltage with respect to ground may be obtained from either transistor to indicate the condition of the flip-flop. *See* TRANSISTOR.

Relays, flip-flops, and similar memory elements provide static, or fixed, memory; they hold the stored information indefinitely, or until they are told to "forget," commonly called "resetting." In contrast, a delay line provides transient memory. A delay line has the property that an electrical signal applied to its input is delayed on its way to the output.

Selecting circuits. A selecting circuit receives the identity (called the address) of a particular item and selects that item from among a number of similar ones. The selectable items are often represented by terminals or leads. Selection usually involves marking the specified terminal or lead by applying to it some electrical condition, such as a voltage or current pulse, or a steady-state dc signal. By means of this electrical condition, the selected circuit is alerted, sized, or controlled.

An electronic selecting circuit using AND gates is the matrix type (**Fig. 5**). In this type of circuit an input signal appears on one of the horizontal input

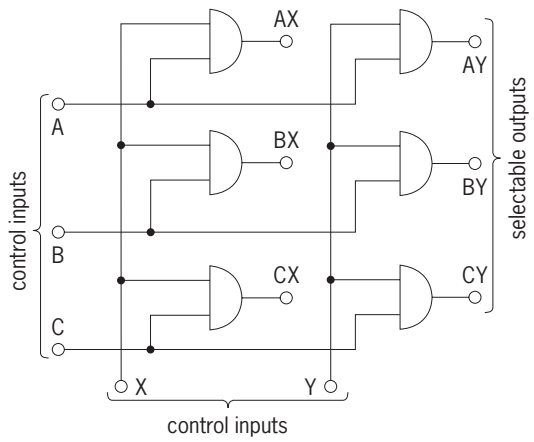


Fig. 5. Matrix selecting circuit using AND gates.

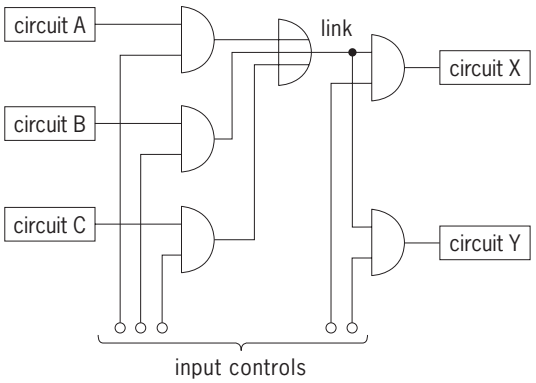


Fig. 6. Connecting circuit using AND and OR gates.

leads and concurrently on one of the vertical input leads. The selected output is at the cross point of these two leads.

Connecting circuits. A switching system is an aggregate of functional circuit units, some of which must sometimes be directly coupled to each other to interchange information. **Figure 6** shows a simple electronic connecting circuit using AND and OR gates. In this arrangement a communication path is provided over a single link from any one of the three functional circuits A, B, C, to either the X or Y circuit by an external control circuit activating the appropriate pair of AND gates. To provide a multilead link, or to provide for other simultaneous interconnections, additional AND gates would, of course, be required. The OR gate maintains separation of the inputs at the common junction point.

Lockout circuits. In switching systems, situations often arise where several similar circuit units are ready at the same instant to request collaboration with another type of functional circuit. Mutual interference among the requesting circuits is prevented by the lockout circuit (sometimes referred to as hunting or finding circuits). In response to concurrent inputs from a number of external circuits, a lockout circuit provides an output indication corresponding to one, and only one, of these circuits at any time.

Figure 7 shows a typical electronic lockout circuit using cold-cathode gas-filled tubes. The external circuits furnish positive potential on the input leads to the control gaps of the tubes as indications of service requests. The operation of the circuit is based on the dynamic negative-resistance characteristics of gas tubes. If such tubes are provided with a common impedance in their conduction paths (the cathode impedance in this circuit), simultaneous input signals will result in the ionization of only one tube. Once the control gap of a tube is ionized, conduction current starts flowing in its main gap and this current through the common impedance instantaneously reduces the voltage across all the other tubes below the value needed to ionize them. This reduced voltage is, however, adequate to keep the single ionized tube in the conducting state until its conduction path is opened. The identity of the particular ionized tube is derived from the anode resistance individual to each tube; the output lead whose potential has been lowered by this resistance represents the circuit whose request has been granted.

Translating circuits. Switching systems process information in coded form; the information is generally in the form of numbers. Numerical codes are many and varied, each with its own characteristics and more or less distinct advantages for different switching circuit situations. Therefore, one of the common functional circuits in switching systems is the translating circuit, which translates information received in one code into the same information expressed in another code. These translating circuits are combinational circuits; a given input signal combination representing a code to be translated always produces the same output signals, which represent the desired code.

Figure 8 is an example of a magnetic-core translating circuit that translates from binary code (1,2,4) to a one out of eight code (0,1,2, . . . ,6,7). The circuit has three flip-flops which are set or reset (not set) according to the binary input code combination. The translating elements are eight magnetic cores, each with five windings, and are represented in

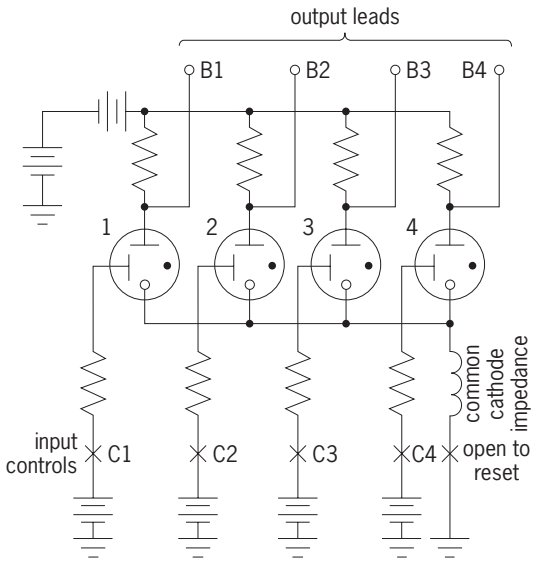
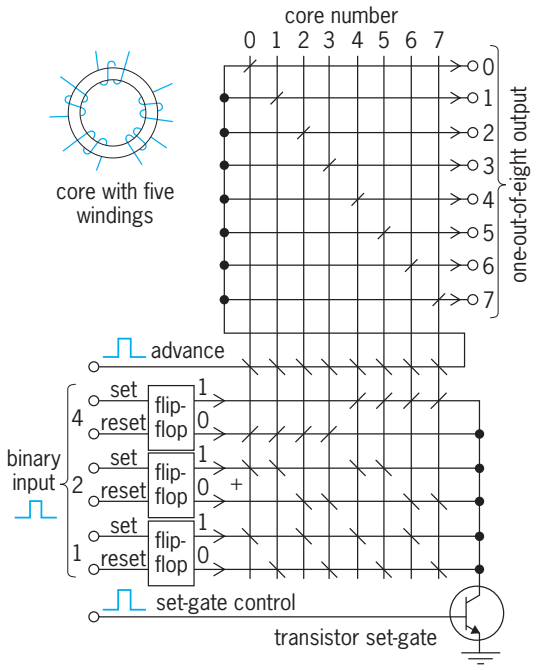


Fig. 7. Lockout circuit using cold-cathode gas tubes.



Output digit desired	Flip-flop 4		Flip-flop 2		Flip-flop 1	
	Set	Reset	Set	Reset	Set	Reset
0		x		x		x
1		x		x	x	
2		x	x			x
3		x	x		x	
4	x			x		x
5	x			x	x	
6	x		x			x
7	x		x		x	

Fig. 8. Code register and translating circuit using magnetic cores, each of which has five windings.

Fig. 8 by a vertical line. Each short, slanting line segment represents a separate winding on a core. These slanting lines also symbolize a mirror action; an input current pulse coming from a flip-flop sets a core if it is reflected upward by the mirror, and prevents setting or resets the core if reflected downward. Once set, the subsequent resetting of a core induces a current which flows upward in the vertical line (in a direction opposite to the resetting current) and is reflected to the left or to the right by each mirror symbol.

With this explanation of the symbolism, the circuit works as follows. The input is binary; that is, it consists of a positive voltage pulse to each of the three flip-flops either on its set or on its reset input lead, according to the **table**. (By adding the numerical designations to those flip-flops which are set in a particular combination, the value of the output digit is determined.)

While the flip-flops are being set, their output current is prevented from flowing into the core windings by the transistor set-gate which is normally nonconducting. Shortly after the binary input combination is recorded in the flip-flops, this set-gate is pulsed for a moment into its conducting state. During this moment, output current will flow from each flip-flop in its "1" output lead (if the flip-flop has been set) or in its "0" output lead (if the flip-flop has been reset). As Fig. 8 shows, the output current of flip-flop 4 is always used to set the cores; that is, the current in the "0" output lead of this flip-flop is used to magnetize the first four cores in the set direction, or the current in its "1" output lead is used to magnetize the last four cores in the set direction. In contrast, the output currents from flip-flops 2 and 1 are always used to magnetize the cores in the opposite or reset direction. Initially, all cores are in the reset condition, and cores that receive both set and reset currents simultaneously will not change this initial condition. An analysis of Fig. 8 will therefore show that, for any desired digit, one and only one of the eight cores will be set by the flip-flops in combination. For instance, if output 3 is desired, the current from flip-flop 4 tends to set cores 0, 1, 2, and 3, but cores 0, 1, and 2 are prevented from being set by the output current from either or both flip-flops 2 and 1. When the translated code is needed, the current pulse on the advance lead resets the single previously set core, and consequently an induced output current pulse appears on the appropriate output lead. (The rectifiers in the input and output portions of the circuit prevent unwanted reverse current.)

Register circuits. Information received by a switching system is not always used immediately. It must be stored in register circuits for future use.

In a register circuit the coded information to be stored is applied as input and retained by memory elements of the circuit, and when needed, the registered information is taken as output in the same code or in a different code. Figure 8 embodies a register function as well as a translating function. Register circuits are devised with a great variety of memory elements and have capacities to store from a few to millions of information bits.

A frequently encountered form of register circuit is the shift register. This type of register has the ability to shift its stored digital information internally to positions representing higher or lower numerical values in the code employed. For example, in decimal code registration a digit may be shifted from the units to the tens position. An obvious use of such registers is in digital computers when, for example, partial multiplication products have to be lined up for addition.

Counting circuits. One of the most frequently encountered circuits in switching systems is the counting circuit whose function, in general, is to detect and count repeated current or voltage pulses which represent incoming information.

Performance characteristics. In theory, functionally equivalent switching circuits can be designed by using any available switching technology; that is, mechanical manual or relay-operated switches, or a number of electronic devices including a wide variety of semiconductors, and vacuum or gas-filled tubes. The selection of a particular technology for use in computer systems is often determined by the time required to perform logical operations. The

Digital integrated circuit families

Logic type	Gate delay (t_d), ns	Power dissipation (P_D), mW/gate	Density, gates/mm^2 (gates/in.^{2})
I. Bipolar—use bipolar junction transistors as basic switching elements			
A. Transistor-transistor logic (TTL)			
1. Standard small-scale integration/medium-scale-integration (SSI/MSI) TTL	5–10	10	20–80 (13,000–50,000)
2. Low-power Schottky (LPS)	5–10	2	20–80 (13,000–50,000)
3. Advanced LPS for large-scale integration (LSI)*	1–2	1–2	200–600 (130,000–400,000)
B. Emitter-coupled logic (ECL)			
1. Standard SSI/MSI ECL	0.5–1	25	20–50 (13,000–32,000)
2. Advanced ECL for LSI*	0.7–1	10	100–200 (65,000–130,000)
3. Current-mode logic (CML)†	0.5–1	0.2–1	200–300 (130,000–200,000)
C. Integrated injection logic (I^2L)			
1. Standard LSI I^2L*	2–5	0.05	200–400 (130,000–250,000)
2. Integrated Schottky logic (ISL)†	2	0.3	500–1000 (320,000–650,000)
3. Schottky transistor logic (STL)†	2	0.3	500–1000 (320,000–650,000)
4. Implanted advanced composed technology (impact)†	1–2	0.1	800 (500,000)
II. Metal oxide semiconductor (MOS)—use MOS field-effect transistors (MOSFETs) as basic switching elements			
A. *P*-channel MOS (PMOS)			
Standard LSI PMOS	50	1	100–200 (65,000–130,000)
B. *N*-channel MOS (NMOS)			
1. Standard LSI NMOS	10–20	1–2	200–400 (130,000–250,000)
2. Very large-scale-integration (VLSI) NMOS	0.5–1	0.5	1500–3000 (1,000,000–2,000,000)
3. Exotic structures			
a. Vertical channel (VMOS)	0.5–1	0.5	500–1000 (320,000–650,000)
b. Double-diffused (DMOS)	0.5–1	0.5	500–1000 (320,000–650,000)
C. Complementary MOS (CMOS)			
1. Standard SSI/MSI CMOS	20–50	0.01–0.1	50–100 (32,000–65,000)
2. LSI*/VLSI CMOS	0.5–1	0.01–0.5	500–2000 (320,000–1,300,000)
3. Silicon-on-sapphire (CMOS/SOS)	0.5–1	0.01–0.5	750–2500 (500,000–1,600,000)

*Uses advanced fabrication technology with feature sizes below 5 μm.
†Feature sizes of 2.5 μm or smaller.

maximum number of logical operations per second which a computer switching circuit can execute is determined by both the time required by the switching device to change state (gate delay, t_d) and the time it takes signals to propagate between switching circuits, the latter being proportional to the distance between the circuits. Device gate delay is dependent upon the active element's principle of operation, and is normally inversely proportional to gate power dissipation (P_D) since larger current supplied to the switching devices reduces the effects of junction storage time, capacitance, and parasitic elements. The product of t_d and P_D is a popular figure of merit for comparing different device technologies.

Circuit density. One of the objectives of large-scale integrated-circuit (LSI) and very large-scale integrated-circuit (VLSI) design is to incorporate fast device technologies into compact packages which reduce, to the greatest extent possible, propagation delay between circuits. Thus, circuit density, in terms of the number of gates per square millimeter, is another important parameter affecting overall logic speed. Thin-film fabrication techniques for both active and passive components result in dense packaging with higher reliability, finer adjustment of component values, and simpler manufacturing processes. The table lists the major semiconductor switching circuit technologies, along with their acronyms and important performance characteristics. *See* INTEGRATED CIRCUITS.

Josephson junctions. Since semiconductors produce heat, dense circuit packaging is hampered by the need to remove heat. The Josephson junction avoids this problem by using superconducting circuits, that is, circuits cooled to near absolute zero where the electrical resistance that generates heat virtually vanishes. Josephson found that a superconducting current between two wires separated by an ultrathin insulating film can be controlled by a magnetic field applied to the insulator. While the heat dissipated is very small (5 nanowatts per gate), gate delays of less than 0.01 nonosecond in miniaturized devices a few micrometers in length and width can be achieved. Implementation problems, including the need to provide cryotats for low operating temperatures, have thus far precluded commercially viable products. *See* JOSEPHSON EFFECT; SUPERCONDUCTING DEVICES.

Magnetic bubble memory. Another advanced technology is based on the fact that the polarization of tiny cylindrical magnetized areas, contained in a thin film of magnetic material, can be used to represent data 1's or 0's. Such devices are called magnetic bubble memories, and when used for data storage, the principal advantages are extremely high storage density

(7000 bits/mm^2 or 4.5×10^6 bits/in.^{2} and the fact that they are nonvolatile (data are not lost when power is shut off). The memory access time of 10 milliseconds is faster than a disk drive but slower than semiconductor memories. Power consumption is extremely low, being 10 microwatts per bit during reading or writing operations. Joseph A. Pecar

Bibliography. W. Anacker et al., Special issue on Josephson technologies, *IBM J. Res. Develop.*, vol. 24, no. 2, March 1980; K. L. Chopra and I. Kaur, *Thin Film Device Application*, 1983; A. Friedman and P. R. Memon, *Theory and Design of Switching Circuits*, 1975; S. Fulton, *Logic and Switching Circuits*, 1994; M. P. Mitchell, *Switching Circuits for Engineers*, 3d ed., 1975; J. Watson, *Analog and Switching Circuit Design*, 2d ed., 1989.

Switching systems (communications)

The assemblies of switching and control devices provided so that any station in a communications system may be connected as desired with any other station. A telecommunications network consists of transmission systems, switching systems, and stations. Transmission systems carry messages from an originating station to one or more distant stations. They are engineered and installed in sufficient quantities to provide a quality of service commensurate with the cost and expected benefits. To enable the transmission facilities to be shared, stations are connected to and reached through switching system nodes that are part of most telecommunications networks. Switching systems act under built-in control to direct messages toward their ultimate destination or address.

Most switching systems, known as central or end offices in the public network and as private branch exchanges (PBXs) when applied to business needs, are used to serve stations. These switching systems are at nodes that are strategically and centrally located with respect to the community of interest of the served stations. With improvements in technology, it has become practical to distribute switching nodes closer to stations. In some cases to serve stations within a premise, switching is distributed to take place at the stations themselves. A smaller number of systems serve as tandem (intermediate) switching offices for large urban areas or toll (long-distance) offices for interurban switching. These end and intermediate office functions are sometimes combined in the same switching system. *See* PRIVATE BRANCH EXCHANGE.

There are many types of telecommunication services. The principal ones are voice, data (record), picture (still), and video (motion pictures). For each service there is a different balance between the relative investment in transmission, switching, and station (terminal) facilities. This article deals primarily with systems for the switching of voice. Some data services use the voice network. Since separate networks for data services are available, a brief discussion will be given of switching exclusively for voice and data services.

The basic form of switching for current services utilizes circuit switching where a path in space or time is allocated to each message or call. The characteristics of telephone messages are quite different from data and imaging, which are either bursty or long and one-way. For these, packet switching has been introduced, where the message are subdivided into uniform-size arrays which need not be received in real time. *See* DATA COMMUNICATIONS; INTEGRATED SERVICES DIGITAL NETWORK (ISDN); PACKET SWITCHING.

Switching System Fundamentals

Telecommunications switching systems generally perform three basic functions: they transmit signals over the connection or over separate channels to convey the identity of the called (and sometimes the calling) address (for example, the telephone number), and alert (ring) the called station; they establish connections through a switching network for conversational use during the entire call; and they process the signal information to control and supervise the establishment and disconnection of the switching network connection.

In some data or message switching when real-time communication is not needed, the switching network is replaced by a temporary memory for the storage of messages. This type of switching is known as store-and-forward switching.

Signaling and control. The control of circuit switching systems is accomplished remotely by a specific form of data communications known as signaling. Switching systems are connected with one another by telecommunication channels known as trunks. They are connected with the served stations or terminals by lines.

In some switching systems the signals for a call directly control the switching devices over the same path for which transmission is established. For most modern switching systems the signals for identifying or addressing the called station are received by a central control that processes calls on a time-shared basis. Central controls receive and interpret signals, select and establish communication paths, and prepare signals for transmission. These signals include addresses for use at succeeding nodes or for alerting (ringing) the called station.

Most electronic controls are designed to process calls not only by complex logic but also by logic tables or a program of instructions stored in bulk electronic memory. The tabular technique is known as action translator (AT). The electronic memory is now the most accepted technique and is known as stored program control (SPC). Either type of control may be distributed among the switching devices rather than residing centrally. Microprocessors on integrated circuit chips are a popular form of distributed stored program control. *See* COMPUTER STORAGE TECHNOLOGY; INTEGRATED CIRCUITS; MICROPROCESSOR.

Common channel signaling (CCS) comprises a network of separate data communication paths used for transmitting all signaling information between offices. It became practical as a result of processor

control. To reduce the number of data channels between all switching nodes, a signaling network of signal switching nodes is introduced. The switching nodes, known as signal transfer points (STPs), are fully interconnected with each other and the switching offices they serve. All links and signal transfer points are duplicated to ensure reliable operation. Each stored-program-control toll switching system connects to the two signal transfer points in its region.

Numbering plan. Telecommunications networks are bound together by a single address coding plan which provides for uniquely identifying every station and terminal. The North American system is based upon decimal digits. (When letters are on the dial, the telephone system actually recognizes the numerals associated with the letters.)

A telecommunications central office customarily has the nominal maximum capacity to serve 10,000 main stations, using the number series 0000–9999. When there are more than 10,000 main stations, more than one central office is provided, sometimes in more than one building or wire center. Each office is given a separate three-digit designation or code. The minimum requirement is for an adequate number of digits or characters in each number to address each main station in the dialing area. When a call reaches the called office, the called telephone station is determined from the last four numerals (the main station code).

A seven-digit numbering plan has adequate capacity for only a small portion of the telephones in North America. Hence, a geographical area, such as a state or a Canadian province, is selected as a numbering plan area (NPA), within which there are no duplications of seven-digit numbers. The more populous states, which have large numbers of central office codes, are divided into two or more numbering plan areas.

Each numbering plan area is given a three-digit NPA code. Examples are 803 for the state of South Carolina and 415 for the portion of California that includes San Francisco. With this plan the equipment uses, first, the NPA code to determine which area is desired; second, the central office code to select the office in that area; and third, the main telephone number to determine the particular telephone being called. To enable the reuse of area codes as central office codes, a 1 is dialed ahead of each 10-digit number.

Area and central office codes have generally represented geographical areas or exchanges for which uniform rate tariffs are enforced. Non-geographical codes are used, such as 700, 800, and 900, for universal service distinctions. Code numbers for special services, such as 411 for directory assistance and 911 for emergency calls, are used.

In the United States, long-distance services are provided by networks of several carriers (interLATA carriers, discussed below). Callers may presubscribe to a particular carrier or may precede the telephone number by a five-digit prefix code, such as 10722, that designates the carrier whose services are desired on a particular call. These are known as carrier identification codes.

When calling, digits, known as prefixes, may precede the telephone number to represent the service desired on a call. Typically, these are to request the services of an operator (0, 00), to choose a particular telecommunication company (8, 9, or 10XXX, where X equals a digit other than 0), or to call outside of North America (01, 011). To distinguish calls not requiring the selection of services, a 1 is frequently used as a prefix.

Switching fabrics. Space and time division are the two basic techniques used in establishing connections. When an individual conductor path is established through a switch for the duration of a call, the system is known as space division. When the transmitted speech signals are sampled and the samples multiplexed in time so that high-speed electronic devices may be used simultaneously by several calls, the switch is known as time division. Space-division switching has been employed since the early manual switchboards in which the connectives were cords, plugs, and jacks.

Most switching is now automatic. Operators are required only for ancillary functions that cannot yet be economically automated. Such systems use the most modern techniques. In the United States, cord, plug, and jack switchboards have almost disappeared, having been replaced by cordless consoles. Calls are distributed automatically to operators seated at cordless consoles or computer keyboards with keys and lamps to permit them to serve calls requiring judgment. The positions may be located many miles from the stored-program-control system through which the connection is established or processed.

The switching fabric frequently comprises two primary-secondary arrangements: first, the line link (LL) frames on which the telephone lines appear and, second, the trunk link (TL) frames on which the trunks appear (**Fig. 1**). A switching entity may grow to a maximum of 60 line link and 30 trunk link frames. Each line link frame is interconnected with every trunk link frame by a network of links called junctors. Each line link frame has a basic capacity for 290 telephone lines and may be supplemented in 50-line increments to a maximum of 590 lines. The

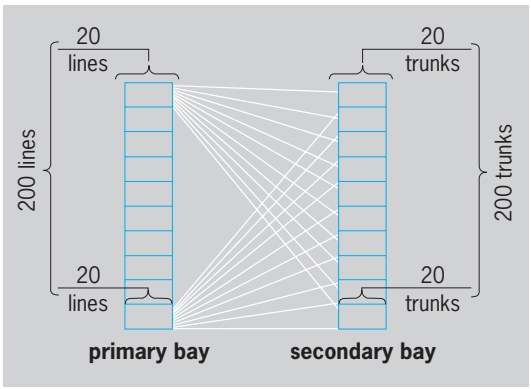


Fig. 1. Primary-secondary link arrangement.

size used in a particular office depends upon the calling rate and holding time of the assigned lines.

Systems use automatic message accounting (AMA) to make call records for billing purposes. Originally the AMA record was made on perforated tape. Now minicomputers located in the office or reached over data links record the AMA data on magnetic tape. On calls within the local area where a message-unit basis of charging applies, frequently only the calling telephone and the number of message units are recorded. For toll calls, the calling and called numbers, answering and disconnect times (from which the length of conversation is computed), and other data are recorded. Automatic data processors at accounting centers convert the recorded information into a form for printing the bill statement.

Centralized automatic message accounting (CAMA) has been applied to the offices used for tandem or toll switching functions. The calling station is identified automatically in the local office, and this number is sent along with the called number over the trunk to the tandem office. Where local offices are not arranged to identify the calling station, an operator is momentarily added to the connection to ask the calling party for his or her number and to key it into the system.

Electronic Switching

In the United States, 10,000 local and 600 toll stored-program-control switching systems have been placed in service, involving more than 120,000,000 lines (as of 1993). Stored program control has become the principal type of control for all types of new switching systems throughout the world, including toll, private branch, data, and Telex systems. About 2000 offices were earlier placed in service, principally in the United Kingdom and France, using electronic logic and memory controls based on the action translator principle (table look-up and proprietary control circuitry).

Service features. Two types of data are stored in the memories of electronic switching systems. One type is the data associated with the progress of the call, such as the dialed address of the called line. Another type, known as the translation data, contains infrequently changing information, such as the type of service subscribed to by the calling line and the information required for routing calls to called numbers. These translation data, like the program, are stored in a memory which is easily read but protected to avoid accidental erasure. This information may be readily changed, however, to meet service needs. The flexibility of a stored program also aids in the administration and maintenance of the service so that system faults may be located quickly.

The availability of large memories and the ease of changing the program residing in them has led to the development and deployment in most stored-program-control systems of many new and optional services such as abbreviated dialing, call waiting, call forwarding, and three-way calling. The wide introduction of common-channel signaling into the network broadens the range of stored-program-control service offerings. These services take advantage of the rapid exchange of information about calls between the originating, intermediate, and terminating offices before connections are established. As a result, calls may be directed within the entire network to suit the individual users and subscribers whose needs are likely to be nationwide rather than residing in a single central office or private branch exchange.

One of the most impressive results of employing electronics in switching is the space savings. Even though the early systems that were installed used mostly discrete semiconductor components, the space savings were as high as 60% when compared with electromechanical switching. Additional space savings have been achieved as integrated circuits are applied.

No. 1A electronic switching system. The high speed of electronics enables the systems to be designed so that all calls in progress are processed and supervised by the same control equipment. The No. 1A electronic switching system widely used in the United States for large local and small toll offices has a capacity as high as 345,000 local calls per hour and may serve as many as 125,000 lines (**Fig. 2**). High-speed peripheral buses connect the stored-program-control processor to the peripheral circuits of the system that access lines, trunks, and the switching network control.

The stored program control comprises a general-purpose assemblage of semiconductor circuits that are structured to interpret the instructions used in programs for the processing of the calls and for the maintenance of the system. These instructions are stored in a memory subsystem as coded programs that are read in sequences that determine actions to be taken.

Portions of the call-processing functions are used repetitively for each call. The programs are relatively fixed and remain in the system memory while the call information is stored for relatively short periods. The No. 1A ESS is provided with integrated circuit (IC) memory for call data, translation, and program storage. Integrated circuit chips are used in a random-access storage. A typical program requires more than 500,000 words of 32 bits each. All control circuits are duplicated to ensure service continuity should a component fail, or to enable the system to grow while in service. Memory redundancy is also provided. Copies of the call processing as well as additional, less frequently used maintenance, administrative, and operations (MAO) programs are stored on magnetic disks. The periphery includes data links to centralized MAO facilities and for common-channel signaling, as well as local maintenance consoles and tape drive units.

Space-division switching fabrics. Initially, most stored-program-control systems employed space-division fabrics. Many unique matrix arrays of devices, both metallic and nonmetallic, have been developed and used in commercial electronic switching systems.

For some electronically controlled switching systems, metallic contacts are used in the space-division

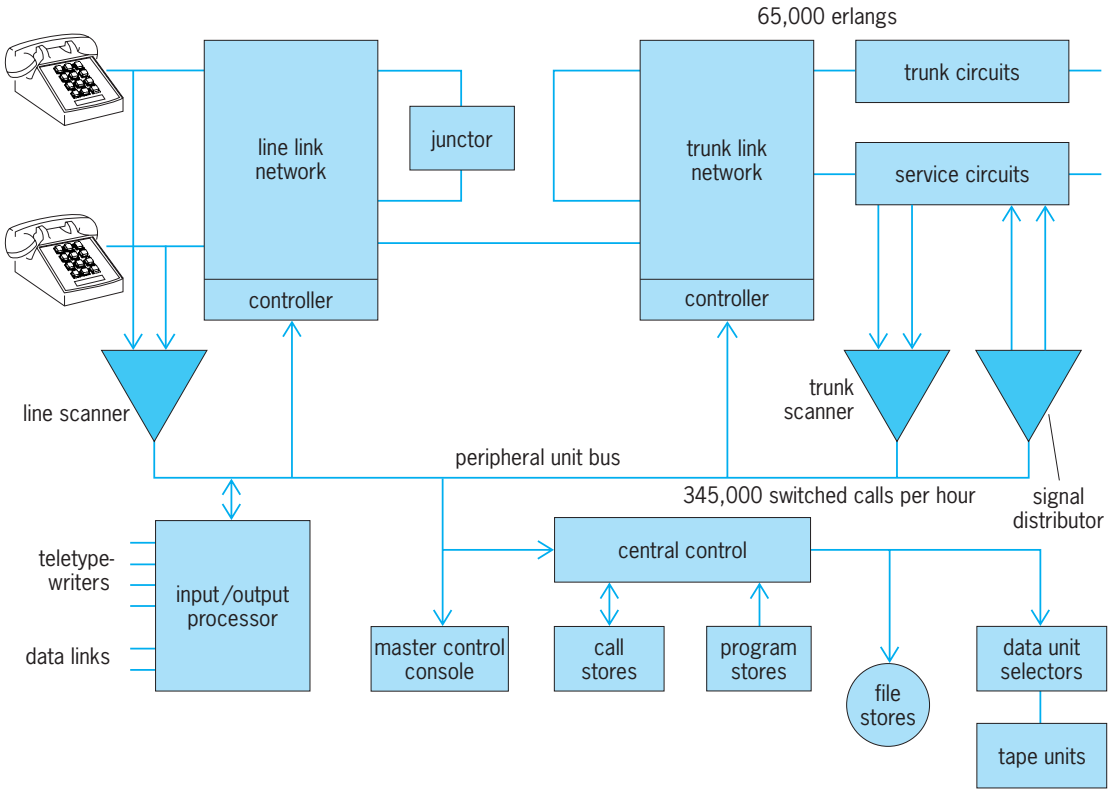


Fig. 2. Block diagram of No. 1A electronic switching system. Junctor is an intraoffice trunk circuit.

fabric. Many of these contacts are magnetic reeds sealed in glass with an inert gas. They latch magnetically when activated by a short pulse of one polarity and release with a pulse of opposite polarity. When the contacts are made of hard magnetic material, such as those employed in the No. 1A ESS, they are known as remreeds. Other systems use miniature crossbar or similar coordinate switches. In all of these space-division fabrics, the switching device is slow in comparison to the speed of the electronic controls. The network controller provides the buffering between high-speed stored program or action translator controls and the slower space division networks.

Electronics in the form of semiconductor or integrated circuit devices are used as crosspoints or gates in the electronic switching fabrics. Generally their use has been confined to smaller systems or PBXs (less than 2000 terminals) or in combination with metallic crosspoints for larger systems.

Semiconductor crosspoints may be used in fabric configurations in the same manner as reed switches. They are designed to be bistable and held actuated over the established speech path. Therefore, these devices combine both transmission and control characteristics. Each crosspoint may use a pair of bistable devices as the two conductors of the connection, or a single device with one wire and a ground return. Some modern semiconductor crosspoint devices are designed to function in the same high-energy environment as metallic devices.

Time-division switching fabrics. Time-division switching is practical only with high-speed electronic techniques. It is used in switching fabrics as well as in the control portions of systems. For time division, analog speech signals are sampled at a rate at least twice the highest frequency to be transmitted. Typically for voice this is 8000 times per second.

Pulse modulation. In a switching system the samples may be pulses of varying amplitude that are analogs to the electrical signals representing the voice. This is known as pulse-amplitude modulation (PAM).

More robust forms of pulse modulation use digital or on-off signals that can be readily encoded from the amplitude pulse, sent over a transmission medium, and periodically reformed to eliminate most of the impairments of transmission and switching. These types of digitally sampled transmission are known by the form of coding employed; the most popular is referred to as pulse-code modulation (PCM). In one type of pulse-code modulation, each sample is coded into one of approximately 256 amplitudes and represented by eight binary (on-off) pulses. The eight pulses representing each speech channel may be placed in sequence or time-multiplexed into groups, typically of 24 or 32 channels, so that the line and repeaters may be used more efficiently. *See* PULSE MODULATION.

Switching techniques. Time-multiplexed coded voice signals reaching a switching system are switched by using two techniques. Assuming signals arrive at the switching fabric on different multiplexed lines, they first need to be synchronized with respect to multiplex channel identification. This usually requires some form of time delay or buffering. This establishes uniform channel periods or time slots. Within

the switch there may be more time slots than in the lines delivering the signals.

One necessary switching fabric function uses further time buffering by placing successive digitized channel samples in a memory in one order and removing them as indicated by the switching selection requirements of each call. This function is known as time slot interchange (TSI).

The other switching fabric function provides for interchanging channels between time-multiplexed lines. Generally this function is carried out by using a high-speed space-division fabric of one or more stages. The space-division fabric control acts to change the actuated crosspoints between time slots so that the successive channels of input lines may be switched to corresponding channels in the same or other lines. Space-division stages operating at time slot rates are time-multiplexed switches (TMS).

Time-division switching fabrics are based upon the use of successive time slot interchanges, called T (for time) stages, and time-multiplexed switches, called S (for space) stages. Typical systems are said to employ T-S-T or S-T-S types of networks. Lines or trunks not reaching the switching system by time-multiplexed facilities must first be multiplexed with similar inputs as part of the time-division switching-fabric function.

Since the coded signals represent only one direction of transmission, the switching-fabric function in a circuit switch is duplicated by reciprocity to provide for the other direction of transmission. Also, since the elements of the network are active and may be used for hundreds of simultaneous calls, redundancy of the network and its controls, similar to that used in the call-processing portions of the system, is usually part of the system architecture.

Digital (PCM) toll systems. Time-division switching is a natural adjunct to digital time-division transmission where the coding is performed for purposes of multiplexing. Pulse-code-modulation transmission was initially economical on interoffice trunk routes from 10 to 50 mi (16 to 80 km). This made time-division digital switching attractive where such trunk facilities were found, namely, for tandem and toll applications.

As an example of a stored-program central time-division digital switching system developed for this application, the No. 4 ESS is the switching system of greatest capability. It has a capacity of 100,000 trunks, and its stored program control is capable of switching 670,000 calls per hour.

The switching fabric for the No. 4 ESS consists of both time-slot-interchange and time-multiplex-switching stages; the latter is reconfigured at the rate of 1,024,000 times per second (128 time slots for each frame of 8000 samples per second). A sample progresses through both the memory (time-slot-interchange) and four-space-division time-multiplexed switch stages for one direction of a typical time-slot connection, with the network control memory being read out cyclically at a rate of 8000 times per second (**Fig. 3**).

The No. 4 ESS serves three types of transmission channels: analog metallic, analog multiplex carrier, and digital multiplex carriers (**Fig. 4**). The digital

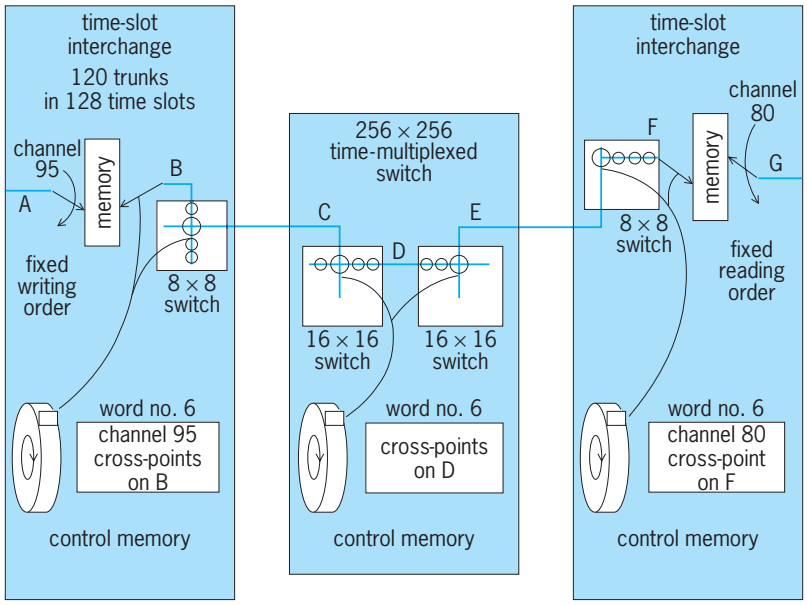


Fig. 3. Diagram of digital time-division switching fabric showing a typical connection. The network has 128 time slots throughout, and the sample shown is in time slot number 6 throughout the network.

time-division switching network routes digital signals from incoming trunks to the desired outgoing trunks. The audio signals on the analog channels are converted by the LT-2 connectors or D4 channel banks into pulse-code-modulation digital samples. The digital interface frame (DIF) processes the digital signals into the format required for the switching network and removes the signaling information. For analog and digital channels, signaling information is detected by the equivalent of trunk circuits in the digital interface frame except where common channel signaling is now being used. Echo suppression on a digital basis is inserted ahead of the switching network.

All operations are directed and supervised by the stored-program-control processor and are aided in routine tasks by the signal processors built into the digital interface frames. The signal processors provide the scanner and distributor functions for a portion of the trunks and, in turn, pass the significant information content in a more compact form to the central control. The use of peripheral or distributed processors enables the central processor to devote its attention to the more critical decisions in the processing of calls, thereby providing for greater call-attempt capacity. System units synchronize their operations under the control of a system clock.

Digital (PCM) local systems. Systems have been designed to terminate ordinary telephone lines carrying analog voice signals directly on analog-to-digital conversion circuits. To use time-division-multiplex switching, voice signals are digitized, assigned time slots, and then multiplexed. The line interface circuit, colloquially known as BORSCHT, provides the system with these capabilities, as well as those normally expected within local space-division switching systems. BORSCHT is an acronym for Battery (to feed the analog telephone transmitter), Overvoltage

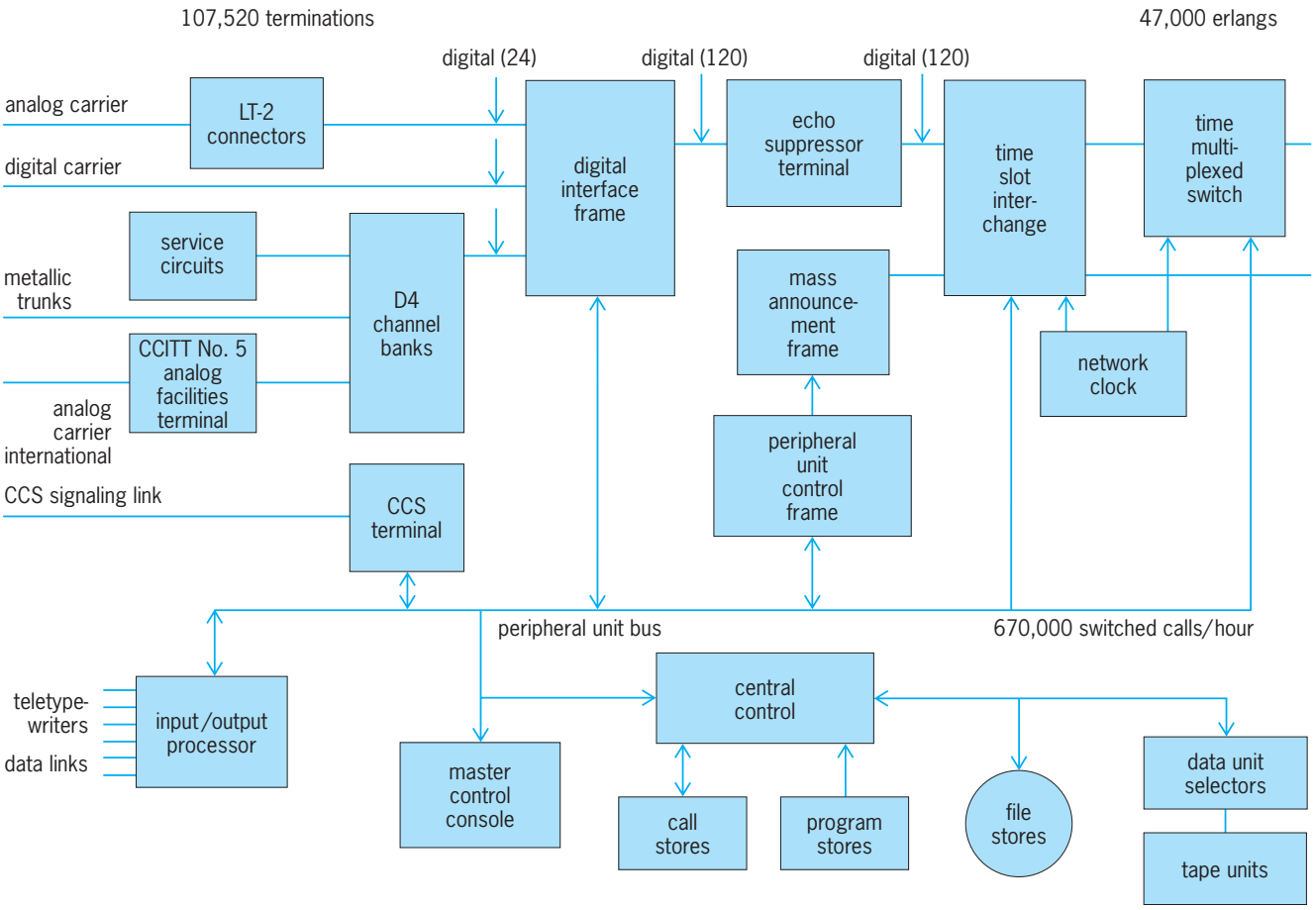


Fig. 4. Block diagram of No. 4 electronic switching system. CCITT = Comité Consultatif International de Téléphonique et Télégraphique.

protection, Ringing, Supervision, digital Coding, Hybrid to separate directions of transmission (a requirement for digital time-division-multiplex systems), and Test access (to reach both the in-office and outside plants).

To provide large call-attempt capacity, particularly for toll system applications, additional control processors are used, usually by providing several identical processors that share the load. These are called multiprocessor systems. Also, microprocessors are associated with the periphery of systems to distribute some of the stored-program-control functions. For smaller systems, some basic control functions are built into logic circuits rather than being stored as programs in the processor memory. These are action translator systems with call sequences determined by the use of memory on a look-up-table basis.

Distributed-control TDM systems. In more advanced local digital switching systems, not only is the interface circuit highly integrated, but line and trunk terminations are grouped into modules that permit the application of distributed control techniques. Some or all of the routine call-processing functions are performed by programs stored in microprocessors associated with the modules. In addition, the first switching stage, usually a time stage, is included in the module. The module therefore includes all of the basic switching functions of the network, interface, and control (**Fig. 5**). Systems with varying degrees of distributed control and network are known as distributed switching systems. The close association of these functions in a unit leads to further opportunities for the use of very large-scale integrated (VLSI) circuits.

Distributed switching has also renewed interest in separating the modules from the time-multiplex-switching and central control system core, which together are sometimes called the host. The modules may be located away from the system core, closer

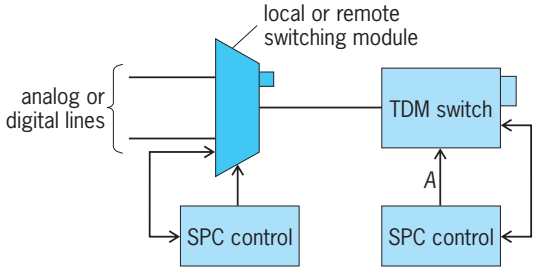


Fig. 5. Second-generation local time-division digital switching system. If the TDM switch is self-directing, the path labeled *A* is omitted.

to the customers. Not only are line facilities saved, but they use digital transmission to connect these modules to the host. Typically these remote units serve from a hundred to several hundred lines. For the larger units, the microprocessor controls used in the switch module are sufficiently autonomous to enable them to function on intralocal calls should the trunk linking the module with the host be disrupted. When remote units provide for some operation independent of the host central office, they are known as remote switching modules, units, or systems.

Changes in the system core have also begun to appear as a result of the application of distributed switching and large-scale circuit integration (LSI). These changes affect both the network and the control. The network employs both time-multiplex switching and time-slot interchange techniques, interchanging channel information between multiplexed links to and from the modules and between time slots. Large-scale integrated circuit chips can be designed with sufficient intelligence that the modules themselves may control the establishment of paths through the switching network. This technique is known as self-selection since the switching actions in the chip are controlled by signals arriving over the links.

Alternative switching techniques. Switching is gaining its own technology with integrated circuits that include complete time slot interchange, line interface (BORSCHT), and other circuits. As a result, many different system architectures and devices are being introduced around the world. Many systems are of the stored-program type with high-capacity random-access memory chips, such as those with 256,000 and 10^6 bits.

Maintenance and administrative functions have been centralized in computers that serve several central offices. Communication with these centers is two-way, with some information for special calls being stored at the central location.

The time-division digital switching technique is being used in newly installed systems as it becomes economically competitive with space division. Distributed switching is applicable to space division as well as time division. These applications may also use powerful low-cost microprocessors for call processing in switching modules remote from the host. A remote switch employing a nonmetallic switching network has been deployed in the United States.

Another technique for the switching of digital signals is the use of existing metallic space-division switching systems. Most modern switching systems will pass 64,000 bit/s digitized voice signals, highlighting the fact that a digital switching system may employ time or space division. But there is interfacing synergy between time-division-multiplex transmission and time-division switching systems. Furthermore, a time-division system provides an opportunity to transmit digital signals from nonvoice sources, such as computers and terminals. This capability gives rise to the concept of networks integrating digital multiplex transmission, switching, and services.

Data switching. For the switching of data and other digitized information, packet switching can be used. Digitizing the voice at the telephones results in end-to-end integrated services digital networking completely replacing the analog network.

Time-division digital switches serve over 65% of the access lines in the United States. As described above, they are limited to individual connections of 64,000 bits/s. By setting up multiple connections, up to 24, greater digital throughput is possible for certain applications.

For future telecommunications networks, a mixture of packet and circuit switching is envisioned with the message information, whether voice, data, or image, digitized and packaged into small cells of 53 bytes. These cells constitute the principal transmitted signals in a technique known as ATM. As the demand for broadband integrated services digital networks (ISDN) builds, ATM switching fabrics will be added to existing public switches. In the meantime, small ATM switches with a number of ports will appear to couple local-area networks, wide-area networks, and other networks serving relatively small groups of terminals on private data networks. *See* LOCAL-AREA NETWORKS; WIDE-AREA NETWORKS.

Toll Calling

In the United States, telephone service is provided by companies serving geographical areas known as exchanges. The local telephone companies are known as local exchange carriers. In much of the country the local exchange carriers serve one or more contiguous exchanges known as local access transport areas (LATAs). Service between LATAs is generally provided by many interLATA carriers (IXCs). There are several methods by which calls are extended to these carriers. A popular arrangement for serving these toll calls, both inward and outward, is known as equal access. Interconnection between the interLATA carriers and the local exchange carriers is usually through an access tandem. The call is switched to the interLATA carrier of the caller's choice either by presubscription or by dialing the prefix 10XXX, where XXX represents identification of the desired carrier. First, the calling line identification number is forwarded through the tandem to the interLATA carrier. The called number is then forwarded from the originating office to the interLATA carrier. Both numbers are sent by using alternating-current trunk signaling. In some cases the interLATA carrier has trunks directly to or from end offices.

The No. 4 ESS is one of the most commonly used systems for the switching of most interLATA toll calls, as well as for some intraLATA calls (**Fig. 6**). There are two methods of placing calls from the calling telephones. In one, the call is routed from a local office through an access tandem to the nearest toll office. Generally a 1 is dialed ahead of the area code on such calls. In the other, a call is routed from the local office through an operator system, when a customer dials 0 before the called number or makes a long-distance call from a public telephone.

The interLATA carriers use the calling number to

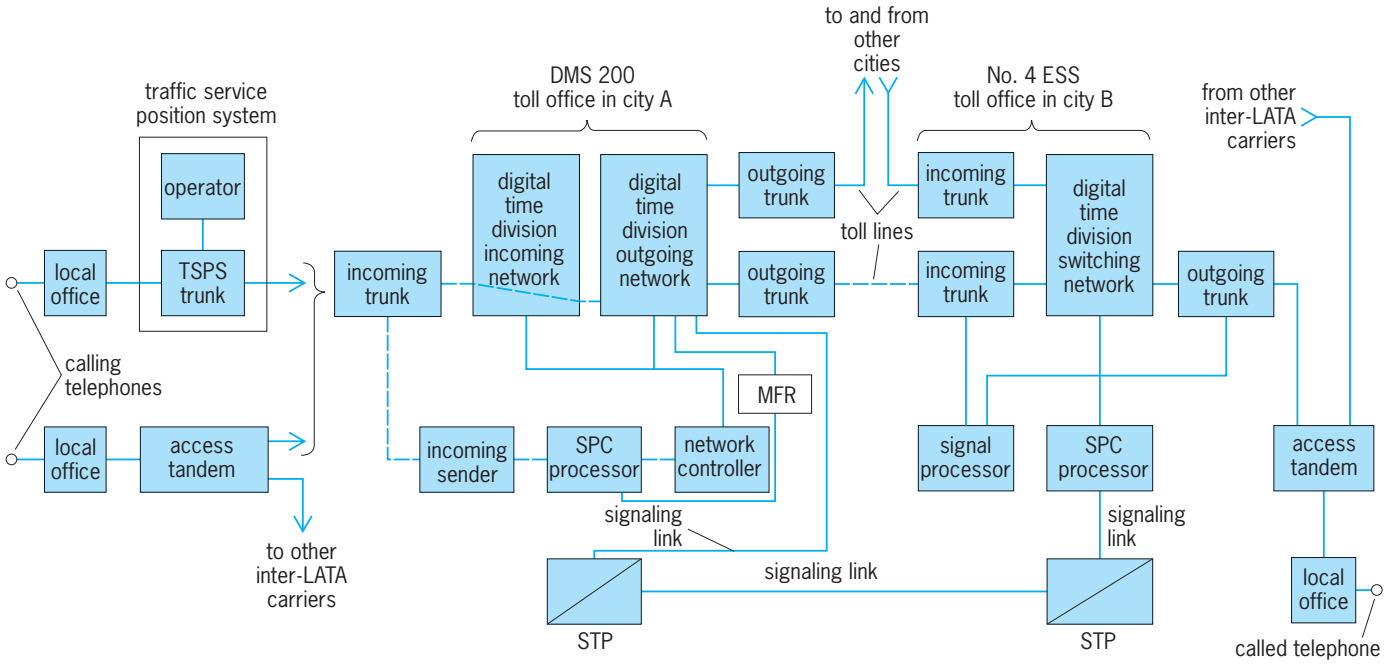


Fig. 6. Diagram of typical toll-call path between distant cities A and B, where the toll office in the city B uses No. 4 ESS and the toll office in city A uses the DMS 200 switching system. Abbreviations are explained in text.

determine if the caller is their customer and use the information for billing purposes. If the number is acceptable, the call is routed from the first interLATA carrier switching office to an office that reaches into the terminating LATA.

Calls arriving at the first interLATA carrier office are connected to a multifrequency receiver (MFR), and signals representing the calling and called numbers are sent from the access tandem.

In most interLATA carrier networks, common channel signaling is used between offices to send information about the call at 56 kilobits per second, including the identity of the outgoing trunk, to the two signal transfer points serving city A. One of these signal transfer points will transfer this information to a signal transfer point serving city B. This signal transfer point will then contact the No. 4 ESS office at which the trunk terminates. The stored program control in the office at B determines the routing and completes the connection to the local office, which in turn sets up the connection to the called telephone. The common-channel-signaling process includes a check that the transmission path selected for the call is viable.

The stored-program-control processor checks the calling number and determines the routing of the call. If all of the direct paths are busy, the interLATA carriers may have alternate routes planned that may carry the calls through intermediate interLATA carrier switches.

An important feature of modern switching systems is their ability to pick an alternate route if the first choice route between cities A and B has no idle toll lines. For example, a route from city A to city C to city B may be selected if no direct trunk to city B is available. Several such alternate routes may be examined, and a call may be routed through several switching points before reaching the terminating toll center. As many as 16 alternate routes may be preplanned. Some alternate routing arrangements are prioritized according to the time of day. The stored-program-control processor then passes information through the network controller to set up a connection from the incoming trunk to outgoing trunks through the incoming and outgoing networks.

Alternate routing ensures good service and economical trunking since, if one route is either busy momentarily or out of service due to an equipment or cable failure, the network is engineered to ensure that there is a good possibility that a toll line in some other route is available. The switching systems in complex networks employing alternate routing must be capable of sending pulses of the required type to operate the switches in the various switching centers. They can also delete digits or add digits to the called telephone number as required to operate the various switches.

Intelligent Networking

Large data storage subsystems, known as service or network control points, are used to verify information provided on credit card calling or for advancing calls, such as 800, to regular addresses in the network. This includes information for routing calls to a toll carrier of choice by the subscriber to the 800 service. The common channel signaling networks for these regional and long-distance carriers have been interconnected, shortening the time from the completion of dialing to the start of ringing on calls independent of their distance.

The information stored at the control points has become very sophisticated, particularly as it applies

to call processing. Calls may be directed differently depending upon time, date, and point of origination. Verbal prompts and voice recognition have been added to request and obtain further information from callers before further call processing. *See* SPEECH RECOGNITION.

The generic name given to these and other network enhancements of this type is intelligent networking. Since the databases are at centralized points, changes in the software in each central office switch may be avoided. Furthermore, the software of these switches is being modified with triggers that, for certain calls from particular stations, will forward the calls to the control points or intelligent peripherals. Intelligent peripherals enable third parties to devise and provide enhanced services to customers who may not have the capability of implementing new services on their own.

Untethered Switched Services

Modern mobile radio service has a considerable dependency on switching. The territory served by a radio carrier is divided into cells of varying geographical size, from microcells that might serve the floors of a business or domicile to cells several miles across that provide a space diversity for serving low-power radios.

Switching systems reach each radio cell site that detects, by signal strength, when a vehicle is about to move from one cell to another. The switching system then selects a frequency and land line for the communication to continue on another channel in a different cell without interruption. This is known as cellular mobile radio service and is used not only for voice but also facsimile and other forms of telecommunication. *See* MOBILE COMMUNICATIONS; TELEPHONE; TELEPHONE SERVICE.

Amos E. Joel, Jr.

Bibliography. R. J. Chapuis and A. E. Joel, Jr., *Electronics, Computers, and Telephone Switching*, 1990; A. E. Joel, Jr., *Electronic Switching: Digital Central Offices of the World*, 1982; F. Mazda, *Switching Systems and Applications*, 1996; J. C. McDonald, *Fundamentals of Digital Switching*, 2d ed., 1990; *Proceedings of the International Switching Symposium*, Florence, Italy, May 1984, AEI Publication, Milan, 1984; *Proceedings of the International Switching Symposium*, Stockholm, May 1990.

Switching theory

The theory of circuits made up of ideal digital devices. Included are the theory of circuits and networks for telephone switching, digital computing, digital control, and data processing.

Switching theory generally is concerned with circuits made of devices or elements that can be in two or more discrete conditions or states. Examples of such devices are switches or relay contacts, which can be opened or closed, rectifying diodes, which can be either forward- or back-biased, solid-state elements (such as transistors), which can be saturated or cut off, and magnetic elements which can be magnetized to saturation in either of two directions. Switching theory establishes an ideal representation of the digital circuit, examines the properties of the representation, then interprets these as properties of the circuit. Switching theory is not concerned with the physical phenomena of action or stability in a particular condition or with the details of transition from one state to another. It takes these as established and proceeds to examine more or less complex combinations of digital devices whose properties are assumed to be ideal.

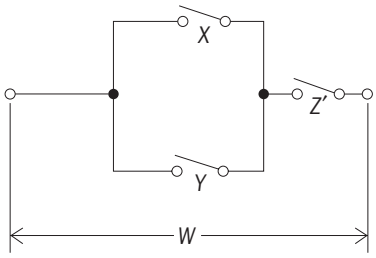
The bulk of switching theory is concerned with circuits that are made of binary (two-valued) devices, since these are most common. Switching theory can be based in part on mathematical logic. *See* BOOLEAN ALGEBRA.

A switching circuit whose outputs are determined only by the concurrent inputs is called a combinational circuit (or logic circuit). A circuit in which outputs at one time may be affected by inputs at a previous time is called a sequential circuit.

Combinational circuits. A rule by which the outputs of a combinational circuit can be determined from its inputs is called a switching function. Since the variables are discrete, a switching function may be expressed in tabular form as a truth table, or may be indicated by a diagram or geometric pattern. If the function and variables are binary, the symbols of 1 and 0 are commonly used to represent the two values. The function may then be represented by a boolean algebraic expression. The two values of a switching function can represent closed and open circuits, as for switches or relay contacts, or high and low or plus and minus voltages, as in electronic circuits.

The simplest combinational switching functions are the NOT function, the AND function, and the OR function. The NOT function is designated by the prime in boolean algebra; $Y = X'$ means that Y is closed (high, plus) when X is open (low, minus), and vice versa. The AND function is designated by the boolean product; $Z = X \cdot Y$ means that Z is closed (high, plus) only if both X and Y are closed. The OR function is designated by the boolean sum; $Z = X + Y$ means that Z is closed if either X or Y or both are closed. All other combinational switching functions can be made by combining these elementary building blocks.

For example, **Fig. 1** shows a switching circuit with three switches, or contacts, X, Y, and Z', each of which can be either open or closed. These can be thought of as input variables. The circuit as a whole will be open or closed depending upon the individual positions of X, Y, and Z'. Its condition can be designated by W, an output variable. Let 0 represent the open condition, and 1 the closed condition. The table in Fig. 1 represents the switching function of the circuit. The boolean expression for this function is $W = Z'(X + Y)$. To interpret this expression the rules of simple boolean algebra must be used: $0 + 0 = 0$ $0 \cdot 0 = 0$ $0' = 1$ $0 + 1 = 1$ $0 \cdot 1 = 0$ $1' = 0$ $1 + 0 = 1$ $1 \cdot 0 = 0$ $1 + 1 = 1$ $1 \cdot 1 = 1$.



X	Y	Z′	W
0	0	1	0
0	0	0	0
0	1	1	1
0	1	0	0
1	0	1	1
1	0	0	0
1	1	1	1
1	1	0	0

Fig. 1. Combinational circuit. $W = Z'(X + Y)$. X and Y are normally open contacts. Z' is a normally closed contact.

Switching theory establishes a number of methods for analysis and synthesis of combinational circuits. A significant problem is minimization, that is, given a switching function, to synthesize the simplest circuit which will realize it. A problem of some theoretical difficulty is that of realizability, that is, given a statement of specifications, to determine whether a switching circuit exists which satisfies them.

Analysis of a series-parallel combination of switches or relay contacts can be carried out by a direct application of boolean algebra. Variables or terms corresponding to contacts, or combinations in parallel, are added, and those in series are multiplied. The values are interpreted according to the rules of boolean algebra. Similar methods can be applied to combinational circuits which employ various solid state switching elements, such as transistors. Circuits that are not series-parallel can be dealt with by an extension of the boolean method, by the use of matrices with discrete-valued elements, or by a number of special methods.

A switching function can be simply synthesized as a series-parallel combination of contacts by giving boolean symbols circuit interpretations explained previously. Electronic logic circuits can be synthesized in a similar fashion. This approach will lead to a method for embodying any switching function expressed in boolean terms. The boolean expression of a function given in tabular or diagrammatic form is easily obtained.

Synthesizing the minimal circuit, or minimization, is more difficult since for every switching function there are many possible circuits. Where the number of variables is small, the minimization problem can often be reduced to one that has already been solved. Tables of minimal or nearly minimal solutions for relay circuits and solid-state circuits are available for circuits with one output and as many as four inputs. Harvard chart methods and Karnaugh map methods utilize geometrical relationships to explore systematically functions with one output and as many as six inputs.

As the number of variables increases, the possible number of functions rapidly becomes large. For example, there are more than 10^{19} different functions of six binary variables. No completely general and practical design methods have been discovered. However, a growing array of special methods for synthesis and minimization is available.

Sequential circuits. Since the outputs of sequential circuits depend on past, as well as present, inputs, they must contain means for remembering or storing the effect of past inputs, such as locking relays, flip-flops, delay lines, or solid-state memories. A device with two stable states can remember one binary digit, or bit. The amount of memory in a circuit can be measured either in bits or in internal states. An internal state of a circuit is a particular configuration of its internal memory devices. The number of internal states is equal to 2^n, where n represents its number of bits. Binary counters and shift registers are examples of sequential circuits.

It is possible to represent a sequential circuit as a combinational circuit with feedback. Thus, the combinational circuit of **Fig. 2** becomes a sequential circuit with two bits of memory if two of its outputs are connected to two of its inputs. Any such closed loop must contain gain and some delay; sometimes additional delay is inserted.

If the combinational circuit and the delays in Fig. 2 are completely specified, the internal description of the circuit is known and its behavior can be analyzed. If the switching function of the combinational circuit is such that $m_1 = M_1$ and $m_2 = M_2$ for a given set of inputs, no change can occur as a result of the action of the memory loops and the circuit is stable; otherwise, it is unstable. If it is unstable, the inputs must cause a transition to a new state, which in turn may

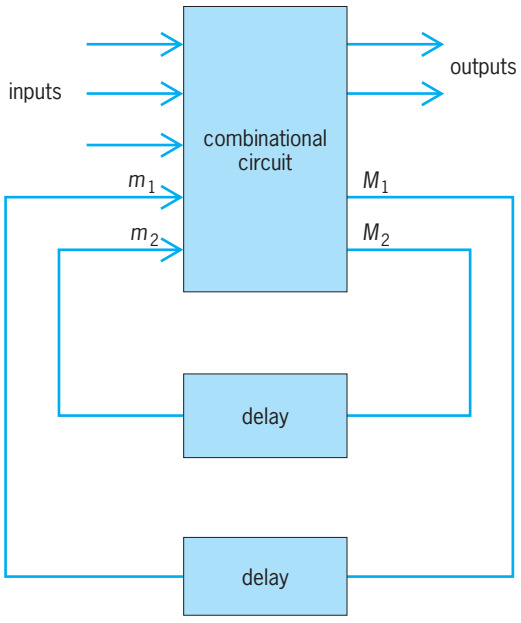


Fig. 2. Sequential circuit with two memory loops.

be stable or unstable. If no stable state is reached, the circuit is said to buzz. If the state to which a circuit may pass depends on which of two or more memory loops acts first, the circuit is said to have a race condition, and its performance may be ambiguous. This difficulty does not occur in circuits in which changes are caused or timed by repetitive clock pulses. Such circuits are called synchronous. Circuits which make transitions at the natural internal rate are known as asynchronous, and these asynchronous circuits must be designed with greater care.

To proceed from external circuit requirements to an internal description of a sequential circuit requires art and skill, as well as knowledge of switching theory. *See* DIGITAL COMPUTER; LOGIC CIRCUITS; SAMPLED-DATA CONTROL SYSTEM; SWITCHING CIRCUIT; SWITCHING SYSTEMS (COMMUNICATIONS).

Willard D. Lewis

Bibliography. E. D. Fabrioius, *Modern Digital Design and Switching Theory*, 1992; J. Y. Hui, *Switching and Traffic Theory for Integrated Broadband Networks*, 1990; Z. Kohavi, *Switching and Finite Automata Theory*, 2d ed., 1978; S. Muroga, *Logic Design and Switching Theory*, 1979, reprint 1990.

Sycamore

American sycamore (*Platanus occidentalis*), a member of the plane tree family, known also as American plane tree, buttonball, or buttonwood, and ranging from southern Maine to Nebraska and south into Texas and northern Florida. Ordinarily this tree is 60–120 ft (18–37 m) in height and has a trunk which is 2–5 ft (0.6–1.5 m) in diameter. Individuals 140 ft (43 m) tall and 14 ft (4.3 m) in diameter have been recorded. It has the most massive trunk of any American hardwood. Characteristic are the white patches which are exposed when outer layers of the bark slough off; the simple, large, lobed leaves whose stalks completely cover the conical winter buds; and the spherical fruit heads that are always borne singly in the American species and persist throughout the winter (**illus.** *a*). The tough, coarse-grained wood is difficult to work, but is useful for butchers' blocks, saddle trees, vehicles, tobacco and cigar boxes, crates, and slack cooperage. The trees are usually scattered through the forest in moist soil. *See* ROSALES.

(a) (b)

Sycamores. (*a*) American sycamore (*Platanus occidentalis*), bud, leaf with fruit ball, and budding twig. (*b*) London plane tree (*P. acerifolia*), leaf with fruit balls.

London plane (*P. acerifolia*) is supposedly a hybrid of *P. occidentalis* and *P. orientalis*, and it is one of the most desirable trees for planting in crowded cities because of its resistance to injury from gases, smoke, dust, and drought. It can be recognized by the usually three-lobed leaves resembling those of a maple, and by the fruit balls borne in groups of two or three (illus. *b*). The sycamore of Europe is usually understood to be a maple, *Acer pseudoplatanus*, known also as the sycamore maple. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Sycettida

An order of the subclass Calcaronea in the class Calcarea. This order comprises a rather diverse group of calcareous sponges, and includes the families Sycettidae, Heteropiidae, Grantiidae, Amphoriscidae, and Lelapiidae. Choanocytes with apical nuclei are limited to flagellated chambers and never occur lining the general spongocoel. The family Sycettidae resembles the contrasting order Leucosoleniida in lacking the true dermal membrane or cortex possessed by the other five families. The most massive skeleton (of bundles of modified triradiate spicules) is found in the family Lelapiidae, which superficially resembles the unrelated pharetronid sponges (abundant in the Permian and Triassic periods). The canal system is never asconoid (as in Leucosoleniida), but can be either syconoid or leuconoid. Leuconoid species seem to be derived from syconoid forms. Genera include *Grantia*, *Sycon*, *Sycetta*, *Heteropia*, *Amphoriscus*, and *Leucilla*. *See* CALCAREA; LEUCOSOLENIIDA.

W. D. Russell-Hunter

Syenite

A phaneritic (visibly crystalline) plutonic rock with granular texture composed largely of alkali feldspar (orthoclase, microcline, usually perthitic) with subordinate plagioclase (oligoclase) and dark-colored (mafic) minerals (biotite, amphibole, and pyroxene). If sodic plagioclase (oligoclase or andesine) exceeds the quantity of alkali feldspar, the rock is called monzonite. Monzonites are generally light to medium gray, but syenites are found in a wide variety of colors (gray, green, pink, red), some of which make the material ideal for use as ornamental stone.

Composition. Syenites may be classed as normal (calc-alkali) syenites or alkali syenites. In the latter the alkali feldspar and mafics are soda-rich. Intergrowths of potash and soda feldspar (perthite) are of various types and are strikingly developed. Feldspar grains usually show fair crystal outlines (subhedral) or may be irregular (anhedral). Many are highly interlocking. Sanidine occurs in some of the finer-grained varieties and in syenite porphyry. Normal syenites generally carry crystals of soda plagioclase

(oligoclase) which are usually subhedral and may be zoned (with calcic cores and sodic rims). Some alkali syenites contain discrete grains of albite. Plagioclase of monzonites may be as calcic as sodic andesine.

Black flakes (microscopically brown) of biotite mica and irregular to stubby prisms of green hornblende are characteristic of normal syenite. Diopsidic augite is the most common pyroxene and frequently forms cores within hornblende crystals. In alkali syenite the mafic minerals show wide variation. Biotite is deeply colored and iron-rich. Amphiboles are soda-rich (arfvedsonite, hastingsite, or riebechite) and are commonly zoned. Diopsidic and titanium-rich augite crystals are commonly encased by shells of aegerine-augite and aegerite.

Minor constituents may include quartz which is usually interstitial. When present in amounts between 5 and 10%, the rock is called quartz syenite; in excess of this amount, the rock becomes granite. Small amounts of feldspathoid (nepheline, sodalite, or leucite) may be present; but if in excess of 10%, the rock becomes a feldspathoidal syenite (nepheline syenite).

Accessory minerals include zircon, sphene, apatite, magnetite, and ilmenite. Accessories in special varieties of syenite include iron-rich olivine, corundum, fluorite, spinel, and garnet.

Texture. The texture of syenite is most commonly even-grained. Very coarse or pegmatitic textures are local. In some syenites numerous, relatively large crystals (phenocrysts) of alkali feldspar give the rock a porphyritic texture. These may be of early or late generation and may range from euhedral (well-formed crystals) to anhedral. They are particularly abundant in finer-grained varieties and in syenite porphyries.

Structure. A variety of directive structures may be present. Banding and parallel wavy streaks (schlieren) of different minerals are seen in some syenites; flow structures due to clustering and parallel orientation of elongate minerals may be present. Euhedral tabular feldspar crystals in parallel arrangement give the rock a distinctive appearance. In some cases these directive features represent effects of magma flow; in others they represent vestigial bedding or foliation in metasomatic or metamorphic rocks.

Occurrence and origin. Syenite is an uncommon plutonic rock and usually occurs in relatively small bodies (dikes, sills, stocks, and small irregular plutons). Normal syenite may be associated with monzonite, quartz syenite, and granite, whereas alkali syenites are associated with alkali granites or feldspathoidal rocks.

Many syenites have crystallized directly from syenitic magma (rock melt); others may have formed by reaction between magma of nonsyenitic composition and abundant contaminating rock fragments. Still others may have formed metasomatically as alkali-rich emanations, perhaps escaping from deeply buried magmas, have permeated rocks of special composition, and have replaced them with abundant alkali feldspar. *See* IGNEOUS ROCKS; MAGMA; METAMORPHISM; METASOMATISM.

Carleton A. Chapman

Symbiotic star

A double star system in the late stage of stellar evolution. A symbiotic star is a binary system and not a single star. The symbiotic phase represents a brief span in the life of the binary. Symbiotic stars are rare objects, and their distances are as a rule many hundreds of parsecs (1 parsec = 3.26 light-years). Thus, on ordinary photographic plates of the sky, symbiotic stars appear pointlike and are not resolved into two individual stars. On closer inspection, symbiotics are always associated with a nebular environment. The "near-official" list of symbiotic stars contains 188 safe entries; 15 of them are extragalactic, of which 1 lies in the dwarf galaxy Draco, 6 lie in the Small Magellanic Cloud, and 8 lie in the Large Magellanic Cloud. The list contains in addition 30 suspected candidates. *See* LOCAL GROUP; MAGELLANIC CLOUDS.

Spectra and variability. Symbiotic stars became targets of research in the early 1930s. Their spectra display titanium oxide absorption bands, indicative of a cool star, together with emission lines of singly ionized helium or doubly ionized oxygen and occasionally even many times ionized iron, which point to a hot nebula. The spectra indicate the presence of a cool M-type star with a surface temperature below 4000 K, and a hot nebula quite similar to a planetary nebula. The light output is variable on time scales of months and years, and the light curve can show abrupt changes, with increases that imply changes of at least a factor 2—but much more on other occasions—in the total energy output on time scales of days and months. *See* ASTRONOMICAL SPECTROSCOPY; LIGHT CURVES; NEBULA; PLANETARY NEBULA; VARIABLE STAR.

Evidence of binary systems. In 1981 the first astrophysical colloquium dedicated to symbiotic stars took place. Of the then more than 100 known symbiotics, only 5 were proven binaries. The proof came from periodic eclipses and from periodic Doppler shifts in wavelength that showed that the emitting object was moving to and from the observer. Binary periods between one and several years were found. During the following years the evidence accumulated that all symbiotics are double stars and that indeed their binary periods are between one year and many dozens of years. Their orbits are sufficiently wide for the two stars not to be in direct contact, and both stars lie safely within their Roche lobe. Thus, for the two stars stellar evolution proceeds over the whole main-sequence lifetime practically uninfluenced by the partner. That changes dramatically in the later stages of their evolution. *See* DOPPLER EFFECT; ECLIPSING VARIABLE STARS.

Dust. Systematic observations in the infrared led to a division into two classes: D-types that show infrared emission indicative of astronomical dust, and S-types that emit infrared radiation typical of red

giant atmospheres. D-types always seem to contain Mira variables, and their binary periods are probably longer than 10 years. The S-types are relatively dust-free and have binary periods shorter than 20 years. In the D-types, dust emits thermal radiation at average temperatures of typically 1000 K. Due to the wide binary separation in these systems, the cool star has been permitted to evolve to a state with substantial dust production in its envelope. *See* GIANT STAR; INFRARED ASTRONOMY; MIRA.

Late-stage stellar evolution. The gas temperature in the symbiotic nebula is approximately 15,000 K. This is much below the temperature that is needed to collisionally ionize the atoms in the nebula to the observed degree. The nebular gas is obviously radiatively ionized. This cannot be due to the cool star; it must be the action of a small, very hot star. The combination of observations and model calculations shows the cool star to be a red giant. Typical radii are 100 ± 50 solar radii, or up to 300 solar radii for Miras, and temperatures of 3500 ± 500 K, or approximately 2500 K for Miras. For the hot star, temperatures between 50,000 and 200,000 K are found. The star's size is that of a white dwarf or of a central star in a planetary nebula, thus between $\sim$0.01 solar radius (approximately the size of the Earth) and $\sim$0.1 solar radius; higher values are found for particularly active phases. Only in a few cases is it possible to determine stellar masses. They are approximately 2.5 ± 1.5 solar masses for the cool giants and 0.6 ± 0.2 solar mass for the hot star. The low mass of the white dwarf makes it unlikely that by further accretion this star could later explode as a supernova of type Ia; however, that issue is not resolved.

The symbiotic nebula has, as a rule, the same chemical composition as expected from red giants. This is taken as evidence that the red giant suffers considerable mass loss. Values between 10^{-8} and 10^{-7} solar mass per year have been found. Wind velocities are between 10 and 30 km/s (16 and 18 mi/s) for the red giant, and $\sim$1000 km/s (600 mi/s) for the white dwarf. This constellation of properties situates symbiotics in the late stage of stellar evolution. They began as stars with masses of the order of 5 solar masses. The more massive of them has gone faster through the main sequence, and when it arrived in the red giant region, after having converted hydrogen in its core to helium, it shed most of its mass through a stellar wind. That material can occasionally still be detected in the wider environment. The star has become a hot white dwarf. The originally less massive star has retained its mass and has now entered the red giant phase, with hydrogen burning in a shell, and a large mass loss. This phase may last on the order of 10^7–10^8 years. At the end of it, a planetary nebula is likely to be created for a few ten thousand years. *See* STELLAR EVOLUTION; WHITE DWARF STAR.

Interactive binary. A fraction of the mass lost by the red giant is captured by the white dwarf. This provides new hydrogen-rich fuel on top of the white dwarf's carbon-oxygen core. The accretion proceeds via some kind of accretion disk, the details of which are not known. It liberates gravitational energy

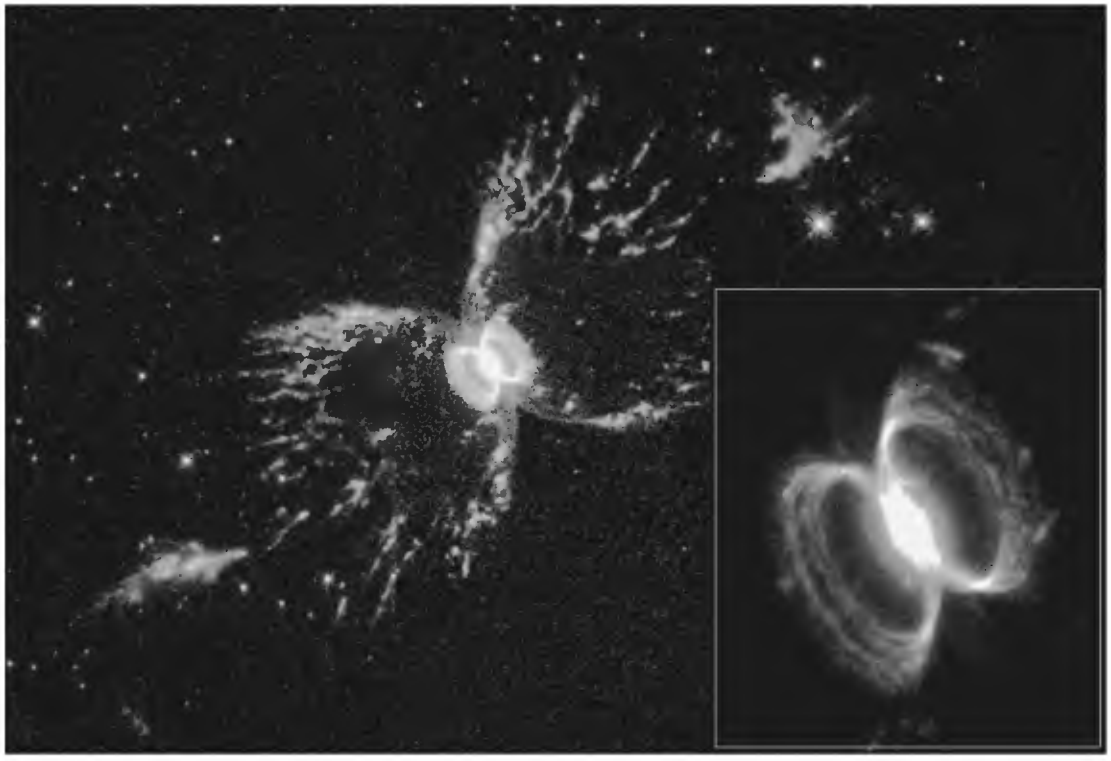


Fig. 1. ***Hubble Space Telescope*** **image of the Southern Crab Nebula, about 4400 parsecs (14,000 light-years) away. It contains a symbiotic system with a Mira-type star and a white dwarf, too close together to be visible as separate objects. The inset shows a bright nebula embedded in the center of a larger one. Both structures are probably due to a symbiotic nova outburst thousands of years ago. (*Courtesy of R. Corradi, M. Livio, U. Munari, H. Schwarz; NASA*)**

which can convert into radiative energy and be responsible for some of the irregular luminosity variations. The hot radiation from the white dwarf ionizes a fraction of the wind and the outer atmosphere of the red giant. This influences the dust production in the red giant.

Symbiotic nova. When the white dwarf has accumulated a critical mass of hydrogen, thermonuclear reactions on its surface lead to an outburst of energy lasting for over 100 years (**Fig. 1**). The energy production during that outburst can reach many thousand times that of the Sun. The mechanism is similar to a nova explosion, except that the nova has a higher peak energy output but a shorter duration. The total energy output of a symbiotic nova is comparable to that of a classical nova. Whereas in symbiotic novae the outburst on the white dwarf is not disturbed by the red giant, in classical novae the two stars are closer to each other, which has repercussions on the outburst.Whether some symbiotics in this way can accumulate sufficient mass to explode as a supernova of type Ia is still an open question. *See* CATACLYSMIC VARIABLE; NOVA; SUPERNOVA.

Nebular environment. Observations have shown traces of the mass formerly lost by the hot star. However, as noted above, the bulk of the matter now detected as an ionized nebula is due to the present mass loss of the red giant in its stellar wind. In its active phase the white dwarf may have a stellar wind of its own. In that case the nebular environment will be strongly structured by the collision of the two winds. This leads to shock zones with temperatures of several million kelvins. The origin of bipolar gaseous jets (**Fig. 2**) observed in several symbiotics is not yet known. Observational evidence indicates that the white dwarf possesses a strong magnetic field which could play a major role.

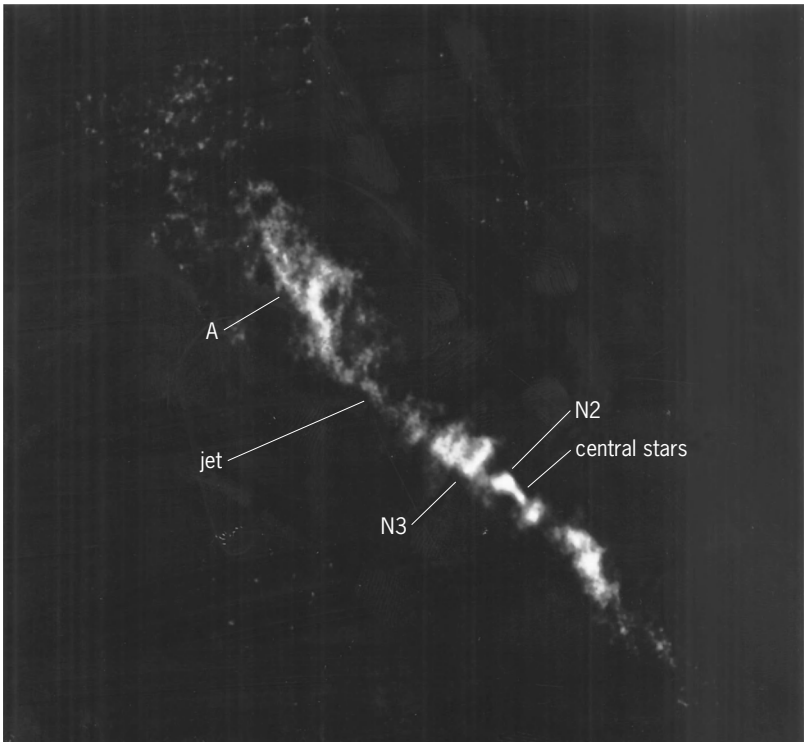


Fig. 2. ***Hubble Space Telescope*** **image of R Aquarii, the closest symbiotic system, about 200 parsecs (650 light-years) away. Its highly collimated bipolar jets are ejected from the central stars, which are too close together to be seen as separate objects. The jets encounter dense regions, such as N2, N3, and A, where they create shock waves, inducing strong radiation emission. (*Courtesy of F. Parece, European Southern Observatory; NASA*)**

Observations. The simultaneous presence of cool dust, a cool star, a hot star, and a hot nebula distributes the spectral information from x-rays through the far-ultraviolet, the visual range, and the infrared, up to radio wavelengths. In addition to ground-based observatories, several artificial satellites—in particular the *International Ultraviolet Explorer* (*IUE*), the *Hubble Space Telescope* (*HST*), the *Roentgen-Satellite* (*ROSAT*), the *Advanced Satellite for Cosmology and Astrophysics* (*ASCA*), the *Far Ultraviolet Spectroscopic Explorer* (*FUSE*), as well as *XMM-Newton*, the *Chandra X-ray Observatory, INTEGRAL, SWIFT*, and the *Infrared Space Observatory* (*ISO*)—have contributed spectra and pictures. Amateur astronomers, following daily light variations, have provided valuable contributions. The discovery of strong Rayleigh and Raman scattering within the symbiotic system, which leads to polarization of the scattered light, has helped to determine the binary movement and the highly complex structure of the symbiotic nebula. *See* BINARY STAR; CHANDRA X-RAY OBSERVATORY; GAMMA-RAY ASTRONOMY; HUBBLE SPACE TELESCOPE; POLARIMETRY; RAMAN EFFECT; SATELLITE (ASTRONOMY); SCATTERING OF ELECTROMAGNETIC RADIATION; STAR; ULTRAVIOLET ASTRONOMY; X-RAY ASTRONOMY.

Harry Nussbaumer

Bibliography. K. Belczynski et al., A catalogue of symbiotic stars, *Astron. Astrophys. Suppl.*, 146: 407–435, 2000; R. L. M. Corradi, J. Mikolajewska, and T. J. Mahoney (eds.), *Symbiotic Stars Probing Stellar Evolution*, ASP Conf. Proc., vol. 303, Astronomical Society of the Pacific, San Francisco, 2003; S. J. Kenyon, *The Symbiotic Stars*, Cambridge University Press, 1986; J. Mikolajewska (ed.), *Physical Processes in Symbiotic Binaries and Related Systems* (Proceedings of an international conference held in 1996), Copernicus Foundation for Polish Astronomy, Warsaw, 1997; J. Mikolajewska et al. (eds.), *The Symbiotic Phenomenon* (Proceedings of the IAU Colloquium held in 1987), Kluwer Academic, 1988.

Symbolic computing

The manipulation of symbols, representing variables, functions, and other mathematical objects, and combinations of these symbols, representing formulas, equations, and expressions, according to mathematical rules.

Giving the solution to an equation using symbols instead of numbers is referred to as symbolic manipulation or symbolic computing. Often, symbolic computing refers to computations or calculations that involve the manipulation of mathematical symbols to obtain solutions. For example, the differentiation and integration of functions in calculus, manipulation of

matrices in linear algebra, or determination of the function that solves a differential equation are now commonly referred to as symbolic computations.

Consider the equation

$$3x - 5 = 0 \tag{1}$$

whose exact solution is $x = 5/3$ or $x = 1^2/_3$ or $x = 1.6\overline{66}$, where the digit 6 repeats forever. All of these solutions are called exact solutions to the equation. If we write $x \approx 1.666$, we say 1.666 is an approximate numerical solution to our equation. Determining exact decimal digit solutions to an equation is not always possible. Consider the equation

$$x^2 - 2 = 0 \tag{2}$$

whose exact solution is $x = \pm\sqrt{2}$. Since $\sqrt{2}$ is a symbol for representing a number, this is referred to as the exact solution in symbolic form. An exact decimal digit representation for $\sqrt{2}$ is not possible, since there is no repeating pattern in the infinite number of digits for $\sqrt{2}$.

Now consider the general form for each of the above two equations. The general form for the first equation would be $ax+b = 0$, where a and b would represent real numbers. In particular, letting $a = 3$ and $b = 5$ in this equation leads to the first equation. The solution to this equation in symbolic form is $x = -b/a$ as long as $a \neq 0$. The second equation is a specific example of the general quadratic equation $ax^2+bx+c = 0$, whose solution in symbolic form is

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{3}$$

One would let $a = 1$, $b = 0$, and $c = -2$ to obtain the second equation. *See* CALCULUS; LINEAR ALGEBRA.

Software packages. In the past, symbolic computation was done by the human mind with paper and pencil or in some similar manner. Now computers are able to do symbolic manipulations, commonly using software packages or hand-held graphing calculators. Some common software packages that can do symbolic computing are Mathematica®, Maple®, and MATLAB®. Texas Instruments and Casio are two of the most common hand-held calculators that have symbolic computing features. Microsoft Office's Excel allows Maple as an add-in so that symbolic computing can be done. Software programming environments that can do symbolic computing are usually called symbolic computing packages or computer algebra systems. *See* CALCULATORS.

Symbolic computing software packages are able do algebra and calculus in any dimension—that is, they have the ability to do scalar or vector calculus (**Fig. 1**). They can do linear algebra operations. They can solve many differential equations and perform transformations on objects, like differential equations. Many packages have components that do abstract analysis. Symbolic computing packages can work with real or complex numbers and usually have some routine for dealing with complex arithmetic and its subtleties. Symbolic computing packages are capable of doing most university undergraduate mathematics and many areas of high-level research mathematics. Most of the standard algorithms used in mathematical and scientific research are incorporated in the package. Sometimes the algorithm is stored in a library inside the software package and has to be loaded into the package from a directory. *See* ALGORITHM; MATHEMATICAL SOFTWARE.

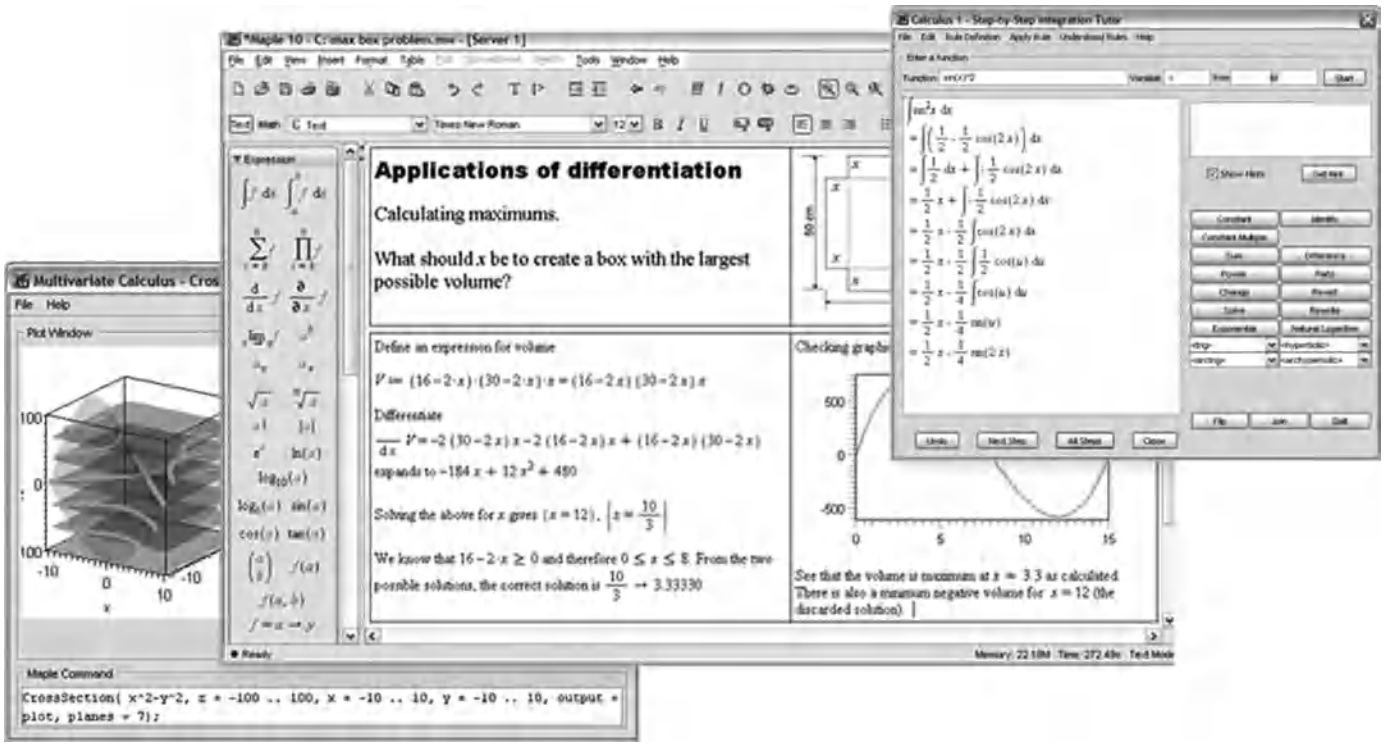


Fig. 1. Example of a calculus software package. (*Courtesy of Maplesoft*)

Symbolic computing packages are equipped with algorithms and tables that can replace a symbol for a number with an accurate or exact decimal digit numerical value. For example, one can ask the package to give the first 200 decimal digits for $\sqrt{2}$. If the package gives the answer as a symbolic expression, one can replace the parameters in the expression with numerical values or ask the package to do it and then ask it to give numerical values to the symbolic expression to as many decimal digits as desired.

The symbolic formulas given by the software package as an answer can usually be visualized in a variety of formats. These packages are able to rotate graphics and view graphics from different viewpoints and in different formats. Graphics frames can be built and then animated in sequence to appear as a movie. This allows visual scientific analysis and discovery. The software packages can also visualize data sets of numerical values in various formats. The graphics for both of these cases can be done in one, two, or three dimensions with different coloring schemes and plotting symbols depending on the user's preferences (**Fig. 2**).

In most cases, the graphics generated by these software packages is of publishable quality. In fact, most symbolic computing packages have features that allow their graphics to be exported in various formats for implementation in a variety of word processors. Some packages will even translate their symbolic output into the format needed by specialized scientific word processors.

Most symbolic computing packages come with a high-level language that the user can program. These programs can usually be translated by the software package to a high-level language like C or Fortran. These C or Fortran programs can then be run on state-of-the-art computers to improve the speed of calculations, the efficiency of calculating, and the robustness (accuracy) of results. The data run by these codes can then be inputted into the software package where they can be visualized. The data can be viewed in a variety of formats. *See* PROGRAMMING LANGUAGES.

Applications. Symbolic computing software packages allow a user to do symbolic manipulations in seconds. These same manipulations can take minutes, hours, or even days for humans to do. Some of the manipulations are not humanly possible. The fact that symbolic computation packages can do manipulations so quickly allows teachers to show students in the classroom environment symbolic computations that used to take several classroom sessions or could not be demonstrated at all in the classroom because

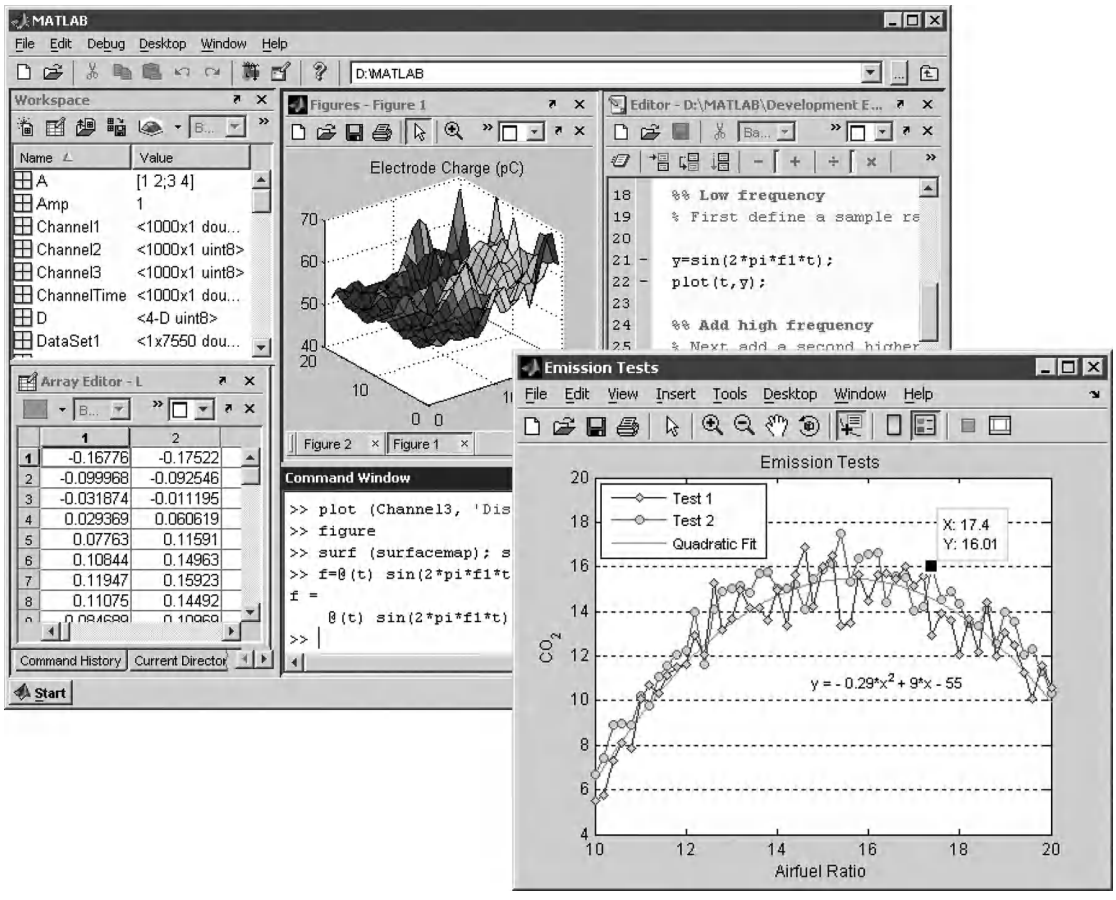


Fig. 2. MATLAB development environment, which allows users to develop algorithms, interactively analyze data, view data files, and manage projects, showing line plots of multiple engine emission test results with a curve fitted to the raw data. (*Courtesy of The MathWorks, Inc.*)

of time considerations. Symbolic computing packages are giving teachers much more flexibility and power in the classroom.

These software packages can also do professional-grade research. They are used to improve and enhance engineering design in many industries, such as electronics, automotive, and aviation. Many software packages have the ability to be run in parallel. This feature allows the increase in the speed of calculations and the speed of visualization formats.

Almost all symbolic computing packages have a new-user learning environment. This tool allows the new user to go through a set of lessons to learn how to use the symbolic computing package and its components. The learning environment usually shows the user how to build sessions, called worksheets, and write programs in the computing packages language that can be used not only by the user but also by other users. The learning environment usually describes how to add features contained in the software package's libraries. Many textbooks are available for learning Maple, Mathematica, or MATLAB. There are also many mathematical, scientific, and engineering textbooks that use them as part of the text and have exercises for which the reader must use one of these packages.

Worksheets. One strength of symbolic computing packages is their ability to allow the user to build worksheets or programs that other users can access as a tool or as a component. This allows teachers, for example, to build a fully documented worksheet. The teacher can then have students work through the worksheet to learn some mathematical or scientific concept. The worksheet can contain all the steps in sequence for the student to go through one at a time. Symbolic computing packages allow the user to put text in the worksheet in textbook format so that it is easy for others to see how to use the worksheet. The packages allow mathematical symbols to be typed in textbook format in the text. The author of the worksheet can use this feature to detail the mathematics that the worksheet can perform. As another example, an engineer could build a worksheet in the symbolic computation package to use as a tool and then make this worksheet available to all the other engineers in the firm as a component to the symbolic computing package.

Most symbolic computing packages also allow the user to make personal worksheets available to everyone on the Internet. Each software package has a specific format for doing this. The Web sites of symbolic computing software package vendors offer worksheets that can be run over the Internet. Most of these are scientific sessions and usually have some sophisticated visualization that displays the symbolic and/or numerical calculation. *See* INTERNET.

James Sochacki

Bibliography. W. Gander and J. Hrebícek (eds.), *Solving Problems in Scientific Computing Using Maple and MATLAB*, 4th ed., 2004; S. Wolfram, *The Mathematica Book*, 5th ed., 2003.

Symmetrodonta

A group of extinct mammals that range from the Late Triassic to Late Cretaceous. Their fossil remains have been found in Mesozoic deposits worldwide. These small insectivorous or carnivorous mammals are the size of a shrew or mouse. They are considered to be distant relatives of the more derived therian mammals, including the extinct eupantotheres, such as dryolestids and the living placental and marsupial mammals.

Symmetrodonts are characterized by a distinctive feature: the main cusps of their cheek teeth are arranged in a symmetrical triangle. The stylocone, paracone, and metacone of the upper tooth form a triangle known as the trigon, while the protoconid, paraconid, and metaconid of the lower tooth form a triangle known as the trigonid. The triangular upper and lower cheek teeth fill in the gaps between the adjacent teeth of the opposite tooth row, and are specialized for crushing insects or slicing worms. However, symmetrodonts lack a basinlike heel in the lower teeth (known as the talonid basin) that would allow the grinding of the ingested food in the more derived living therian mammals and their kin. Symmetrodonts are also distinguished from more derived therians by lacking the angle on the mandible.

A newly discovered symmetrodont skeleton from China has a straight cochlea, an auditory structure in the ear that is less derived than those in living therians. Its shoulder girdle was mobile and more derived than those of living monotreme mammals. However, symmetrodonts retain a primitive interclavicle bone that no longer exists as a separate bone in the adults of living therian mammals. Its forelimb had a more splayed posture than those of the living therians.

Symmetrodonts are classified as the order Symmetrodonta of the class Mammalia. Three families are notable among the diverse symmetrodonts. The Kuehneotheriidae, found in the Upper Triassic to the Lower Jurassic deposits of Europe and India, are considered to be close to the ancestry of all other therian mammals. Kuehneotheriids have primitive feature characters in the mandible. The Amphidontidae are from the Upper Jurassic deposits of North America and England. The Spalacotheriidae, which are characterized by the highly acute triangular cheek teeth, are the most diverse symmetrodont family and have been found on most continents (except Antarctica) in the Upper Jurassic to the Upper Cretaceous sediments.

Zhexi Luo

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1987; R. L. Cifelli and S. K. Madsen, Spalacotheriid symmetrodonts (Mammalia) from the medial Cretaceous (upper Albian or lower Cenomanian) Mussentuchit local fauna, Cedar Mountain Formation, Utah, USA, *Geodiversitas*, 21(2):167–214, 1999; P. Ensom and D. Sigogneau-Russell, New symmetrodonts (Mammalia, Theria) from the Purbeck Limestone Group, Lower Cretaceous, southern England, *Cretaceous Res.*, 21:767–779, 2000; Y. Hu et al., A new

symmetrodont mammal from China and its implications for mammalian evolution, *Nature*, 390:137–142, 1997; M. C. McKenna and S. K. Bell, *Classification of Mammals Above the Species Level*, 1997; D. Sigogneau-Russell and P. Ensom, Thereuodon (Theria, Symmetrodonta) from the Lower Cretaceous of North Africa and Europe, and a brief review of symmetrodonts, *Cretaceous Res.*, 19:445–470, 1998.

Symmetry breaking

A deviation from exact symmetry. According to modern physical theory the fundamental laws of physics possess a very high degree of symmetry. Several deep insights into nature arise in understanding why specific physical systems, or even the universe as a whole, exhibit less symmetry than the laws themselves.

Spontaneous symmetry breaking. This mechanism occurs in quite diverse circumstances and brings a cluster of important observable consequences whenever it occurs. The most symmetrical solutions of the fundamental equations governing a given system may be unstable, so that in practice the system is found to be in a less symmetrical, but stable, state. When this occurs, the symmetry is said to have been broken spontaneously.

For example, the laws of physics are unchanged by any translation in space, but a crystalline lattice is unchanged only by special classes of translations. A crystal does retain a large amount of symmetry, for it is unchanged by those finite translations, but this falls far short of the full symmetry of the underlying laws. *See* CRYSTAL STRUCTURE.

Another example is provided by ferromagnetic materials. The spins of electrons within such materials are preferentially aligned in some particular direction, the axis of the poles of the magnet. The laws of physics governing the interactions among these spins are unchanged by any rotation in space, but the aligned configuration of spins has less symmetry. Indeed, it is left unchanged only by rotations about the polar axis. *See* FERROMAGNETISM.

In both these examples, the loss of symmetry is associated with the appearance of order. This is a general characteristic of spontaneous symmetry breaking.

These examples can provide some understanding of why spontaneous symmetry breaking is so common. For concreteness, ferromagnets will be discussed. The laws of physics operating between two spins attach a particularly low energy, and therefore special stability, to the configuration wherein both point in the same direction. This in itself does not break the rotation symmetry, because the spins can be rotated together by any amount, without changing the energy. If a third spin is added, it is found to be particularly stable when it is aligned with the previous two, and so forth. In this way, a system of many spins aligned in some particular direction is built up, for which rotation symmetry is effectively lost. For while small disturbances from the outside world, or even quantum fluctuations, will cause a system of a few spins to rotate as a whole, with many spins this becomes increasingly difficult, and eventually a definite stable direction is fixed, breaking the symmetry.

Consequences. As mentioned above, there is a cluster of important observable consequences associated with spontaneous symmetry breaking.

Nambu-Goldstone bosons. This is a class of low-energy excitations associated with gentle variations of the order. In the case of ferromagnets, an overall rotation of all the spins at once does not change the energy. If the direction of alignment is changed slowly in space, each spin will be nearly aligned with its neighbors, and the overall cost in energy will be small. Thus, there is a class of excitations of the ferromagnet, the magnons, that exist as a consequence of the spontaneous symmetry breaking, and that have very low energy. Similarly, in the case of crystals, phonons are associated with gentle distortions of the lattice structure. *See* MAGNON; PHONON.

Phase transitions. At high temperatures the energy gained by assuming an ordered structure is increasingly outweighed by the entropy loss associated with the constraints it imposes, and at some point it will no longer be favorable to have spontaneous symmetry breaking in thermal equilibrium. Changes from broken symmetry to unbroken symmetry are marked by phase transitions. For a magnet, the transition occurs (by definition) at the Curie temperature, above which the material no longer displays a magnetic moment. For a crystal, it is melting into a liquid or sublimation into a gas. *See* CURIE TEMPERATURE; ENTROPY; PHASE TRANSITIONS; THERMODYNAMIC PRINCIPLES.

Defects. These are imperfections in the ordering. The most familiar examples are domain walls in magnets. On either side of the domain wall the spins are aligned, but the direction of alignment changes in crossing the wall. *See* CRYSTAL DEFECTS; DOMAIN (ELECTRICITY AND MAGNETISM).

Systems with long-range forces. In systems with long-range forces as well as spontaneous symmetry breaking, it need no longer be true that gradual changes require only a small input of energy, because even distant regions interact significantly. Thus, the Nambu-Goldstone bosons no longer have very low energies, and they are not easily excited. Conversely, the system will exhibit a special rigidity, with strong correlations between distant points. These ideas are central to modern theories of superconductivity and of particle physics (the Higgs mechanism). *See* ELECTROWEAK INTERACTION; STANDARD MODEL; SUPERCONDUCTIVITY.

Other forms. There are several other sources of symmetry breaking in nature, besides spontaneous symmetry breaking.

Frozen accidents. Sometimes the processes tending to enforce an equilibrium configuration that has some symmetry are so slow that, for practical purposes, this configuration is never achieved. Then so-called frozen accidents occurring during the formation of the system are locked in, and generally ruin its

potential symmetry. Examples are provided by glasses and by the handedness of biological molecules. *See* GLASS; MOLECULAR ISOMERISM; STEREOCHEMISTRY.

Asymptotic symmetries. Sometimes symmetries are imperfect, but are in a well-defined sense as good as they can be, subject to limitations imposed by fundamental laws of physics. Under resizing symmetry, also known as scale invariance, physical laws operate in the same way if the sizes of all bodies are multiplied by a common factor. In quantum chromodynamics, resizing symmetry becomes more nearly valid as smaller subatomic objects are considered. Perfect resizing symmetry could be true only in a physically trivial theory, but quantum chromodynamics comes as close as a nontrivial theory can. In inflationary universe models, the universe is as homogeneous as it can be, subject to irreducible quantum fluctuations in the original state. *See* INFLATIONARY UNIVERSE COSMOLOGY; QUANTUM CHROMODYNAMICS.

Symmetry limited by partial vision. A symmetry may appear to be broken to an observer with only a partial view of the system under consideration. For example, a magnet just above its Curie temperature has large regions where nonzero average alignments persist for long times, and only by averaging over extremely large regions or very long times are averages to zero obtained.

In quantum mechanics, a highly symmetric wave function may describe, when applied to any particular set of experiments, much less symmetry. For example, the wave function describing a photon emitted in atomic decay may be completely symmetric under rotations. Nevertheless, any particular measurement of the photon finds it moving in some definite direction, breaking the symmetry. According to one interpretation of quantum mechanics, the wave function potentially describes many worlds—that is, many possible results of macroscopic experiments—and the limited vision of occupants of any one world prevents them from seeing the complete symmetry. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; SYMMETRY LAWS (PHYSICS).

Frank Wilczek

Bibliography. P. Anderson, *Basic Notions of Condensed Matter Physics*, 1984; M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond*, 1987.

Symmetry laws (physics)

The physical laws which are expressions of symmetries. The term symmetry, as it is used in mathematics and the exact sciences, refers to a special property of bodies or of physical laws, namely that they are left unchanged by transformations which, in general, might have changed them. For example, the geometric form of a sphere is not changed by any rotation of the sphere around its center, and so a sphere can be said to be symmetric under rotations. Symmetry can be very powerful in constraining form. Indeed, referring to the same example, the only sort of surface which is symmetric under arbitrary rotations is a sphere.

The concept that physical laws exhibit symmetry is more subtle. A naive formulation would be that a physical law exhibits symmetry if there is some transformation of the universe that might have changed the form of the law but in reality does not. This formulation is much too rigid, however, since the comparison of different universes is generally not feasible or desirable. A more fruitful definition of the symmetry of physical law exploits a profound feature of successful descriptions of nature: locality. Locality is the principle that the behavior of a given system is only slightly affected by the behavior of other bodies far removed from it in space or time. Because of locality, it is possible to define symmetry by using transformations that do not involve the universe as a whole but only a suitably isolated portion of it. Thus the statement that the laws of physics are symmetric under rotations means that (say) astronauts in space would not be able to orient themselves—to determine a preferred direction—by experiments internal to their space station. They could do this only by referring to weak effects from distant objects, such as the light of distant stars or the small residual gravity of Earth. Any experiment that is not sensitive to these small effects will give the same result, whether it is done before or after a rotation of the space station and its contents as a whole.

Symmetries of space and time. Perhaps the most basic and profound symmetries of physical laws are symmetry under translation in time and under translations in space.

The statement that fundamental physical laws are symmetric under translation in time is equivalent to the statement that these laws (unlike human laws) do not change or evolve. Time-translation symmetry is supposed to apply, fundamentally, to simple isolated systems. Large complicated systems, and in particular the universe as a whole, do of course age and evolve. Thus in constructing the big-bang model of cosmology, it is assumed that the properties of individual electrons or protons do not change in time, although of course the state of the universe as a whole, according to the model, has changed quite drastically.

More precise and controlled tests of time-translation symmetry are based on observations of the spectral lines in light from distant stars, which was emitted long ago. Such observations indicate that the atoms in distant stars long ago obeyed, with great accuracy, the same laws as the atoms of the same elements on Earth do today. Other tests involve searches for anomalies in the Moon's orbit, which could indicate changes in the strength of gravity; and geochemical evidence involving naturally occurring nuclear reactions early in the Earth's history, whose character is highly sensitive to the powers of the fundamental interactions.

The statement that fundamental laws are symmetric under translations in space is another way of formulating the homogeneity of space. It is the

statement that the laws are the same throughout the universe. It says that the astronauts in the previous example cannot infer their location by local experiments within their space station. To do this, they must look to weak effects from distant objects. The power of this symmetry is that it makes it possible to infer, from observations in laboratories on Earth, the behavior of matter anywhere in the universe. Used in conjunction with symmetry under time translation, it is the foundation of astrophysics, and indeed of all scientific predictions about the behavior of matter at places and times besides here and now.

The symmetry of physical law under rotations, mentioned above, embodies the isotropy of space. It is very accurately tested by measurements of radiation from atoms, which reveal no preferred direction.

In the mathematical formulation of dynamics, there is an intimate connection between symmetries and conservation laws. Symmetry under time translation implies conservation of energy; symmetry under spatial translations implies conservation of momentum; and symmetry under rotation implies conservation of angular momentum. *See* ANGULAR MOMENTUM; CONSERVATION LAWS (PHYSICS); CONSERVATION OF ENERGY; CONSERVATION OF MOMENTUM.

The fundamental postulate of the special theory of relativity, that the laws of physics take the same form for observers moving with respect to one another at a fixed velocity, is clearly another statement about the symmetry of physical law. The idea that physical laws should be unchanged by such transformations was already clearly discussed by Galileo, who illustrated it by an observer's inability to infer motion while on a calm sea voyage in an enclosed cabin. The novelty of Einstein's theory arises from combining this velocity symmetry with a second postulate, deduced from experiments, that the speed of light is a universal constant and must take the same value for both stationary and uniformly moving observers. To accommodate both postulates, it was necessary to recognize that what the moving observer identifies as the interval of time between two events is a mixture of the distance and time intervals identified by the stationary observer. In the mathematical formulation of the theory, the space-time transformations which implement velocity symmetry (the Lorentz transformations) are closely analogous to the transformations among purely spatial coordinates which implement ordinary rotation symmetry. *See* GALILEAN TRANSFORMATIONS; LORENTZ TRANSFORMATIONS; RELATIVITY.

The existence of objects with particular locations invalidates complete translation symmetry, but for crystals a large and useful residue remains. A very small displacement of a crystal lattice does not lead to an identical object, but a finite translation, which takes the lattice into itself, does. Whereas complete translation symmetry implies the conservation of momentum (as mentioned above), the partial symmetry characteristic of a lattice gives a weaker conservation law. Specifically, the crystal symmetry allows a discrete set of possible changes in momentum for electromagnetic radiation that scatters from it. However, insofar as the crystal does not change its state, it provides a time-translation invariant background, and the energy of the radiation (that is, its frequency and wavelength) cannot change. Thus x-rays scattering from a crystal are deflected in a discrete set of directions, determined by the underlying crystal symmetry. This is the principle underlying the determination of crystal structures by x-ray diffraction. *See* CRYSTAL STRUCTURE; X-RAY DIFFRACTION.

Discrete symmetries. Before 1956, it was believed that all physical laws obeyed an additional set of fundamental symmetries, denoted P, C, and T, for parity, charge conjugation, and time reversal, respectively. At that time, subtle experiments involving the particles known as K mesons led to the suggestion that P might be violated in the weak interactions, and violations were indeed observed. This discovery eventually led to questioning—and in some cases overthrow—of other cherished symmetry principles.

Parity, P, roughly speaking, transforms objects into the shapes of their mirror images. If P were a symmetry, the apparent behavior of the images of objects reflected in a mirror would also be the actual behavior of corresponding real objects. This abstract concept can be made concrete by referring to a specific experimental arrangement, which tests the validity of P symmetry. A K^+ meson is observed to decay into a positron and a neutrino. The positron has the property of spin, which for present purposes may be visualized as an intrinsic rotation of the positron about an axis passing through its center. It is observed that positrons emitted in K^+ decay are almost always right-handed in the sense that their rotation is in the direction the fingers of a right hand curl if the thumb points in the direction of motion. *See* SPIN (QUANTUM MECHANICS).

If this whole situation were reflected in a mirror, the image positrons would appear to be left-handed. But the real positrons were right-handed. Hence the mirror-image particles do not behave in the same way as real particles. Thus P is not a valid symmetry of physical law. *See* PARITY (QUANTUM MECHANICS).

Charge conjugation, C, changes particles into their antiparticles. It is a purely internal transformation; that is, it does not involve space and time. If the laws of physics were symmetric under charge conjugation, the result of an experiment involving antiparticles could be inferred from the corresponding experiment involving particles. Thus, from the fact that K^+ decays into right-handed positrons and neutrinos, it would be inferred from C symmetry that K^- decays into right-handed electrons and antineutrinos at the same rate. In fact, the decay of K^- into electrons and antineutrinos does occur at very nearly the predicted rate, but the electrons are left-handed. Thus C is not a valid symmetry of physical law.

Remarkably, by combining the transformations P and C, thus simultaneously performing a mirror reflection in space and a change of particles into antiparticles, a result is obtained, CP, which is much

more nearly a valid symmetry than either of its components separately. However, in 1964 it was discovered experimentally that even *CP* is not quite a valid symmetry.

Although the preceding discussion has emphasized the failure of *P*, *C*, and *CP* to be precise symmetries of physical law, both the strong force responsible for nuclear structure and reactions and the electromagnetic force responsible for atomic structure and chemistry do obey these symmetries. Only the weak force, responsible for beta radioactivity and some relatively slow decays of exotic elementary particles, violates them. Thus these symmetries, while approximate, are quite useful and powerful in nuclear and atomic physics. Furthermore, the violation of these symmetries within the weak interactions is intricately patterned and structured. Elucidating this pattern was a key step in arriving at the modern theory of these interactions. It was found, in a sense that can be made precise, that *P* and *C* symmetry violation by the weak interaction is maximal. Indeed, at a fundamental level, only left-handed quarks and leptons, and right-handed antiquarks and antileptons, participate in this interaction. *See* ELECTROWEAK INTERACTION; FUNDAMENTAL INTERACTIONS; LEPTON; QUARKS; WEAK NUCLEAR INTERACTIONS.

The operation of time-reversal symmetry, *T*, involves changing the direction of motion of all particles. For example, it relates reactions of the type $A + B \rightarrow C + D$ to their reverse $C + D \rightarrow A + B$. No direct violation of *T* has been detected. A powerful test of *T* symmetry is obtained by studying the response of neutrons to an electric field. If the spin axes of neutrons were to align with the direction of an applied electric field, *T* symmetry would be violated, since reversing the direction of motion changes the direction of spin. So far, no such effect has been found. *See* TIME REVERSAL INVARIANCE.

Time-reversal symmetry, even if valid, applies in a straightforward way only to elementary processes. It does not, for example, contradict the one-way character of the second law of thermodynamics, which states that entropy can only increase with time. The consequences of *T* for macroscopic systems are built up from its microscopic foundation, by using the locality of physical interactions. They include the Onsager reciprocity relations, which relate different transport processes. *See* THERMODYNAMIC PRINCIPLES; TIME, ARROW OF.

Fundamental principles of quantum field theory suggest that the combined operation *PCT*, which involves simultaneously reflecting space, changing particles into antiparticles, and reversing the direction of time, must be a symmetry of physical law. Existing evidence is consistent with this prediction. *See* CPT THEOREM.

Internal symmetry. Internal symmetries, like *C*, do not involve transformations in space-time but change one type of particle into another. An important, although approximate, symmetry of this kind is isospin or i-spin symmetry. It is observed experimentally that the strong interactions of the proton and neutron are essentially the same. This regularity explains, for example, why the most stable light nuclei contain equal numbers of protons and neutrons. Such combinations are the most symmetric, and therefore it is understandable that they are the most energetically favorable. *See* I-SPIN.

Now, the most long-range component of the strong interaction can be considered as being generated by the exchange of π mesons or pions. Charged π^+ and π^- mesons are relatively long-lived, and were discovered soon after pions were proposed. However, it was shown theoretically that these two mesons were not enough to allow a proper symmetry transformation that transformed neutrons into protons and vice versa, and also transformed π mesons among themselves, in such a way as to leave the strong interaction invariant. A third, electrically neutral π^0 meson was therefore proposed, and a particle with the predicted properties was duly found. *See* MESON.

This was the prototype of several other successful predictions of the existence and properties of new particles, based on postulates of internal symmetries. Perhaps the most notable was the prediction of the mass and properties of the Ω^- baryon, based on an extension of the symmetry group SU(2) of isospin to a larger approximate SU(3) symmetry acting on strange particles as well. These symmetries were an important hint that the fundamental strong interactions are at some level universal, that is, act on all quarks in the same way, and thus paved the way toward modern quantum chromodynamics, which does implement such universality. *See* BARYON; GROUP THEORY; UNITARY SYMMETRY.

Much simpler mathematically than SU(2) internal symmetry, but quite profound physically, is U(1) internal symmetry. It illustrates the general relationship between symmetry and conservation laws alluded to above, for it is uniquely related to the conservation of some quantum number. For concreteness, the important case of the electric charge quantum number will be considered. The action of the U(1) internal symmetry transformation with parameter λ is to multiply the wave function of a state of electric charge q by the factor $e^{i\lambda q}$. An amplitude between two states with electric charges q_r and q_s will therefore be multiplied by a factor $e^{i\lambda(q_s-q_r)}$. Since the physical predictions of quantum mechanics depend on such amplitudes, these predictions will be unchanged only if the phase factors multiplying all nonvanishing amplitudes are trivial. This will be true, in turn, only if the amplitudes between states of unequal charge vanish; that is, if charge-changing amplitudes are forbidden, which is just a backhanded way of expressing the conservation of charge. Conversely, given an additive conservation law, the same argument could be reversed to construct an internal U(1) symmetry. Thus symmetry under an internal U(1) transformation is equivalent to the existence of an additive conservation law. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; QUANTUM NUMBERS.

Localization of symmetry. The concept of local gauge invariance, which is central to the standard model of fundamental particle interactions, and in

a slightly different form to general relativity, may be approached as a generalization of the U(1) internal symmetry transformation, where a parameter λ independent of space and time appears. Although it is mathematically consistent to have such a parameter, it goes against the spirit of locality, according to which each point in space-time has a certain independence. There is therefore reason to consider the possibility of a more general symmetry, involving a space-time-dependent transformation in which the wave function is multiplied by $e^{i\lambda(x,t)q(x,t)}$, where $q(x,t)$ is the density of charge at the space-time point (x,t). These transformations are much more general than those discussed above, and invariance under them leads to much more powerful and specific consequences.

It might be expected that a space-time-dependent symmetry requires the amount of charge at every point in space to be conserved separately; that is, charge can never move. If, in addition, symmetry is demanded under velocity transformations, as required by special relativity, then the charge must simply vanish everywhere. Although this is not valid for electromagnetic charge, it does crudely describe the behavior of color charges in the strong interaction. Indeed, the fundamental quarks and gluons, which carry color charges, cannot exist as free particles—they are said to be confined—essentially as a result of local color symmetry. *See* GLUONS; QUANTUM CHROMODYNAMICS.

For electromagnetism, the situation is more subtle. Roughly speaking, charged particles can move, consistent with the space-time-dependent symmetry transformation specified above, if they leave behind a compensating disturbance in the electromagnetic field. Indeed, if this very restrictive and demanding idea is implemented, the required interactions of matter with the electromagnetic field are predicted precisely. Thus, the theory of the electromagnetic field—Maxwell's equations and quantum electrodynamics—can be said to be the unique ideal embodiment of the abstract concept of a space-time-dependent symmetry, that is, of local gauge symmetry. *See* MAXWELL'S EQUATIONS; QUANTUM ELECTRODYNAMICS.

As mentioned above, it is similarly true that the modern theory of the strong interaction arises from implementing a conceptually simple generalization of this symmetry. A mathematical discussion of isospin symmetry involves symmetry transformations that are rotations in a two-dimensional space. The approximate conservation laws and relations among interactions required for invariance under this symmetry follow from a generalization of U(1) symmetry, with parameter λ, where λ is allowed to become a fixed 2×2 (hermitian) matrix. By analogy, it is interesting to consider allowing this λ-matrix to depend on space and time. This sort of theory was first proposed in 1954, but had serious difficulties. These were resolved by later theoretical work, especially in the early 1970s, that led directly to the modern standard model. *See* STANDARD MODEL.

Likewise, general relativity is closely associated with localization of symmetries, namely the symmetries of space-time translation, rotation, and velocity. The construction of a theory invariant under this wide symmetry group involves new issues, and in particular the specification of a preferred space-time field for the measure of distance, the metric $g_{\mu\nu}$. Nevertheless, in spite of important technical differences between gravity and the other interactions, the forms of all the known fundamental interactions can be said to be dictated by local symmetry principles. *See* GAUGE THEORY.

Guidance to physical law. As mentioned above, symmetry arguments have played an important, if implicit, role in the discussion of physical law since the beginnings of modern science. The conscious use of symmetry as a guide to physical law is, however, a twentieth-century phenomenon. Symmetry provided guidance toward the current synthesis, the standard model, discussed above. Today symmetry dominates much thinking about possible new physics beyond the standard model.

Modern gauge theories of the strong, weak, and electromagnetic interactions have very similar mathematical structures, suggesting that their symmetries might be synthesized into a larger encompassing symmetry. The idea is that the larger "unified" symmetry is spontaneously broken, so that while it is not valid for experiments done at the comparatively low energies which are practically accessible, it becomes more accurate for experiments at high energy or in describing events during the early moments of the big bang. *See* GRAND UNIFICATION THEORIES; SYMMETRY BREAKING.

Supersymmetry postulates invariance under transformations which change the spin of elementary particles. As for unifying gauge symmetries, it is fruitful to consider the possibility that supersymmetry, although not accurately valid, is spontaneously broken. There is striking semiquantitative evidence that the observed strong, weak, and electromagnetic interactions derive from spontaneous breakdown of a theory that is not only unified but also supersymmetric. Finally, superstring theory promises to incorporate these ideas in a framework that also allows for unification with the gravitational interaction. *See* ELEMENTARY PARTICLE; SUPERSTRING THEORY; SUPERSYMMETRY.

Frank Wilczek

Bibliography. S. Dimopoulos, S. Raby, and F. Wilczek, Unification of couplings, *Phys. Today*, 44(10):25–33, October 1991; J. E. Lannutti and P. K. Williams (eds.), *Current Trends in the Theory of Fields: A Symposium in Honor of P. A. M. Dirac*, 1978; J. Sakurai, *Invariance Principles and Elementary Particles*, 1969; E. Wigner, *Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra*, 1959.

Symmorphosis

A theory of structural design of biological organisms that postulates that structure is quantitatively matched to functional demand as a result of regulated

morphogenesis during growth and maintenance. Symmorphosis is a theory of economic design. In biological organisms, all functions depend on structural design, specifically on the morphometric characteristics of the organs. In general terms, the larger the structure the greater the functional capacity. A central postulate of symmorphosis, as a theory of economic design, is that differences in the functional demand on an organ require quantitative adjustments of its structural design parameters in order to match functional capacity to (maximal) functional demand.

Many functional systems comprise a sequence of steps, each supported by a different set of structures. For example, to support cell energetics oxygen must be supplied to the cells; this involves a chain of events from gas exchange in the lung, through cardiovascular transport of blood, to oxidative metabolism on the membranes of the cells' mitochondria. Symmorphosis postulates that quantitative design of each step is matched to overall functional demand. There should be enough structure but not too much, as it would be wasteful to design one step vastly in excess of the others.

The notion that animals, and humans, should be designed economically follows from common sense, but it is also supported by many observations. Thus blood vessel architecture ensures blood flow distribution with minimal energy loss; bone structure is patterned according to stress distribution and also quantitatively adapted to total stress; with training, athletes can specifically adjust the structure of their muscles and of their cardiovascular system to higher functional demands, and these modifications are soon reversed when training is stopped. These examples suggest optimization of structural design based on a cost-to-benefit relationship: the larger an organ, the more it will cost in terms of construction, maintenance, and load on the body. However, these examples are anecdotal, and one can clearly come up with similar anecdotal evidence that proves the opposite.

The validity of the concept of economic design as an important principle of structure-function relationship in biology can be tested by the hypothesis of symmorphosis which predicts that if the functional requirements on an organ system vary, structural design should vary in parallel. This hypothesis has been tested mainly on the mammalian respiratory system, where large natural variations in functional capacity are observed: the maximal rate of oxidative metabolism is much larger in small animals than in large animals (allometric variation) and in athletic species compared to more sedentary species (adaptive variation). The general result is that the hypothesis of symmorphosis is acceptable for most steps of the functional cascade, but the lung is a case where a certain, though not excessive, redundancy in design is noted.

The theory of symmorphosis postulates ideal adaptation of structural design to functional capacity. This is, however, hardly a reasonable assumption, because good engineering design of complex systems requires some redundancies as safety factors in view of imperfections in functional performance and variable boundary conditions. By using the concept of symmorphosis as a hypothesis, such deviations from idealized economic design can be detected.

Ewald R. Weibel

Bibliography. E. R. Weibel, C. R. Taylor, and H. Hoppeler, The concept of symmorphosis: A testable hypothesis of structure-function relationship, *Proc. Nat. Acad. Sci. USA*, 88:10357–10361, 1991; E. R. Weibel, C. R. Taylor, and H. Hoppeler, Variations in function and design: Testing symmorphosis in the respiratory system, *Resp. Physiol*, 87:325–348, 1992.

Sympathetic nervous system

The portion of the autonomic nervous system concerned with nonvolitional preparation of the organism for emergency situations. *See* AUTONOMIC NERVOUS SYSTEM.

The sympathetic nervous system is best understood in mammals. This system consists of two neuron chains from the thoracic and lumbar regions of the spinal cord to viscera and blood vessels. The first or preganglionic neuron has its cell body in the spinal cord and sends its axon to synapse with a postganglionic sympathetic neuron, which lies either in a chain of sympathetic ganglia paralleling the spinal cord or in a sympathetic ganglion near the base of the large blood vessels vascularizing the alimentary

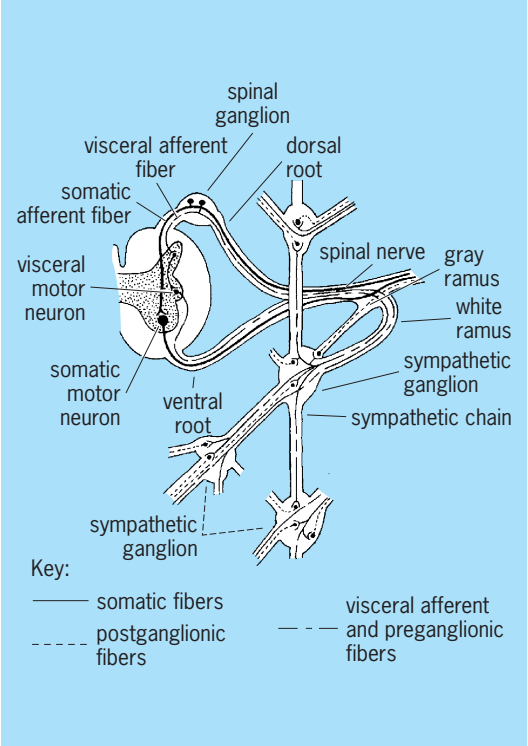


Fig. 1. Visceral reflex arc and the sympathetic chain. (*After B. A. Houssay et al., Human Physiology, 2d ed., McGraw-Hill, 1955*)

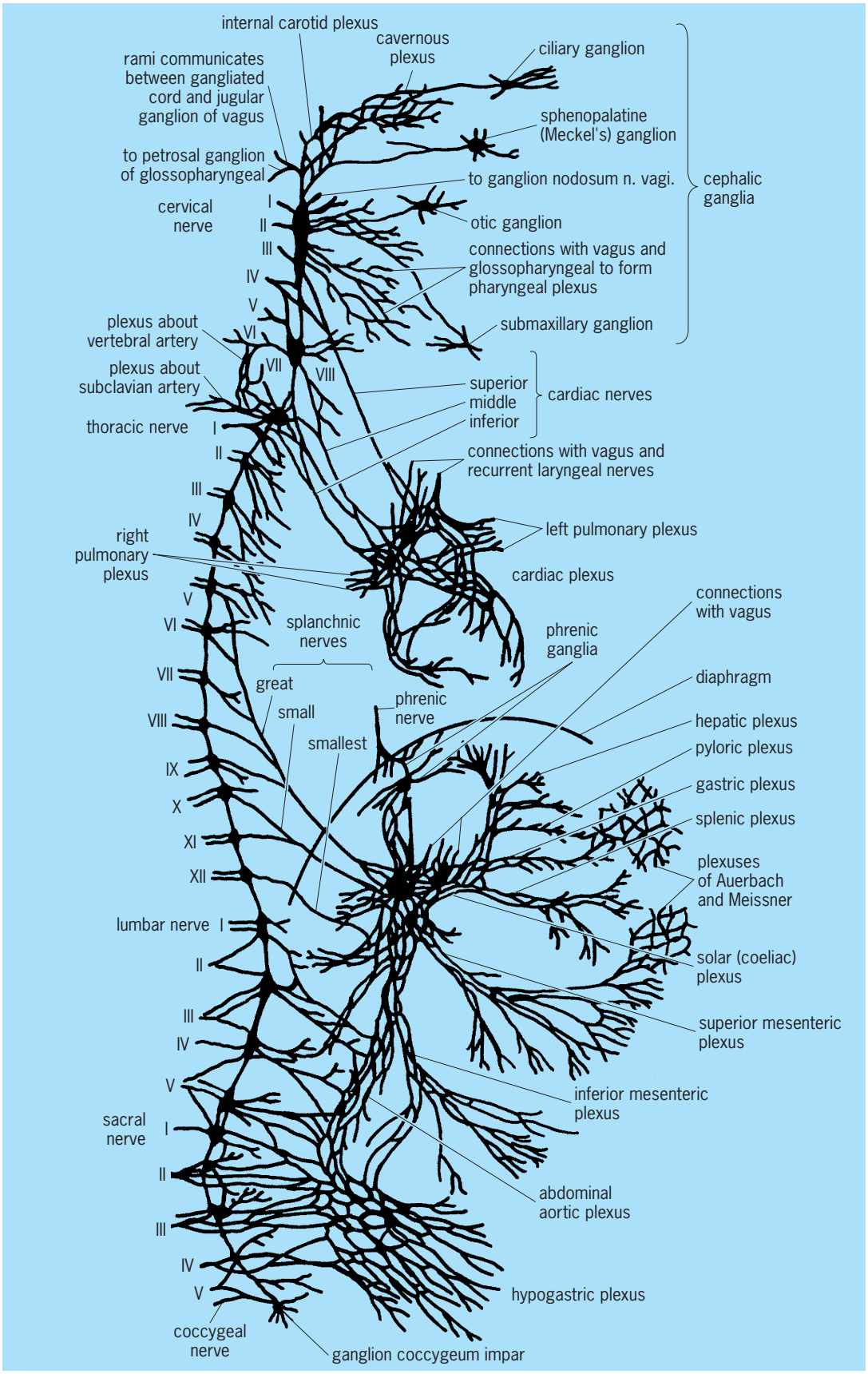


Fig. 2. Diagram showing the chain of ganglia and the great plexuses of the sympathetic system. (*After J. P. Schaeffer, ed., Morris' Human Anatomy, 11th ed., Blakiston-McGraw-Hill, 1953*)

viscera. The postganglionic axons are longer than the preganglionic axons and extend to glands or smooth muscles of viscera and blood vessels (**Fig. 1**). Sensory visceral nerve fibers innervate blood vessels and viscera and carry sensory information to the spinal cord, thus providing a visceral reflex arc (**Fig. 2**).

Among nonmammalian vertebrates, the sympathetic nervous system of birds and reptiles is similar to that in mammals, but it is somewhat different in anamnoites. In amphibians and bony fish the sympathetic neurons arise from the entire length of the cord. In cartilaginous fish (such as sharks) and agnathans (such as lampreys) the sympathetic nervous system is still not well understood and apparently not discretely separated. *See* NERVOUS SYSTEM (VERTEBRATE); PARASYMPATHETIC NERVOUS SYSTEM.

Douglas B. Webster

Sympathetic vibration

The driving of a mechanical or acoustical system at its resonant frequency by energy from an adjacent system vibrating at this same frequency. Examples include the vibration of wall panels by sounds issuing from a loudspeaker, vibration of machinery components at specific frequencies as the speed of a motor increases, and the use of tuned air resonators under the bars of a xylophone to enhance the acoustic output. Increasing the damping of a vibrating system will decrease the amplitude of its sympathetic vibration but at the same time widen the band of frequencies over which it will partake of sympathetic vibration. *See* MECHANICAL VIBRATION; RESONANCE (ACOUSTICS AND MECHANICS); VIBRATION.

Lawrence E. Kinsler

Symphyla

A class of the Myriapoda. The symphylans, like the pauropods, are tiny, pale, centipedelike creatures that inhabit humus or soil, or live under debris; in general, they live wherever there is sufficient moisture to preclude excessive water loss. They are similar to the Pauropoda and Diplopoda in being progoneate (having reproductive tracts open at anterior end of body) and anamorphic (gradually increasing in complexity of form and function). Each of their mandibles, like those of millipedes, but unlike the simple pauropod mandible, bears a movable gnathal lobe; at the same time their two pairs of maxillae are more reminiscent of the chilopods and lower insects than of the singly maxillate millipedes and pauropods. Additional signal characteristics include the following: The antennae are unbranched and simple; there is 1 pair of spiracles arising in the head and opening into tracheae; there are 12 pairs of legs, 1 pair per body segment; most of the legs have peculiar basal eversible vesicles with associated styli: the tergites number at least 15 and do not form diplotergites; there is a prominent pair of terminal spinnerets.

As is the case for the Pauropoda, little is known about symphylan biology. It is established that they feed upon decaying material as well as upon living plants; their role as greenhouse pests is widely appreciated by agriculturalists. Presumably all symphylans hatch with a reduced number of legs, six or seven in the species investigated, and thereafter undergo molts not only until the adult complement is gained but throughout life, which means 4–5 years in some forms.

The class consists of three families to which not more than 60 species have been assigned.

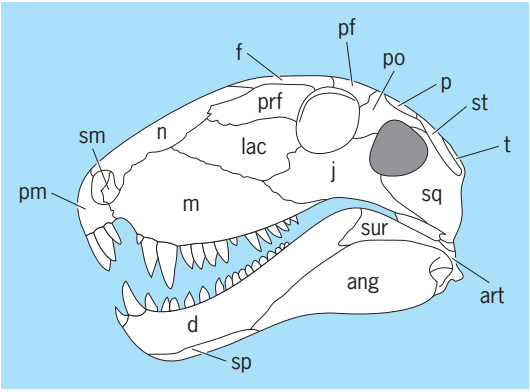
Ralph E. Crabill

Bibliography. G. Eisenbeis and W. Wichard, Symphyla, pp. 170–179 in *Atlas on the Biology of Soil Arthropods*, Springer-Verlag, Berlin, 1987; G. O. Poinar and C. A. Edwards, First description of a fossil symphylan, *Scutigerella dominicana* sp. n. (Scutigerellidae, Symphyla), in Dominican amber, *Experientia*, 51:391–393, 1995; U. Scheller, A new troglobitic species of Hanseniella Bagnall (Symphyla: Scutigerellidae) from Tasmania, *Austral. J. Entomol.*, 35:203–207, 1996; U. Scheller and J. A. Adis, Pictorial key for the Symphylan families and genera of the Neotropical region south of Central Mexico (Myriapoda, Symphyla), *Stud. Neotrop. Fauna Environ.*, 31:57–61, 1996.

Synapsida

The clade that includes all living and extinct species that share a closer relationship with modern mammals than with modern reptiles. The evolutionary history of synapsids is usually described in terms of three major radiations. The oldest synapsids are Middle Pennsylvanian in age (approximately 320 million years ago [Ma]) and belong to the earliest, "pelycosaur" radiation. Pelycosaur-grade synapsids are known primarily from Texas, Oklahoma, and surrounding states, although several fossils have been discovered in Europe. By the Middle Permian (260 Ma), a group of advanced synapsids, the therapsids, had achieved nearly complete distribution across the supercontinent of Pangea. As with their pelycosaur forebears, therapsids diversified into both carnivorous and herbivorous species, some growing up to 3–4 m (10–13 ft) in length.

Therapsid fossils are best known from Russia and South Africa, although finds in Antarctica, Brazil, China, India, and North America demonstrate their cosmopolitan nature. At the end of the Permian Period (251 Ma), a major mass extinction killed off numerous therapsid groups and left others with far fewer representatives. One group of therapsids that did survive, however, was the cynodonts. This group is of particular importance because it includes the ancient ancestors of mammals. The final radiation of synapsids is the mammals, which first appear in the fossil record in the Late Triassic and continue to the present. Traditional taxonomies group the



Skull of the pelycosaur-grade synapsid *Dimetrodon* in lateral view. The shaded region represents the lateral temporal fenestra, a diagnostic feature of all synapsids, including modern mammals. Abbreviations: ang, angular; art, articular; d, dentary; f, frontal; j, jugal; lac, lacrimal; m, maxilla; n, nasal; p, parietal; pf, postfrontal; pm, premaxilla; po, postorbital; prf, prefrontal; sm, septomaxilla; sp, splenial; sq, squamosal; st, supratemporal; sur, surangular; t, tabular. (*Reconstruction modified from B. S. Rubidge and C. A. Sidor, Evolutionary patterns among Permo/Triassic therapsids, Annu. Rev. Ecol. Systemat., 32:449–480, 2001*)

pelycosaurs and therapsids together as the "mammal-like reptiles," emphasizing that they possess some, but not all, of the features of mammals. All synapsid skulls, however, are characterized by a single temporal opening, bordered in early forms by the jugal, squamosal, and postorbital bones (see **illustration**). *See* ANIMAL EVOLUTION; MAMMALIA; PELYCOSAURIA; REPTILIA; THERAPSIDA.

Christian A. Sidor

Bibliography. M. J. Benton, *Vertebrate Palaeontology*, 2005; R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1988.

Synaptic transmission

The physiologic mechanisms by which one nerve cell (neuron) influences the activity of an anatomically adjacent neuron with which it is functionally coupled. Brain function depends on interactions of nerve cells with each other and with the gland cells and muscle cells they innervate. The interactions take place at specific sites of contact between cells known as synapses. The synapse is the smallest and most fundamental information-processing unit in the nervous system. By means of different patterns of synaptic connections between neurons, synaptic circuits are constructed during development to carry out the different functional operations of the nervous system. In order to understand nervous function, it is first necessary to understand the mechanism of transmission at a single synapse, and then the varieties of synapses out of which synaptic circuits are formed. *See* NEURON.

Electrical synapses. The simplest type of synapse is the electrical synapse (**Fig. 1***a*), which consists of an area of unusually close contact between two cells packed with channels that span the two membranes and the cleft between them. Because the cleft persists at these sites, electrical synapses are also known as gap junctions. Electrical and metabolic communication between two cells is established by the components of the gap junctions. Six subunits of the polypeptide connexin form a half-channel in each membrane, and the two half-channels are joined end to end across the intercellular gap to form the complete channel. A variety of influences, including calcium ions (Ca^{2+}), pH, membrane potential, neurotransmitters, and phosphorylating enzymes, may act on the channels to regulate their conductance in one direction (rectification) or both directions. For example, the channels are closed by high concentrations of calcium ions or by low pH.

Electrical synapses are present throughout the animal kingdom in a variety of functional settings. In vertebrates, they are numerous in the central nervous systems of fish as well as in certain nuclei of the mammalian brain, in regions where rapid transmission and synchronization of activity is important. Electrical synapses also interconnect glial cells in the brain. *See* BIOPOTENTIALS AND IONIC CURRENTS.

Chemical synapses. The most prevalent type of junction between nerve cells is the chemical synapse (Fig. 1*b*). At chemical synapses there are no gap junctions. Instead, neurotransmitters are released from the presynaptic cell, diffuse across the synaptic cleft, and bind to receptors on the postsynaptic cell. Chemical synapses are found only between nerve cells or between nerve cells and the gland cells and muscle cells that they innervate.

Neuromuscular junction. The neuromuscular junction is a prototypical synapse. The mechanisms of transmission at a single synapse have historically been most closely studied at the neuromuscular junction, that is, the junction between the axon terminals of a motoneuron and a muscle fiber (**Fig. 2**). Muscle, like nerve cells, is electrically excitable and contracts in response to depolarization. The mechanisms responsible for neurotransmission at this synapse are representative of chemical synapses in general. Three basic elements constitute the synapse: a presynaptic process (in this case, the motoneuron axon terminal) containing synaptic vesicles; an end plate (a specialized site of contact between the cells); and a postsynaptic process (in this case, the muscle cell). In motoneurons the nerve terminal is large and contains many individual active zones, each consisting of a cluster of presynaptic vesicles along a dense barlike structure. Opposite each of these is a fold of the postsynaptic membrane containing receptors for the transmitter substance released from the presynaptic terminal.

The chemical substance that serves as the transmitter at the vertebrate neuromuscular junction is acetylcholine (ACh). Within the nerve terminal, acetylcholine is concentrated within small spherical vesicles (Fig. 2). At rest, these vesicles undergo exocytosis at low rates, releasing their acetylcholine in quantal packets to diffuse across the cleft and bind to and activate the postsynaptic receptors. Each quantum gives rise to a small depolarization of the postsynaptic membrane. These miniature end-plate

potentials are rapid, lasting only some 10 milliseconds, and small in amplitude, only some 500 microvolts, below the threshold for effecting any response in the muscle. They represent the resting secretory activity of the nerve terminal.

When an organism wants to move its muscles, electrical impulses known as action potentials are generated in the motoneurons. These are conducted along the axon and invade the terminal, causing a large depolarization of the presynaptic membrane. This opens special voltage-gated channels for calcium ions, which enter the terminal and bind to special proteins, causing exocytosis of one or two vesicles at many active zones simultaneously. Calcium ions are thus the necessary and crucial link between the electrical signals in the presynaptic neuron and the chemical signals sent to the postsynaptic neuron. An action potential causes as many as 100 of the 500 or so active zones at the neuromuscular junction of the frog to release a quantum of acetylcholine. The combined action of this acetylcholine on postsynaptic receptors sets up a large postsynaptic depolarization known as the end-plate potential, which has a time course of 20–30 milliseconds and an amplitude of 40–50 millivolts. This exceeds the threshold for generating an impulse in the surrounding membrane, and causes the muscle to contract. The action of acetylcholine at its receptor is terminated by an enzyme, acetylcholinesterase, which is present in the synaptic cleft and hydrolyzes the acetylcholine to acetate and choline.

In order for neurotransmission to continue after a round of exocytosis, synaptic vesicles have to be regenerated. Vesicles are rapidly and efficiently reformed in the nerve terminal by endocytosis, the process by which cells internalize extra plasma membrane. The machinery responsible for this endocytosis must sort proteins to include in the reforming synaptic vesicle membrane, generate the inward curvature of the nascent vesicle, and finally pinch the neck of the bud to release the newly formed vesicle into the nerve terminal. Specific neurotransmitter transporters within the membrane then fill the synaptic vesicle with appropriate neurotransmitter. The regenerated vesicle either returns to the plasma membrane where it rejoins the releasable pool of vesicles, or remains in the nerve terminal as part of a reserve pool. *See* ACETYLCHOLINE; MUSCLE.

Molecular mechanisms. Synaptic transmission requires coordinated interactions among many proteins found in the sending (presynaptic) and receiving (postsynaptic) compartments of the synapse. In the synaptic terminal, these include a special category of proteins known as SNAREs (synaptosome-associated protein receptors), which are small membrane-anchored proteins that are essential for membrane fusion and exocytosis. Specific SNAREs are found on synaptic vesicles (synaptobrevin or VAMP, molecular weight 18,000) and the presynaptic plasma membrane (syntaxin, molecular weight 35,000, and SNAP-25, molecular weight 25,000). SNAREs form tight alpha-helical complexes with each other that are thought to pull the membranes

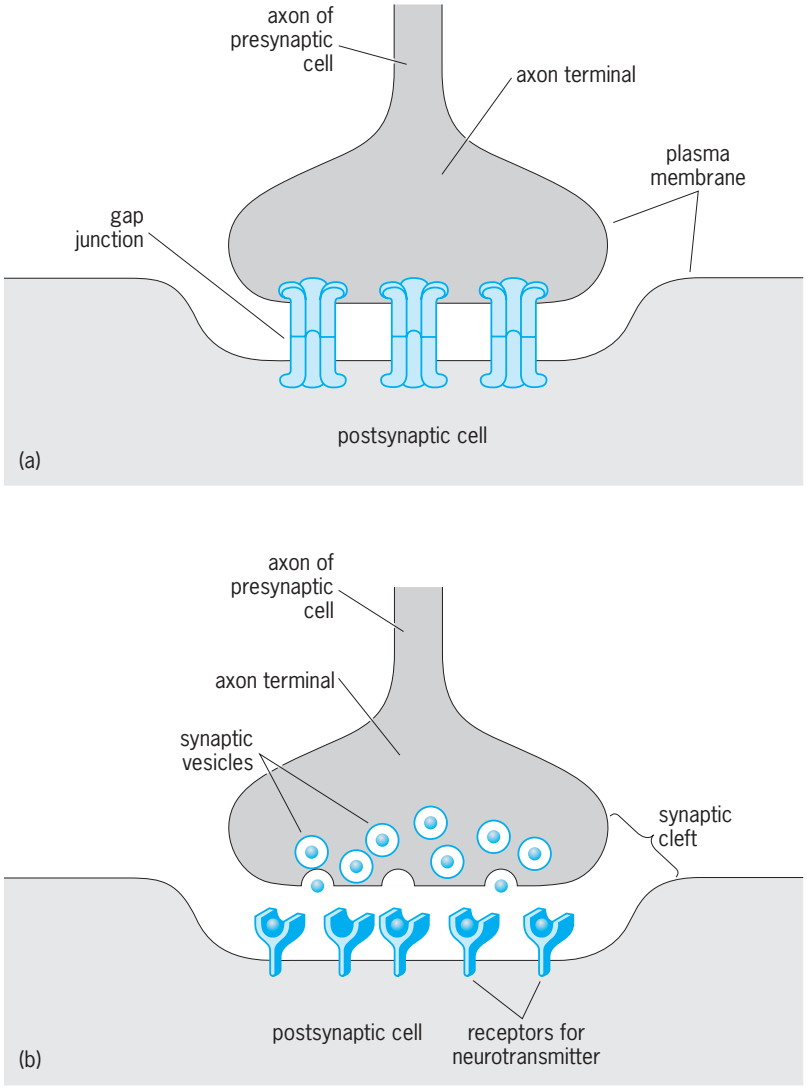


Fig. 1. Types of synapses. (*a*) An electrical synapse, showing the plasma membranes of the presynaptic and postsynaptic cells linked by gap junctions. (*b*) A chemical synapse, showing neurotransmitters released by the presynaptic cell diffusing across the synaptic cleft and binding to receptors on the postsynaptic membrane.

in which they are anchored close together and either facilitate or directly promote fusion, and hence exocytosis. A large number of other proteins, including a calcium-responsive trigger protein known as synaptotagmin, are also important in regulating vesicle exocytosis. Additional sets of proteins are responsible for endocytosis and regeneration of synaptic vesicles.

In the postsynaptic membrane, the target of the axon terminal's chemical signal is a neurotransmitter receptor; in the case of vertebrate muscle, this is the acetylcholine receptor, a membrane protein composed of five separate polypeptide chains or subunits. There are two alpha subunits, and beta, gamma, and delta subunits, totaling a molecular weight of approximately 260,000. The five subunits span the membrane and form a ring around a central Pore, which acts as a channel for the passage of ions across the membrane. When activated, the channel is cation-selective, allowing sodium (Na^+) and

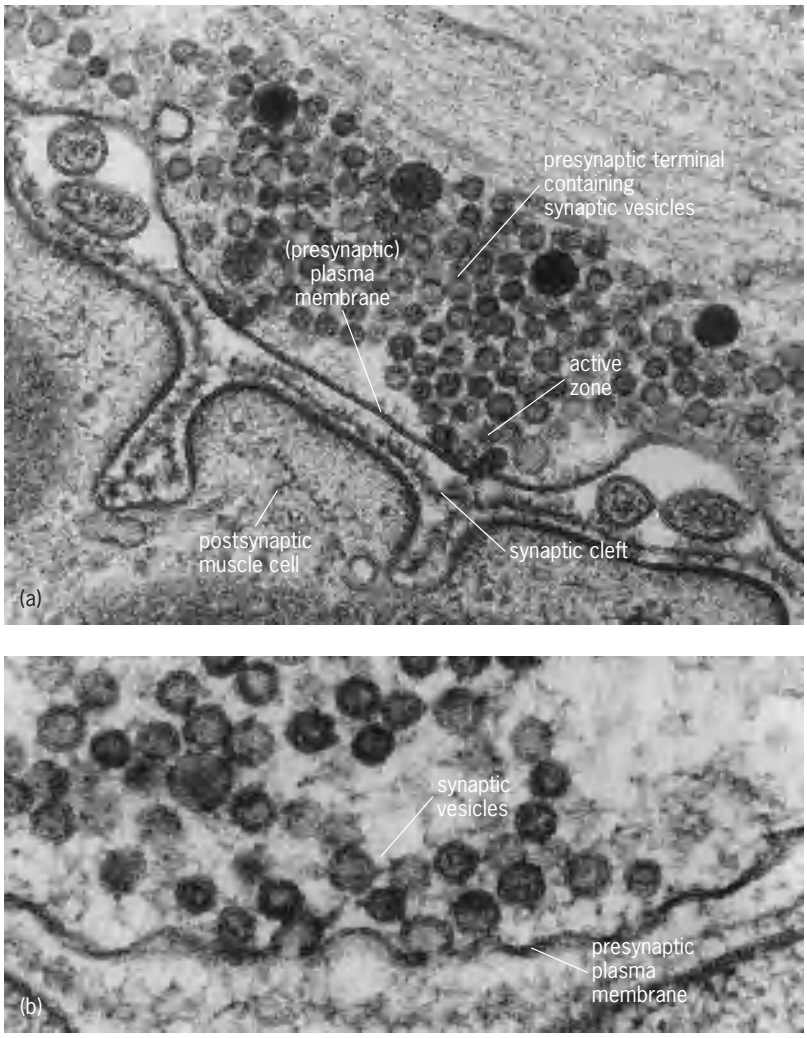


Fig. 2. Ultrastructure of exocytosis at the neuromuscular junction. (*a*) Thin-section electron micrograph of a neuromuscular junction from resting frog sartorius muscle. Vesicles are clustered in the presynaptic terminal at one of many active zones; some are docked at the plasma membrane (arrows). The postsynaptic muscle cell is seen across the synaptic cleft from the nerve terminal. (*b*) A similar presynaptic terminal frozen shortly (5 milliseconds) after electrical stimulation. Vesicles are seen in the process of fusing with the plasma membrane. (*Reproduced from J. Cell Biol., 88:564–580, 1981; copyright by The Rockefeller University Press*)

potassium (K^+) ions to pass through; under normal conditions this depolarizes the membrane potential toward zero.

Activation of the receptor occurs when two molecules of acetylcholine bind to each of the alpha subunits, causing an allosteric conformational change in the protein structure, which increases the ionic conductance of the channel. The acetylcholine receptor channel opens abruptly, has a relatively constant open conductance, and closes abruptly. These properties suggest that the protein jumps directly from being open to being closed in response to the presence or absence of acetylcholine.

Central synapses. The neuromuscular junction is an example of a large terminal with multiple synaptic sites. Most terminals, however, are small and contain essentially one active zone. In the central nervous system the presynaptic process containing synaptic vesicles is most often an axon terminal and the postsynaptic process a dendrite, making an axodendritic synapse, but other relationships are also seen. The effect of transmitter on a postsynaptic cell is either excitatory or inhibitory, meaning that it either depolarizes or hyperpolarizes the membrane. Whether a transmitter has an excitatory or inhibitory effect on a cell is determined by the type of ion able to pass through the cell's receptor channels. Most central synapses consist of two principal types.

Amino acid transmitters. Type 1 central synapses are characterized by small round presynaptic vesicles and a densification of the two apposed membranes that is asymmetrical, being much heavier postsynaptically. This type of synapse commonly releases an amino acid transmitter, such as glutamate, whose action produces an excitatory postsynaptic potential. At low levels of activity the glutamate binds to the glutamate receptor and activates a relatively small conductance increase for sodium and potassium ions. Glutamate is the transmitter at many excitatory synapses throughout the central nervous system.

GABA transmitters. The other type of central synapse, type 2, is characterized by presynaptic vesicles with varied shapes and a symmetrical densification of the apposed membranes. This type is usually associated with inhibitory synaptic actions. The most common inhibitory transmitter is gamma-amino butyric acid (GABA). Most inhibitory interneurons in different regions of the brain make these kinds of synapses on relay neurons in those regions. The GABA receptor is a complex channel-forming protein with several types of binding sites and several conductance states.

Other transmitters. Other types of central synapses utilize different transmitter substances, and there may be several types of receptor molecules for each of these substances. These can be distinguished from each other by the fact that they bind different agonistic substances, or are blocked by different antagonists. In the case of acetylcholine, at the neuromuscular junction the acetylcholine receptor binds nicotine, and is therefore called the nicotonic acetylcholine receptor, whereas at other synapses another type of acetylcholine receptor binds muscarine, and is known as the muscarinic acetylcholine receptor. At muscarinic synapses, second messengers phosphorylate a separate channel protein, bringing about a conformational change that activates a conductance that leads to an excitatory or inhibitory postsynaptic potential. This response has a slow time course, on the order of tenths of a second to several seconds. Muscarinic receptors are found especially at motoneuron synapses on gland cells and smooth muscle cells, and on specific populations of neurons in the forebrain. At these sites, the synapse may be much less clearly defined morphologically, and the transmitter may diffuse over long distances to act on receptor molecules.

Other classes of synaptic transmitter substances include the biogenic amines, such as the catecholamines, norepinephrine and dopamine, and

indoleamine 5-hydroxytryptamine (also known as serotonin). *See* DOPAMINE; NORADRENERGIC SYSTEM; SEROTONIN.

A final type of transmitter substance consists of a vast array of neuropeptides. These vary in size from simple dipeptides, such as carnosine, to complex molecules of 20 or more amino acids, which include some kinds of hormones. They are a diverse group and include substances that stimulate the release of hormones (such as thyrotropin-releasing hormone); those that act at synapses in pain pathways in the brain (the endogenous morphinelike substances, enkephalins and endorphins); and many of still undetermined functions (for example, cholecystokinin and vasoactive intestinal peptide). The release mechanisms and receptors for these peptides are still largely unknown, nor is it known whether they are confined to the sites of morphologically defined synapses. Peptides may act not only directly on receptors but also indirectly, modifying the state of a receptor in its response to other transmitter substances, and they may do this in an activity-dependent manner. Most of the known responses to peptides have long latencies of onset and long durations of action, lasting for several minutes or more. In view of the complexity and slow time course of many of their effects, these peptides are often referred to as neuromodulators. *See* ENDORPHINS; HORMONE.

G-proteins. At central synapses, rapid responses to transmitters are most commonly the result of direct synaptic transmission, in which the receptor itself is the ion channel. Such receptors (for example, the nicotinic acetylcholine receptor) are referred to as ionotropic, and typically have effects within milliseconds of being activated. There are, however, other receptor molecules for each of the neurotransmitters, and many of these are not themselves ion channels. These receptors (for example, the muscarinic acetylcholine receptor) are known as metabotropic receptors, and they affect neurotransmission indirectly via a set of intermediary proteins called G-proteins. G-proteins are so named because they bind to guanine nucleotides and change their molecular state depending on whether guanosine triphosphate (GTP) or guanosine diphosphate (GDP) is bound to the G-protein's alpha subunit. Activated G-proteins have a variety of effects on synaptic processes. Some are mediated by direct interactions with ion channels. However, many other effects are mediated by activation of cellular second messenger systems involving messengers such as calcium and cyclic adenosine monophosphate (cyclic AMP). The time courses for effects caused by activation of G protein-coupled receptors is much longer than that of ionotropic channels (milliseconds), reflecting the lifetime of the activated G-protein subunits (seconds) and second messengers (seconds to minutes). Such longer-lasting signals greatly increase the complexity of chemical neurotransmission and synaptic modulation.

Activity-dependent changes. The synapse is a dynamic structure whose function is very dependent on its activity state. At the neuromuscular junction, following vesicle release, the active zone undergoes periods of depression and potentiation that are increased in intensity by high-frequency stimulation. The number of receptors appears to be regulated by a mechanism in which a small percentage of receptors, when bound by acetylcholine, become desensitized; that is, they are unable to bind acetylcholine again. Increased synaptic activity causes an increase in the numbers of desensitized receptors, leading to decreases in the total number of receptors, a process known as down regulation. In this way the synapse is constantly adjusted for its information load.

At glutamatergic synapses, high levels of input activity bring about a different transmission state. The buildup of postsynaptic depolarization relieves the normal block of a specialized glutamate receptor channel by external magnesium ions, permitting influx of calcium ions into the postsynaptic process. Since the conductance of the channel is dependent on the depolarization state of the membrane, it is said to have a voltage-gated property, in addition to being ligand-gated by its transmitter. The calcium ion acts as a second messenger to bring about a long-lasting increase in synaptic efficacy, a process known as long-term potentiation. The conjunction of increased pre- and postsynaptic activity to give an increase in synaptic efficacy is called Hebbian, and is believed to be the type of plasticity mechanism involved in learning and memory. *See* LEARNING MECHANISMS; MEMORY.

Drug actions. The synapse is one of the primary targets of drug actions. The first example to be identified was the arrow poison curare, which blocks nicotinic neuromuscular transmission by binding to acetylcholine receptor sites. Many toxic agents have their actions on specific types of receptors; organic fluorophosphates, for example, are widely used pesticides that bind to and inactivate acetylcholinesterase. Most psychoactive drugs exert their effects at the synaptic level. Minor tranquilizers, such as the benzodiazapine antianxiety drugs, block GABA receptors among their actions. Antipsychotic drugs, on the other hand, appear to have in common the fact that they block dopamine transmission. This led to the hypothesis that schizophrenia is the result of increased activity at central dopamine synapses. Another theory has held that mental depression is due to a deficiency of catecholamine transmitters, especially norepinephrine; in this theory, mania would be due to increased activity at these synapses. However, all these mental syndromes are much more complex, and most of the transmitter systems are involved to varying extents. *See* NERVOUS SYSTEM (VERTEBRATE); NEUROBIOLOGY.

Gordon M. Shepherd; Phyllis I. Hanson

Bibliography. J. G. Nicholls et al., *From Neuron to Brain*, 4th ed., 2000; G. M. Shepherd, *Neurobiology*, 3d ed., 1994; G. Siegel et al., *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, 5th ed., 1994; M. J. Zigmond et al., *Fundamental Neuroscience*, 1999.

Synbranchiformes

An order of eel-like fishes in the class Actinopterygii, commonly called swamp eels. Synbranchiforms are among the most highly specialized fishes. They are characterized by a usually scaleless eel-like body lacking pelvic fins with pectoral fins present or absent; gill openings entirely ventral and often confluent across the breast; premaxillae lacking ascending processes and nonprotrusible; ectopterygoid bones enlarged; endopterygoids reduced or absent; swim bladder present or absent; and gills poorly developed, with respiration in some species accomplished in part by highly vascularized buccopharyngeal pouches. The order has no fossil record and comprises three families, 15 genera, and about 99 species, all but three of which are freshwater inhabitants. Synbrachiformes are thought to form a monophyletic group with the smegmamorphs. *See* ACTINOPTERYGII.

Family Synbranchidae (swamp eels). Swamp eels (**Fig. 1**) are characterized by a lack of pectoral fins, except in the early development of some species; vestigial dorsal and anal fins; a caudal fin, if present, that is small or vestigial; scales usually absent; small eyes with some species blind; widely separated nostrils; a gill opening that is a small slit or pore; absence of a swim bladder and ribs. They are the only teleosts to possess an amphistylic jaw suspension (the upper jaw is attached directly to the skull). They have a maximum total length of 150 cm (59 in.). Most species are protogynous hermaphrodites (first females, then males). Most species can breathe air, with some of them highly evolved in this capacity by having lunglike suprabranchial pouches. Swamp eels are primarily freshwater (all but three species), living in swamps, in caves (with blind species found in the Yucatan Peninsula of Mexico and Liberia), and in sluggish waters of tropical and subtropical western Africa, Asia, the Indo-Australian Archipelago, and Mexico to South America. The species *Monopterus albus* was introduced into inland waters of the United States. Swamp eels are capable of overland excursions, and some species can live out of water for extended periods by burrowing in the substrate at the onset of a dry season. The family comprises four genera and 17 species.

Families Chaudhuriidae and Mastacembelidae. These two families have a physoclistic swim bladder (that is, the air bladder is not connected with the digestive tract by an open duct); the pectoral girdle is attached to the vertebral column by ligaments; the posttemporal bone is absent; and the dorsal and anal fins are continuous with a small caudal fin. *See* SWIM BLADDER.

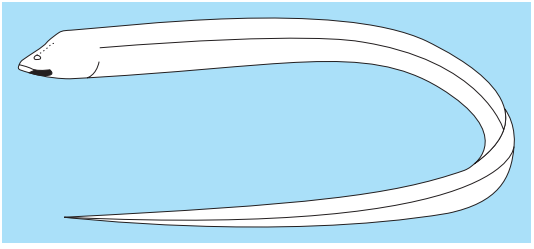


Fig. 1. Example from the Synbranchidae (swamp eels).

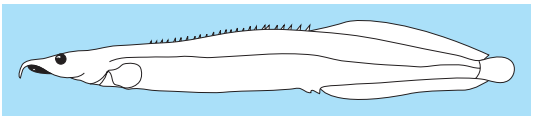


Fig. 2. Example from the Mastacembelidae (spiny eels).

Members of the Chaudhuriidae (earthworm eels) are distinguished by a lack of dorsal and anal fin spines; a usually naked body; absence of a lateral line, a rostral appendage, and endopterygoid and epineural bones; presence of a basisphenoid bone; and a maximum length of about 8 cm (3.2 in.). The six genera and nine species inhabit freshwaters of northeastern India and through Thailand to Korea, including parts of Malaysia and Borneo.

The Mastacembelidae (spiny eels) (**Fig. 2**) differ from other synbranchiforms by having from 9 to 42 individual (that is, isolated, not attached to each other by membranes) spines preceding the soft dorsal fin, which has from 52 to 131 rays and a fleshy rostral appendage. They are further distinguished by a body covered with small scales (naked body in perhaps three species); no basisphenoid bone; and maximum length of 90 cm (35 in.). In addition to occupying a wide variety of habitats, some species can remain buried in the substrate for long periods of time in order to escape dry periods. The five genera and 73 species occupy the freshwaters of tropical Africa and through Syria to the Malay Archipelago, China, and Korea.

Herbert Boschung

Bibliography. R. Britz, S. Favorito, and G. D. Johnson, The hyopalatine arch of a 25 mm larva of *Synbranchus* and homology of the single pterygoid in the Synbranchidae (Teleostei: Synbranchiformes), *Proc. Biol. Soc. Wash.*, 116(2):337–340, 2003; P. T. Fuller, L. G. Nico, and J. D. Williams, Nonindigenous fishes introduced into inland waters of the United States, *Amer. Fish. Soc. Spec. Publ.*, no. 27, Bethesda, MD, 1999; G. D. Johnson and C. Patterson, Percomorph phylogeny: A survey of acanthomorphs and a new proposal, *Bull. Mar. Sci.*, 52(1):554–626, 1993; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Syncarida

A unique superorder of malacostracan Crustacea, noted particularly for the total lack of a carapace or carapace shield. Two orders are recognized, Anaspidacea and Bathynellacea. The Stygocarididae, at one time considered a distinct order, are now assigned familial rank within the Anaspidacea.

The syncarid body plan is simple, consisting of a cephalon (head and sometimes one thoracic somite), thorax of seven or eight somites, and an abdomen of six somites and telson or five somites and pleotelson (fused sixth somite-telson). Eyes, with apposition

optics, may be stalked or sessile, or sometimes absent. There are six to eight thoracic appendages (thoracopods), usually two to five abdominal appendages (pleopods), and uropods. Pleopods may be entirely lacking, or the first one or two pairs may be modified as copulatory appendages. Sexes are separate. Young hatch as immature adults and pass through several molts before reaching maturity. Studies have suggested that bathynellids, at least, arose by neoteny from a zoealike malacostracan ancestor similar to the Penaeidae.

Extant syncarids are specially adapted to freshwater, often interstitial environments. Whereas the Anaspidacea are restricted to the Southern Hemisphere (Australia, Tasmania, New Zealand, South America), bathynellids have been found worldwide, except in Antarctica. *See* BATHYNELLACEA; CRUSTACEA; MALACOSTRACA. Patsy A. McLaughlin

Bibliography. L. Botosaneanu (ed.), *Stygofauna munidi*, 1986; P. A. McLaughlin, *Comparative Morphology of Recent Crustacea*, 1980; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 1982; H. K. Schminke, Adaptation of Bathynellacea (Crustacea, Syncarida) to life in the interstitial ("Zoea Theory"), *Int. Rev. Ges. Hydrobiol.*, 66(4):575–637, 1981.

Synchronization

The process of maintaining one operation in step with another. The commonest example is the electric clock, whose motor rotates at some integral multiple or submultiple of the speed of the alternator in the power station. In television, synchronization is essential in order that the electron beams of receiver picture tubes will be at exactly the same spot on the screen at each instant as is the beam in the television camera tube at the transmitter. Synchronism in television is achieved by transmitting a synchronizing pulse at the end of each scanning line, to make all receivers move simultaneously to the start of the next line. A similar vertical synchronizing pulse is transmitted when the camera beam reaches the bottom of the picture, to make all beams go back to the top for the start of the next field. *See* OSCILLOSCOPE; TELEVISION. John Markus

Synchronous converter

A synchronous machine used to convert alternating current (ac) to direct current (dc), or vice versa. The ac-to-dc converter has been superseded by solid-state converters using silicon controlled rectifiers (SCRs) or power MOSFETs (for reasons of efficiency, lower maintenance costs, and less trouble) or by motor-generator sets. Converters are no longer manufactured, but there are converters still in use. *See* POWER INTEGRATED CIRCUITS; SEMICONDUCTOR RECTIFIER.

The synchronous converter has an armature similar to those used in dc machines, with connections to a commutator and brushes on the dc end and to slip rings on the ac end. The field is similar to that of a dc machine and may have shunt and series windings on the main poles, interpoles with series windings to assist commutation, and amortisseur windings to aid starting and to prevent hunting. The machine is thus a combination of a synchronous motor and a dc generator, having all the essential elements of each but with no duplication of parts.

As compared with a motor-generator set consisting of a synchronous motor driving a dc generator, the synchronous converter has the following advantages: (1) a single magnetic circuit comprising field poles, yoke, armature core, and teeth; (2) a single armature in whose conductors the ac and dc currents tend to cancel, thereby reducing greatly the copper losses; (3) a single pair of bearings and a smaller and shorter shaft; (4) higher efficiency, lower first cost, and lighter weight. There are, however, attendant disadvantages: (1) The ratio of dc to ac voltage is fixed and cannot be changed materially by adjusting the field excitation. (2) The machine is quite sensitive to sudden load changes and may hunt or flashover its commutator. (3) The heating in the armature is not the same for all conductors. (4) Power-factor correction by field adjustment is quite limited. *See* DIRECT-CURRENT GENERATOR; SYNCHRONOUS MOTOR. Loyal V. Bewley

Synchronous motor

An alternating-current (ac) motor which operates at a fixed synchronous speed proportional to the frequency of the applied ac power supply. A synchronous machine can be operated as either a generator or a motor (and sometimes as a source of reactive power, fulfilling the function of a capacitor), depending only on its applied shaft torque (whether positive, negative, or zero) and its excitation. There is no fundamental difference in the theory, design, or construction of a machine intended for any of these roles, although certain design features are stressed for each of them. In use, the machine may change its role from instant to instant. As a result, it is preferable to set up a common general theory for synchronous generators, motors, and capacitors. The distinction between generator and motor is merely a difference in the direction of the currents and the sign of the torque angles. *See* ALTERNATING-CURRENT GENERATOR; ALTERNATING-CURRENT MOTOR.

Basic theory. A single-phase, two-pole synchronous machine is shown in **Fig. 1**. The coil is on the pole axis at time $t = 0$, and the sinusoidally distributed flux Φ linked with the coil at any instant is given by Eq. (1), where ωt is the angular displacement of the coil and Φ_{max} is the maximum value of the flux. This flux will induce in a coil of N turns an instantaneous voltage e, given by Eq. (2).

$$\phi = \Phi_{max} \cos \omega t \tag{1}$$

$$e = -N\frac{d\phi}{dt} = \omega N \Phi_{max} \sin \omega t = E_{max} \sin \omega t \tag{2}$$

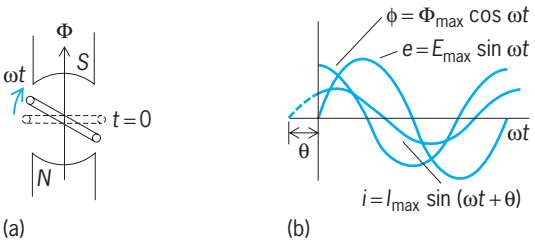


Fig. 1. Single-phase, two-pole synchronous machine. (*a*) Configuration of coil and poles. (*b*) Variation of flux Φ linking coil, voltage *e*, and current *i* with time.

The effective (rms) value E of this voltage is given by Eq. (3).

$$E = \frac{E_{\max}}{\sqrt{2}} = \sqrt{2}\,\pi f N \Phi_{\max} = 4.44 f N \Phi_{\max} \quad (3)$$

Armature reaction. Assume that the impedance of the coil and its external circuit of resistance R_t and reactance X is given by Eq. (4). There will flow a current,

$$Z = R_t \pm jX = Z\underline{/\pm\theta} \quad (4)$$

with a value given by Eq. (5), in which the phase

$$\mathbf{I} = \frac{\mathbf{E}}{\mathbf{Z}} = \frac{E}{Z}\underline{/\mp\theta} \quad (5)$$

angle ϕ is taken positive for a leading current. This current will develop a sinusoidal space distribution of armature reaction as in Eq. (6). If this single-phase

$$A = 0.8 N I_{\max} \sin(\omega t + \theta) \quad (6)$$

mmf is expressed as a space vector and resolved into direct (in line with the pole axis) A_d and quadrature A_q components, it is given by Eq. (7).

$$\begin{aligned}\mathbf{A} &= A_d + jA_q \\ &= 0.4 N I_{\max} \{[\sin\theta + \sin(2\omega t + \theta)] \\ &\quad + j[\cos\theta - \cos(2\omega t + \theta)]\}\end{aligned} \quad (7)$$

Polyphase armature reaction. In a three-phase machine, assuming balanced currents, the phase currents are given by Eqs. (8). Upon writing Eq. (7) for ωt,

$$\begin{aligned} i_a &= I_{\max} \sin(\omega t + \theta) \\ i_b &= I_{\max} \sin(\omega t + \theta - 120^\circ) \\ i_c &= I_{\max} \sin(\omega t + \theta - 240^\circ) \end{aligned} \quad (8)$$

$\omega t - 120^\circ$, and $\omega t - 240^\circ$, respectively, and adding, Eq. (9) results for the polyphase armature reaction.

$$\mathbf{A} = A_d + jA_q = 1.2 N I_{\max} (\sin\theta + j\cos\theta) \quad (9)$$

Equation (10) gives the three-phase power of the machine, and Eq. (11) gives the developed torque.

$$P = 3EI\cos\theta \quad (10)$$

$$T = \frac{P}{\omega} = \frac{3}{\omega} EI\cos\theta \quad (11)$$

The above equations constitute the essential description of the operation of a synchronous generator. The same equations apply for a motor if the currents are reversed, that is, by changing the sign of the current I. They may also be interpreted in the form of phasor diagrams, and show the two cases of a smooth-rotor and a salient-pole machine.

Smooth-rotor synchronous machine. In the smooth-rotor machine, the reluctance of the magnetic path is essentially the same in either the direct or quadrature axes. In **Fig. 2***a* let the flux Φ be selected as reference phasor and drawn vertically. Then comparing Eqs. (1) and (2) it is seen that the induced voltage E_f lags the flux by 90°. By Eq. (5) the current I lags the voltage by an angle θ for an inductive circuit, and by Eq. (9) causes a constant mmf of armature reaction A in phase with the current. This armature reaction causes a flux ϕ_a, stationary in space with respect to the field poles, which in turn induces a voltage E_a lagging it by 90°. The two induced voltages E_f (due to the field flux Φ) and E_a (due to the armature reaction flux ϕ_a) combine vectorially to give the resultant voltage E'. But the terminal voltage V is less than E' by the resistance and reactance drops, $R\mathbf{I}$ and $jx_l\mathbf{I}$ in the winding, and Eq. (12) applies.

$$\mathbf{V} = \mathbf{E}' - (R + jx_l)\mathbf{I} \quad (12)$$

The leakage reactance drop $jx_l\mathbf{I}$ lags the current by 90° as does the armature reaction voltage E_a. If a fictitious reactance of armature reaction x_a is introduced to account for E_a, it is obvious that Eq. (12) may be rewritten to give Eqs. (13), in which

$$\begin{aligned}\mathbf{V} &= E_f - jx_a\mathbf{I} - (R + jx_l)\mathbf{I} \\ &= E_f - R\mathbf{I} - j(x_a + x_l)\mathbf{I} \\ &= E_f - (R + jX_s)\mathbf{I}\end{aligned} \quad (13)$$

$X_s = x_a + x_l$ is called the synchronous reactance of the machine.

Salient-pole synchronous machine. In a similar fashion the phasor diagram for a salient-pole machine, Fig. 2*b*, may be set up. Here the effects of saliency result in proportionately different armature reaction fluxes in the direct and quadrature axes, thereby necessitating corresponding direct, X_d, and quadrature, X_q, components of the synchronous reactance. The angle δ in Fig. 2 is called the torque angle. It is the angle between the field-induced voltage E_f and the terminal voltage V and is positive when E_f is ahead of V.

The foregoing equations and phasor diagrams were established for a generator. A motor may be regarded as a generator in which the power component of the current is reversed 180°, that is, becomes an input instead of an output current. The motor vector diagram is shown in **Fig. 3**. Here the torque angle δ is reversed, since V is ahead of E_f in a motor (it was behind in the generator). Therefore a motor differs from a generator in two essential respects: (1) The currents are reversed, and (2) the torque angle has changed sign. As a result the power input, Eq. (10), for a motor is negative, or has become a power output, and the torque is reversed in sign.

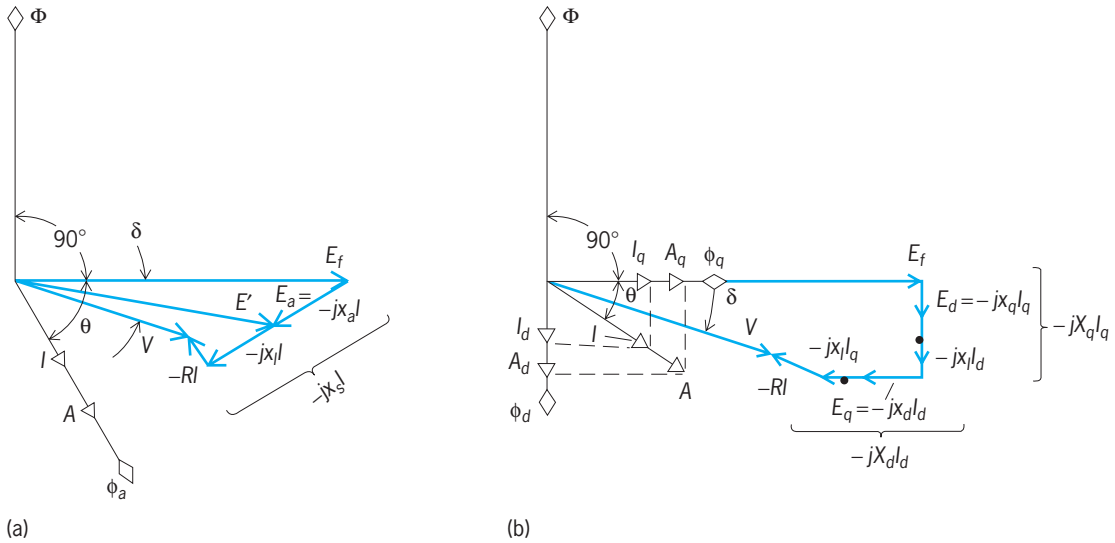


Fig. 2. Vector diagrams of synchronous generators. (*a*) Smooth-rotor machine. (*b*) Salient-pole machine.

When the current **I** is 90° out of phase with the terminal voltage V, the torque angle δ is nearly zero, being just sufficient to account for the power lost in the resistance.

Therefore, a synchronous machine is a generator, motor, or capacitor, depending on whether its torque angle δ is positive, negative, or zero. For these conditions the output current is, respectively, at an angle in the first or fourth quadrant, second or third quadrant, or essentially $\pm 90°$ with respect to the terminal voltage at 0°. For any given power input or output, the machine can be made to operate at either leading or lagging power factor by changing the magnitude of the field current producing the flux Φ.

Synchronous capacitor. A synchronous capacitor can be made to draw a leading current and to behave like a capacitance by overexciting its field. Or, it will draw a lagging current on underexcitation. This characteristic thus presents the possibility of power-factor correction of a power system by adjusting the field excitation. A machine so employed at the end of a transmission line permits a wide range of voltage regulation for the line. One used in a factory permits the power factor of the load to be corrected. Of course, a synchronous motor can also be used for power-factor correction, but since it must also carry the load current, its power-factor correction capabilities are more limited than for the synchronous condenser. *See* STATIC VAR COMPENSATOR.

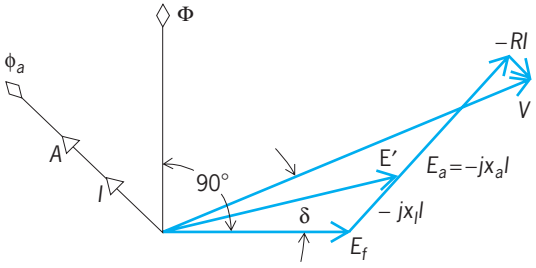


Fig. 3. Phasor diagram of synchronous motor.

Power equations. The power output P_o and reactance power output Q_o of a round-rotor synchronous machine are given by Eqs. (14), in which $\tan \alpha = R/X_s$ and $\mathbf{Z}_s = R + jX_s = Z_s < 90° - \alpha$.

$$P_o = \frac{VE_f}{Z_s} \sin(\delta + \alpha) - \frac{RV^2}{Z_s^2}$$
$$Q_o = \frac{VE_f}{Z_s} \cos(\delta + \alpha) + \frac{X_s V^2}{Z_s^2} \tag{14}$$

For a round-rotor motor, both the torque angle δ and the electrical power output are negative.

For a salient-pole machine, neglecting resistance, Z_s is equal to the direct-axis reactance X_d, α is zero, and P_o is given by Eq. (15). Thus the power or torque

$$P_o = \frac{VE_f}{X_d} \sin\delta + V^2 \frac{X_d - X_q}{2X_d X_q} \sin 2\delta \tag{15}$$

depends essentially on the product of the terminal and induced voltages and sine of the torque angle δ; but in the case of the salient-pole machine there is also a second harmonic term which is independent of the excitation voltage E_f. This term, the so-called reluctance power, vanishes for nonsaliency when $X_d = X_q$. The small synchronous motors used in some electric clocks and other low-torque applications depend solely on this reluctance torque.

Excitation characteristics. The so-called V curves of a synchronous motor are curves of armature current plotted against field current with power output as parameter. Usually a second set of curves with input power factor (pf) as parameter is superimposed on the same plot. Such curves (**Fig. 4**), where armature current is plotted against generated voltage, can be determined from design calculations or from test; they yield a considerable amount of data on the performance of the motor. Thus, given any two of the four variables E_f, I, pf, mechanical power P_m, the remaining two may be easily determined, as well as the conditions of maximum power,

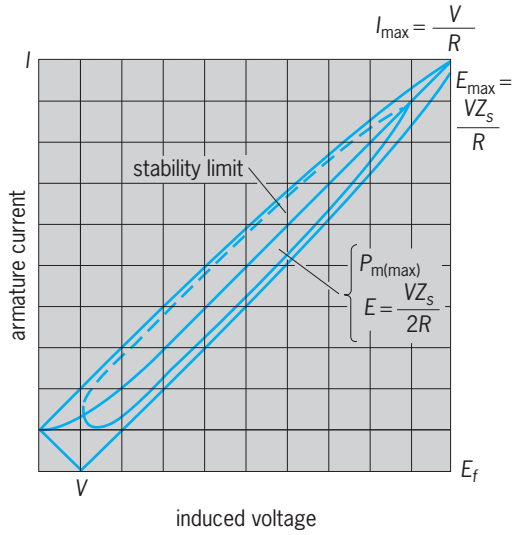


Fig. 4. V curves (armature current versus induced voltage) of synchronous motor.

constant pf, minimum excitation, stability limit, and so forth.

Losses and efficiency. The losses in a synchronous motor comprise the copper losses in the field, armature, and amortisseur windings; the exciter and rheostat losses of the excitation system; the core loss due to hysteresis and eddy currents in the armature core and teeth and in the pole face; the stray loss due to skin effect in conductors; and the mechanical losses due to windage and friction. The efficiency of the motor is then given by Eq. (16).

$$\text{Eff} = \frac{\text{output}}{\text{input}} = \frac{\text{output}}{\text{output} + \text{losses}} \tag{16}$$

Mechanical oscillations. A synchronous motor subjected to sudden changes of load, or when driving a load having a variable torque (for example, a reciprocating compressor), may oscillate about its mean synchronous speed. Under these conditions the torque angle δ does not remain fixed, but varies. As a result the four separate torques expressed in Eq. (17)

$$\begin{pmatrix}\text{Synchronous}\\ \text{motor torque}\\ \text{Eq. (15)}\end{pmatrix} + \begin{pmatrix}\text{induction motor}\\ \text{torque of}\\ \text{amortisseur}\end{pmatrix} = \begin{pmatrix}\text{torque to}\\ \text{overcome}\\ \text{inertia}\end{pmatrix} + \begin{pmatrix}\text{torque}\\ \text{required}\\ \text{by the load}\end{pmatrix} \tag{17}$$

act on the machine rotor. The possibility exists that cumulative oscillations will build up and cause the motor to fall out of step.

Starting of synchronous motors. Synchronous motors are provided with an amortisseur (squirrel-cage) winding embedded in the face of the field poles. This winding serves the double purpose of starting the motor and limiting the oscillations or hunting. During starting, the field winding is either closed through a resistance, short-circuited, or opened at several points to avoid dangerous induced voltages. The amortisseur winding acts exactly as the squirrel-cage winding in an induction motor and accelerates the motor to nearly synchronous speed. When near synchronous speed, the field is excited, and the synchronous torque pulls the motor into synchronism. During starting, Eq. (17) applies, since all four types of torque may be present. Of course, up to the instant when the field is excited the portion of the synchronous motor torque depending on E_f does not exist, although the reluctance torque will be active.

Other methods of starting have been used. If the exciter is direct-connected and a dc source of power is available, it may be used to start the synchronous motor. In the so-called supersynchronous motor the stator is able to rotate in bearings of its own, and is provided with a brake band. For stator allowed to come up to nearly synchronous speed by virtue of the amortisseur windings; the field is then excited and the stator brought to synchronous speed, the rotor remaining stationary. Then as the brake band is tightened, the torque on the rotor causes it to accelerate while the speed of the stator correspondingly slackens; finally the stator comes to rest and is locked by the brake band. In this way maximum synchronous motor torque is made available for acceleration of the load.

For other types of synchronous motors *see* HYSTERESIS MOTOR; RELUCTANCE MOTOR.

Richard T. Smith

Bibliography. D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; A. E. Fitzgerald, C. Kingsley, Jr., and S. D. Umans, *Electric Machinery*, 6th ed., 2003; L. W. Matsch and J. D. Morgan, *Electromagnetic and Electromechanical Machines*, 3d ed., 1986.

Synchrotron radiation

Electromagnetic radiation emitted by relativistic charged particles following a curved path in magnetic or electric fields. Synchrotron radiation is generally produced in laboratories by a beam of relativistic (moving near the speed of light) electrons or positrons in a generally circular path in a specialized synchrotron called a storage ring. A storage ring produces electromagnetic radiation with high flux, brightness, and coherent power levels. Synchrotron radiation is used for a wide variety of basic and applied research in biology, chemistry, and physics, as well as for applications in medicine and technology. *See* ELECTROMAGNETIC RADIATION; PARTICLE ACCELERATOR; RELATIVISTIC ELECTRODYNAMICS.

Electron storage rings provide radiation over a very broad range of photon energies or wavelengths. Radiation is generally produced in the infrared, visible, near-ultraviolet, vacuum-ultraviolet, soft x-ray, and hard x-ray parts of the electromagnetic spectrum, depending on the energy of the electrons in the storage ring and the strength of the magnetic field used to produce the radiation. Lasers are a much

Storage ring synchrotron radiation facilities*

Country	City	Facility	Energy, GeV
Armenia	Yerevan	CANDLE (Center for the Advancement of Natural Discoveries using Light Emission)	3.2‡
Australia	Melbourne	Australian Synchrotron	3†
Brazil	Campinas	LNLS (Laboratorio Nacional de Luz Sincrotron)	1.35
Canada	Saskatoon	CLS (Canadian Light Source)	2.9†
China	Beijing	BSRF (Beijing Synchrotron Radiation Facility)	1.5–2.8
China	Hefei	NSRL (National Synchrotron Radiation Laboratory)	0.8
China	Shanghai	SSRF (Shanghai Synchrotron Radiation Facility)	3.5‡
China, Republic of (Taiwan)	Hsinchu	NSRRC (National Synchrotron Radiation Research Center)	1.3-1.5†
Denmark	Aarhus	ASTRID (Aarhus Storage Ring in Denmark)	0.6
France	Grenoble	ESRF (European Synchrotron Radiation Facility)	6†
France	Gif-sur-Yvette	SOLEIL	2.75†
Germany	Berlin	BESSY II (Berliner Elektronenspeicherring-Gesellschaft für Synchrotronstrahlung)	1.7–1.9†
Germany	Bonn	ELSA (Electron Stretcher Accelerator)	1.5–3.5
Germany	Dortmund	DELTA (Dortmund Electron Test Accelerator)	1.5
Germany	Hamburg	DORIS (originally Double Ring Store) [at HASYLAB (Hamburger Synchrotronstrahlungslabor)]	4.5
Germany	Karlsruhe	PETRA (originally Positron-Electron Tandem Ring Accelerator Facility) [at HASYLAB]	7–14
Germany	Karlsruhe	ANKA (Angstromquelle Karlsruhe)	2.5†
India	Indore	INDUS I	0.45
India	Indore	INDUS II	2.5‡
Italy	Frascati	DAFNE	0.5
Italy	Trieste	ELETTRA	2–2.4†
Japan	Nishi Harima	SPRing-8	8†
Japan	Nishi Harima	New SUBARU	1.5†
Japan	Okasaki	UVSOR (Ultraviolet Synchrotron Orbital Radiation Facility)	0.75
Japan	Sendai	TSRF (Tohoku Synchrotron Radiation Facility)	1.5
Japan	Tsukuba	Photon Factory	2.5
Japan	Tsukuba	PF-AR (Photon Factory Accelerator Ring)	6.5
Jordan	Allaan	SESAME (Synchrotron-light for Experimental Science and Applications in the Middle East)	2.5‡
Korea	Pohang	PLS (Pohang Light Source)	2†
Russia	Dubna	DELSY (Dubna Electron Synchrotron)	0.95
Russia	Novosibirsk	VEPP-2	0.7
Russia	Novosibirsk	VEPP-3	2.2
Russia	Novosibirsk	VEPP-4	5–7
Singapore	Singapore	SSLS (Singapore Synchrotron Light Source): HELIOS 2	0.7
Spain	Barcelona	ALBA	3‡
Sweden	Lund	MAX I	0.55
Sweden	Lund	MAX II	1.5†
Switzerland	Villigen	SLS (Swiss Light Source)	2.4†
Thailand	Nakhon Ratchasima	NSRC (National Synchrotron Research Center)	1–1.3
Ukraine	Kiev	ISI-800	1‡
United Kingdom	South Oxfordshire	Diamond	3†
United States	Argonne, IL	APS (Advanced Photon Source)	7†
United States	Baton Rouge, LA	CAMD (Center for Advanced Microstructures & Devices)	1.3–1.5†
United States	Berkeley, CA	ALS (Advanced Light Source)	1.9†
United States	Gaithersburg, MD	SURF III (Synchrotron Ultraviolet Radiation Facility)	0.386
United States	Ithaca, NY	CHESS (Cornell High-Energy Synchrotron Source)	5.5
United States	Stanford, CA	SPEAR 3 (Stanford Synchrotron Radiation Center)	3†
United States	Stoughton, WI	SRC-Aladdin (Synchrotron Radiation Center)	0.8–1
United States	Upton, NY	NSLS (National Synchrotron Light Source)	0.8
United States	Upton, NY	NSLS II (National Synchrotron Light Source)	2.5–2.8‡

*Partial list of facilities as of June 2006, compiled from www.lightsources.org and other references.
†More recently constructed source, usually with performance of a "third-generation" facility.
‡Planned or under construction as of June 2006.

brighter source of radiation in the visible than are storage rings, so synchrotron radiation is only rarely used in the visible part of the spectrum. However, synchrotron radiation is particularly useful for producing x-ray beams.

Numerous storage ring sources of synchrotron radiation are in operation, construction, or design in many countries (see **table**). The flux [photons/(second, unit bandwidth)], brightness (or brilliance) [flux/(unit source size, unit solid angle)], and coherent power (important for imaging applications and proportional to brightness) available for experiments—particularly in the vacuum-ultraviolet, soft x-ray, and hard x-ray parts of the spectrum—are steadily increasing as higher-performance storage rings are constructed. A beam from a modern storage ring has flux, brightness, and coherent power many orders of magnitude higher than is available from other sources (**Fig. 1**).

Experimental Facilities

Synchrotron radiation has many features (natural collimation, high intensity and brightness, broad spectral bandwidth, high polarization, pulsed time

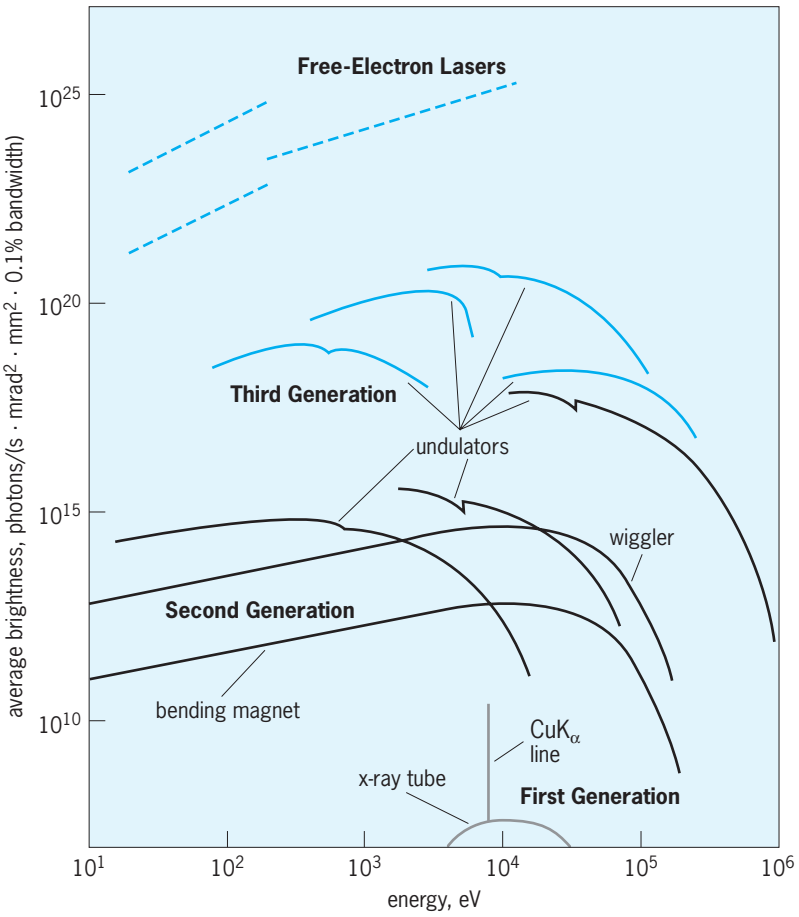


Fig. 1. Representative average brightness of synchrotron radiation sources and free-electron lasers as a function of photon energy. Present and extrapolated performances are shown.

structure, small source size, and high-vacuum environment) that make it ideal for a wide variety of applications in experimental science and technology. Synchrotron radiation was predicted theoretically early in the twentieth century, and first observed in 1947 at General Electric Laboratories. Very powerful sources of synchrotron radiation in the ultraviolet and x-ray parts of the spectrum became available when high-energy physicists began operating electron synchrotrons in the 1950s. Although synchrotrons produce large amounts of radiation, their cyclic nature results in pulse-to-pulse variations, rendering difficult their use in many types of research. By contrast, the electron-positron storage rings developed for colliding-beam experiments starting in the 1960s offered a constant electron-beam intensity and much better stability. Beam lines to bring synchrotron radiation from a storage ring to users were constructed on both synchrotrons and storage rings to allow the radiation produced in the bending magnets of these machines to leave the ring vacuum system and reach experimental stations. In most cases the research programs were pursued on a parasitic basis, secondary to the high-energy physics programs. These parasitic light sources are now referred to as first-generation light sources.

The brightness of x-ray sources has continued to increase since synchrotron radiation surpassed conventional x-ray sources (**Fig. 2**). The proportional rate of increase of the brightness of synchrotron-radiation sources is greater than that of computer processing speed.

Second-generation light sources. Storage rings dedicated as synchrotron light sources have been completed in many countries around the world (see table). They are called second-generation light sources to distinguish them from the first-generation synchrotrons that were built for research in high-energy physics.

Special arrays of magnets may be inserted into the straight sections between storage-ring bending magnets to produce beams with extended spectral range or with higher flux and brightness than is possible with the ring bending magnets. These devices, called wiggler and undulator magnets, utilize periodic transverse magnetic fields to produce transverse oscillations of the electron beam with no net deflection or displacement. They provide orders-of-magnitude increase in flux and brightness over storage-ring bending magnets (Fig. 1), again opening up new research opportunities.

Third-generation light sources. Third-generation light sources are storage rings with many straight sections for wiggler and undulator insertion device sources and with a smaller transverse size and angular divergence of the circulating electron beam. The product of the transverse size and divergence is called the emittance. The lower the electron-beam emittance, the higher the photon-beam brightness and coherent power level. With smaller horizontal emittances and with straight sections that can accommodate longer undulators, third-generation

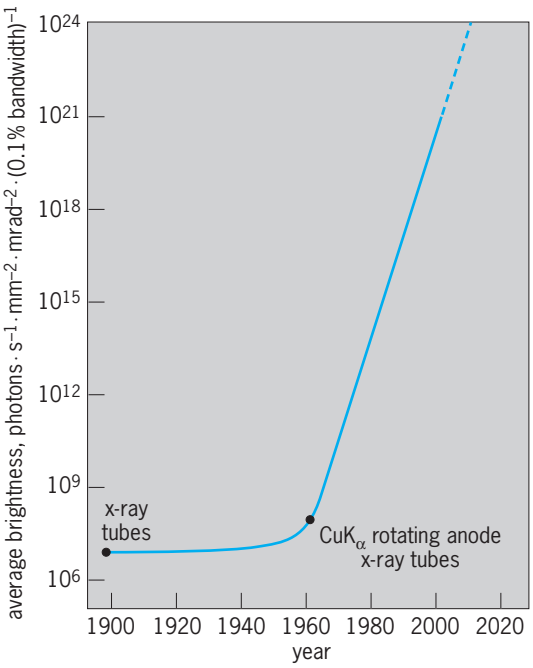


Fig. 2. Average brightness of x-ray sources from discovery of x-rays to the present (extrapolated to future).

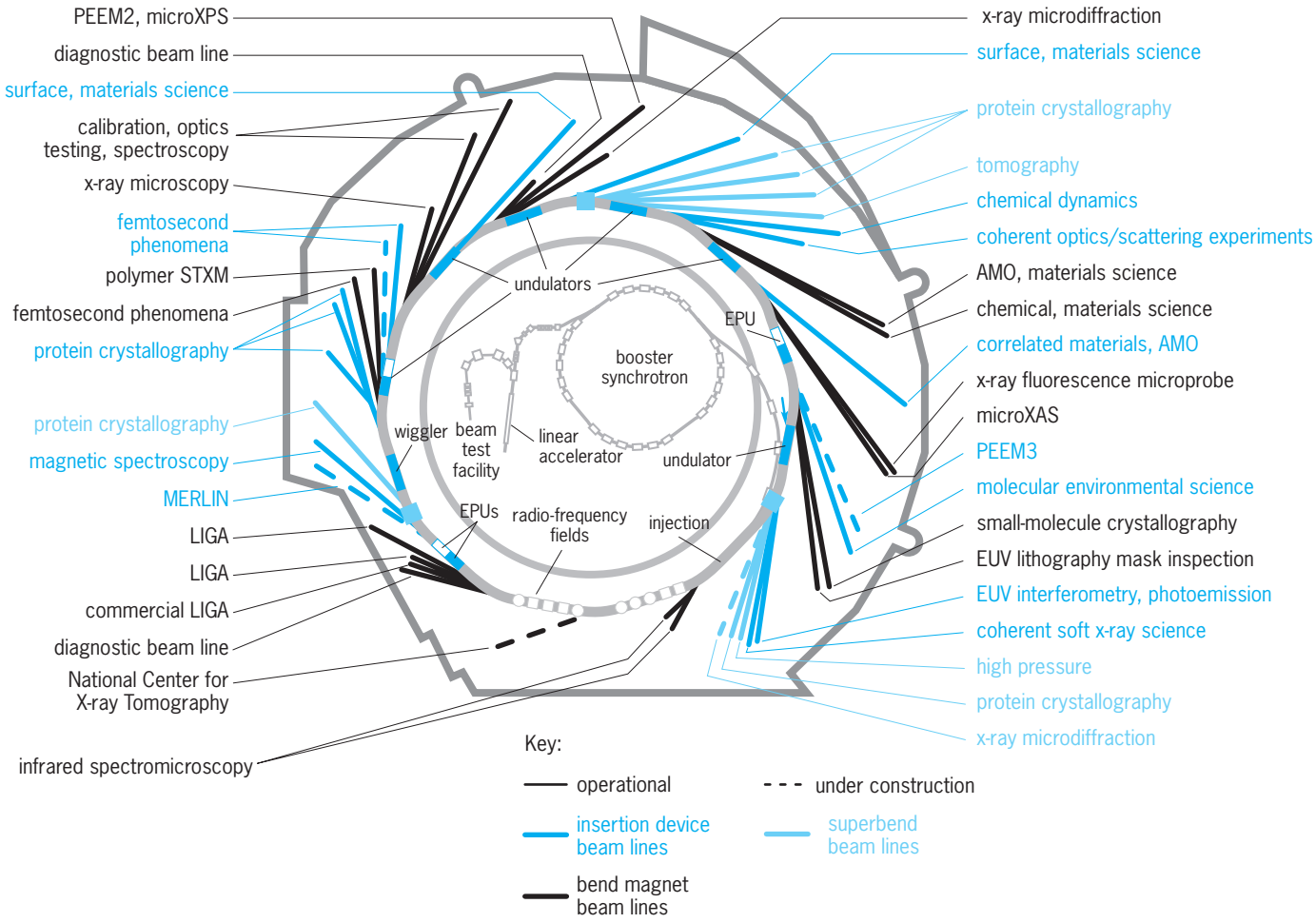


Fig. 3. Layout of the 1.9-GeV Advanced Light Source at Lawrence Berkeley Laboratory, a low-energy, third-generation synchrotron radiation source, as of 2006. Applications of experimental stations on beam lines are indicated. EPU = elliptically polarizing undulator; PEEM = photoemission electron microscope; XPS = x-ray photoelectron spectroscopy; STXM = scanning transmission x-ray microscopy; LIGA = Lithographie, Galvanoformung, und Abformung (lithography, electroforming, and molding); AMO = atomic, molecular, and optical physics; XAS = x-ray absorption spectroscopy; EUV = extreme ultraviolet.

rings provide two or more orders of magnitude higher brightness and coherent power level than earlier sources. One consequence of the extraordinary brightness (brilliance) of these light sources is that the x-ray beam is partially coherent, allowing research in interferometry and speckle spectroscopy, among other uses. *See* COHERENCE; INTERFEROMETRY; SPECKLE.

A considerable number of third-generation rings are in operation (see table). Third-generation light sources with a low-energy electron beam (typically 1–2 GeV) [**Fig. 3**] are generally optimized to produce high-brightness radiation in the vacuum ultraviolet (VUV) and soft (low-energy) x-ray spectral range, up to photon energies of about 2–3 keV. (Lower-brightness light extends to above 10 keV, and the use of superconducting bending magnets can extend the photon energy range to considerably higher values.) Third-generation light sources with a high-energy electron beam (6–8 GeV) can produce harder (higher-energy) x-rays, with energies of tens of kiloelectronvolts and above. Some of the newest storage rings that are intermediate in size, with beam energies up to 3 GeV, have even lower emittances than the first third-generation machines, and are equipped with advanced insertion-device technology with which they can also reach high brightness in the hard x-ray region.

Free-electron lasers. The next logical step in the evolution of synchrotron radiation sources will be a fully coherent source, namely an x-ray laser. A fully coherent source will have a peak brightness 10 orders of magnitude greater than existing third-generation light sources. To produce a source this brilliant requires a very small and well-collimated electron beam; the best way to achieve a particle beam with the required characteristics is not with a storage ring but with a linear accelerator (linac). The beam accelerated by the linac will pass through a very long undulator (approximately 100 m or 300 ft in length) and generate x-ray beams of very high brightness through a process called self-amplified stimulated emission (SASE). Along with a 10,000-fold increase in average x-ray brightness over third-generation sources, fourth-generation free-electron laser sources will have subpicosecond pulse

durations, resulting in x-ray beams with unparalleled instantaneous power levels. Several free-electron lasers are being built to produce vaccum ultraviolet and x-ray beams, such as the LCLS (Linac Coherent Light Source) at Stanford University. The unique properties of x-ray free-electron lasers, such as the instantaneous or peak brightness of a single ultrashort pulse, will surely open new fields of scientific research. *See* LASER.

Properties of Synchrotron Radiation

The radiation produced by an electron in circular motion at low energy (speed much less than the speed of light) is weak and rather nondirectional. At relativistic energies (speed close to the speed of light), the radiated power increases markedly, and the emission pattern is folded forward into a cone with a half-opening angle in radians given approximately by mc^2/E, where mc^2 is the rest-mass energy of the electron (0.51 MeV) and E is the total energy of the electron. Thus, at electron energies of the order of 1 GeV, much of the very strong radiation produced is confined to a forward cone with an instantaneous opening angle of about 1 mrad (0.06°). At higher electron energies this cone is even smaller. The large amount of radiation produced combined with the natural collimation gives synchrotron radiation its intrinsic high brightness. Brightness is further enhanced by the small cross-sectional area of the electron beam, which is as low as 0.01 mm^2 in the third-generation storage rings.

Bending magnet sources. As an electron moves in an arc through a bending magnet (**Fig. 4***a*), the instantaneous forward cone sweeps out a horizontal fan of continuum, polarized radiation. The small vertical opening angle, $2/\gamma$, is preserved, where γ is the ratio of the moving mass of the electron to its rest mass, but the horizontal angle is determined by the acceptance of the beam pipe or optical elements (for example, mirrors, gratings, or crystals). By accepting horizontal angles larger than $2/\gamma$, more flux can be utilized. The brightness, however, is not increased.

Wiggler magnet sources. The electron beam in a wiggler magnet executes transverse oscillations with an angular excursion much larger than $1/\gamma$, meaning excursions through angles larger than the instantaneous radiation. The result is an incoherent sum of radiation from each bend of the electrons. A fan of radiation is produced (Fig. 4*b*), which is similar to that produced by a bending magnet but with an enhancement of intensity due to the accumulation of radiation from each pole. The magnetic field can be higher than the storage-ring bending magnet field, resulting in a spectrum that extends to higher photon energy. Superconducting wigglers with fields up to 7.5 teslas have been used to extend the spectrum deeper into the hard x-ray region. Wigglers (and undulators) can be electromagnets or permanent magnets (**Fig. 5**). Permanent magnets are commonly used, particularly for devices with short periods and many poles, but superconducting electromagnets are useful when the highest magnetic fields are desired. *See* MAGNET.

Undulator magnet sources. An undulator magnet is designed to cause the electron beam to execute transverse oscillations with an angular excursion of the order of $1/\gamma$. The electron-beam angular deflection of the same order as the natural emission angle of synchrotron radiation results in a coherent superposition of radiation from each deflection of the electron beam. The radiation emerging from an undulator is concentrated into a narrow, fowarddirected cone with the smallest possible opening angle (Fig. 4*c*). Brightness is increased over that of bend-magnet radiation because of the collimation of the radiation in both horizontal and vertical planes (Fig. 1). Undulators thus produce the brightest radiation possible from an electron beam. Furthermore, the spectral distribution of radiation from an undulator is qualitatively different from that of a bending magnet or wiggler. Nearly all undulators used in third-generation synchrotron light sources are fabricated from permanent magnets; however, superconducting devices are becoming popular when higher photon energies are desired.

Developments in undulator technology have considerably improved their performance. An example is an undulator in which the magnetic material is inside (rather than outside) the vacuum chamber, allowing considerably higher magnetic fields at the electron beam than were previously obtainable. Other new types of undulators can produce variable polarization. For example, devices with four arrays of magnets (two above and two below the beam) can produce circular, helical, or linear oscillations of the

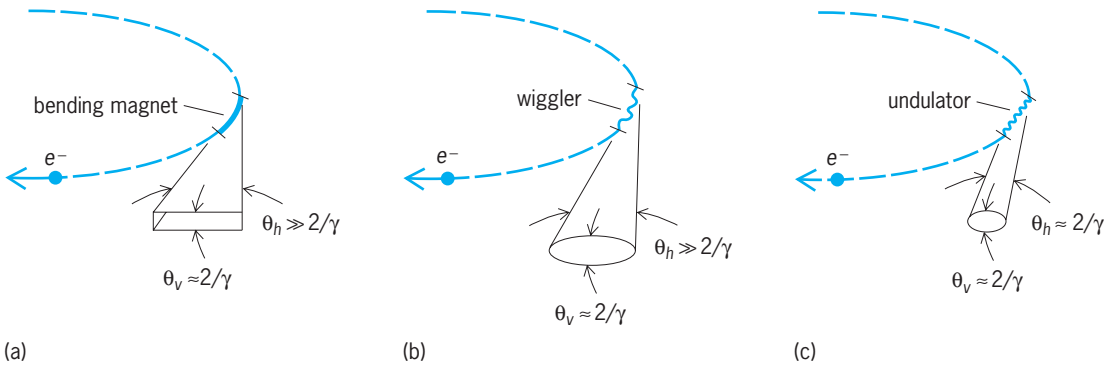


Fig. 4. Spatial characteristics of photon beams from (*a*) bending magnets, (*b*) wigglers, and (*c*) undulators.

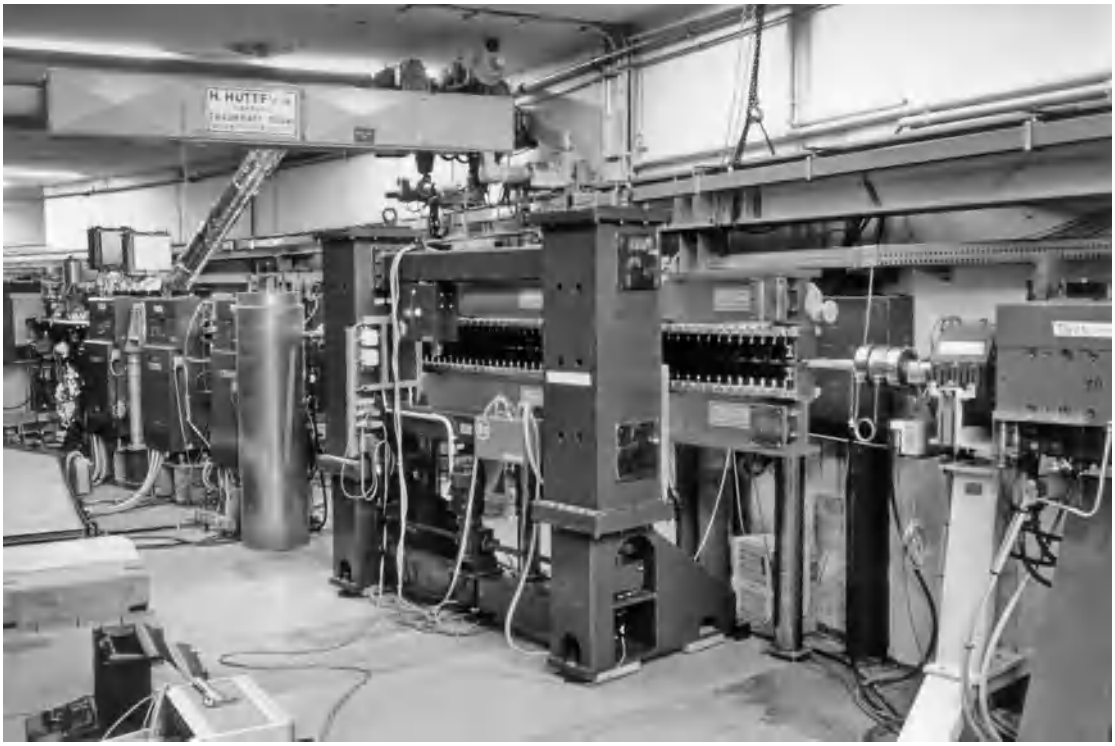


Fig. 5. Open wiggler based on arrays of permanent magnets in the storage ring DORIS at the Hamburger Synchrotronstrahlungslabor (HASYLAB) at DESY (Deutches Elektronen-Synchroton). The spectrum of emitted synchrotron radiation can be varied by varying the distance between the two magnet arrays. The massive structure is required to maintain precision alignment between the jaws. (*DESY*)

electrons, depending on the relative phases of the fields; the radiation produced by this type of undulator has controllable polarization, including linear polarization in any plane and elliptical or circular polarization.

Spectrum. The main spectral features of the radiation produced by bending magnets, wigglers, and undulators are compared in **Figs. 6**, **7**, and **8**. Figures 6 and 7 show the spectrum (flux and brightness) of bending magnets of the Advanced Light Source at the Lawrence Berkeley Laboratory and of a wiggler operating in that ring. Notably higher flux is provided by the wiggler. The flux spectra are characterized by a single parameter—the critical energy ε_c, given in kiloelectronvolts by $2.2E^3/R$ or $0.665BE^2$, where E is the electron energy in gigaelectronvolts, R is the radius of curvature in meters, and B is the magnetic flux density in teslas. Half the power is radiated above the critical energy and half below. Typically, useful flux is available at energies up to four or five times the critical energy. The smooth, continuous nature of the spectra of bending magnets and wigglers over a broad range of photon energy makes it possible to select the particular energy of interest (for example, the absorption edge energy of a particular element) or to study a process such as x-ray absorption or photoemission as a function of photon energy. In use, the broadband radiation is monochromatized by a grating or crystal so that the bandwidth ($\Delta\varepsilon/\varepsilon$, where ε is the energy of the radiation and $\Delta\varepsilon$ is the energy width of the radiation) is 10^{-3} or smaller. Thus, highly monochromatic radiation at any energy within the range of the synchrotron radiation source is readily available.

The spectrum of radiation from an undulator consists of one or more peaks in energy. Interference effects from the coherent superposition of radiation from each deflection of the electron beam produce a spectrum with peak frequencies that are integral multiples of a fundamental frequency. The spectrum is characterized by the undulator strength parameter, K.

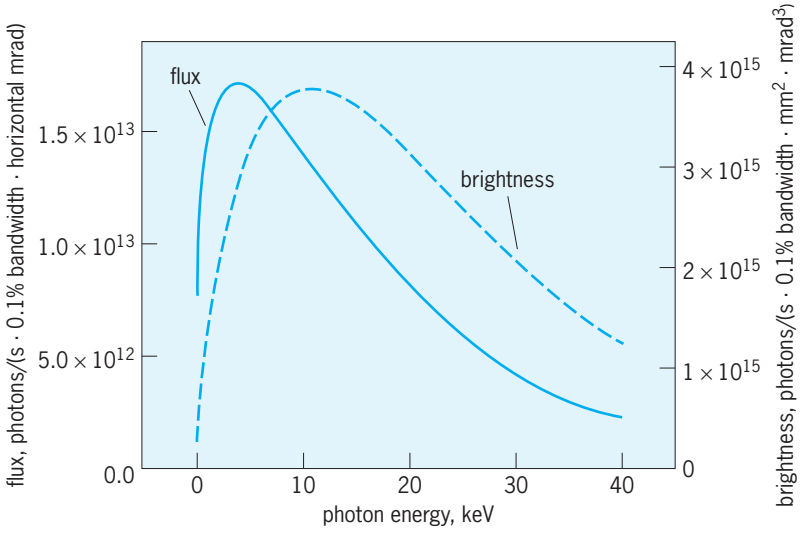


Fig. 6. Photon flux and brightness as functions of photon energy from superconducting bending magnets at the Advanced Light Source.

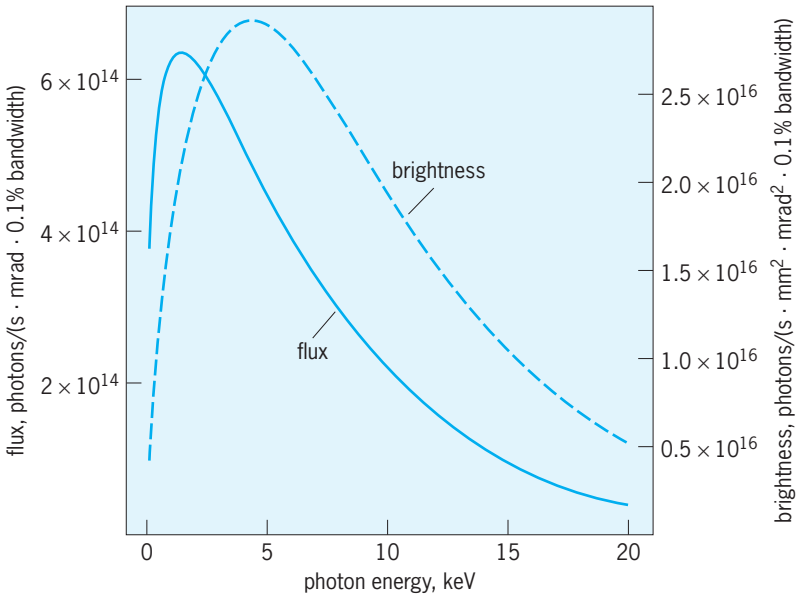


Fig. 7. Photon flux and brightness as functions of photon energy from a wiggler at the Advanced Light Source.

The spectrum from an undulator depends on the value of K. For values of K much less than 1, only the fundamental peak is important. For K approximately equal to 1, the power in the fundamental is a maximum and the first few harmonics have appreciable intensity. For K greater than 1, the wavelength of the fundamental becomes longer and more harmonics appear. For K much greater than 1, the fundamental has very long wavelength and there are many closely spaced harmonics. In this limit, the device is a wiggler, with the envelope of the spectrum approaching the continuous spectrum of that device. Figure 8 shows the spectral brightness of the radiation from an undulator in the Advanced Light Source (ALS), and shows the extremely high brightness offered by the undulator at discrete photon energies that can be varied over a range of values. The spectrum is the locus of the discrete spectral peaks as the undulator magnetic field is varied by changing the gap distance between the magnet arrays.

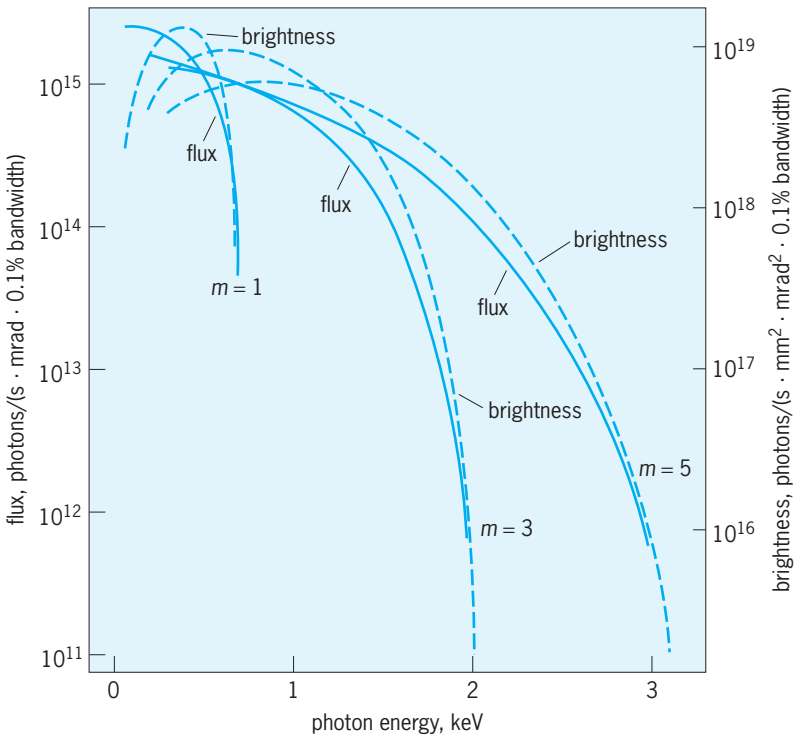


Fig. 8. Photon flux and brightness for an undulator at the Advanced Light Source. The letter *m* indicates the harmonic produced by the undulator. (Only the first few odd harmonics are useful.)

Polarization. The radiation from bending magnets and from wigglers and undulators with vertical magnetic fields is nearly 100% linearly polarized in the plane of the electron orbit, with the electric vector parallel to the electron acceleration. Out of the plane of the orbit, bending magnets produce elliptically polarized radiation. Undulators that produce circular, helical, or linear electron oscillations produce radiation whose polarization is controllable, allowing linear polarization in any plane and elliptical or circular polarization. Circular and helical polarization are used in many kinds of studies, especially of magnetic materials. *See* POLARIZATION OF WAVES.

Pulsed time structure. Radiation from a synchrotron storage ring is produced in pulses as each electron bunch sweeps by the observation point, due to the bunching of the electron beam by the radio-frequency accelerating system of the ring. The pulse duration can be as short as about 50 picoseconds and, in large rings, the interval between pulses can be 1 microsecond or more. This pulse structure facilitates study of time-dependent phenomena and is also used to enhance the data collection rate in some experiments (for example, by measuring photoelectron velocities using time-of-flight techniques).

Although femtosecond x-ray pulses will be most thoroughly exploited at free-electron lasers, it is now possible to produce 100-fs x-ray pulses (that is, pulses shorter than the electron bunch length), in lower fluxes in conventional storage rings. Femtosecond lasers can interact with electrons in a storage ring to produce femtosecond pulses of x-rays. Some ideas for manipulating the electron beam also may make it possible to produce pulses of around 1 ps duration with high flux, but such pulses have not been demonstrated so far. Ultrashort pulses are useful in many interesting studies, for example, investigations of the dynamics of phase changes in matter and the time evolution of electronic and magnetic processes.

Brightness. Brightness is a different property than photon flux, or intensity, which is measured in photons per second. Brightness is directed intensity and includes the source size and the solid angle of the radiation. In this regard, the comparison of a lightbulb with a laser pointer is instructive: The lightbulb radiates photons in most directions, while a laser pointer directs photons into a small spot and a narrow cone. Furthermore, a lightbulb produces a broad spectrum of radiation, while a laser pointer produces radiation in a very narrow range of wavelengths (for example, the red beam from a helium-neon laser). A conventional x-ray tube is analogous to a lightbulb, while

radiation from an undulator is analogous to the beam from a laser pointer.

Brightness is a conserved quantity and cannot be increased by an optical system such as a lens. Brightness is thus a property related to the radiation source, such as the low-emittance beam of electrons in a third-generation storage ring that is coupled with an undulator to produce the radiation.

A high-brightness beam can be focused to a small spot. This property is very important in spectroscopy with high spectral resolution, as high flux can be passed through a narrow slit in a monochromator. High brightness is also important in x-ray microscopy, such as in a scanning x-ray microscope, where spatial resolution is limited by the size of the photon beam. Finally, as noted above, a high-brightness beam can be partially coherent, allowing such applications as interferometry and speckle spectroscopy.

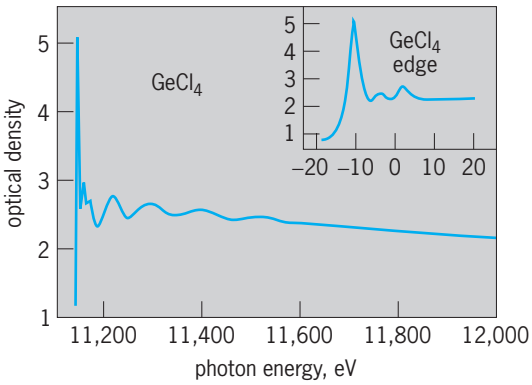


Fig. 9. X-ray absorption spectrum above the germanium K-edge in $GeCl_4$ vapor, showing oscillatory structure known as EXAFS. The inset shows the edge region with a shifted energy scale, displaying XANES spectrum. (*After B. M. Kincaid and P. Eisenberger, Synchrotron radiation studies of the K-edge photo-absorption spectra of Kr, Br_2, and $GeCl_4$: A comparison of theory and experiment, Phys. Rev. Lett., 34:1361–1364, 1975*)

Research Applications

Many of the conventional techniques and processes employing ultraviolet and x-radiation in research and technology have experienced considerable improvement as a result of one or more of the special properties of synchrotron radiation described above. For example, the wavelengths corresponding to the photon energies typically exploited in synchrotron radiation span length scales ranging from the atomic level to biological cells, a range that well suits advanced research in materials science, physical and chemical sciences, metrology, geosciences, environmental sciences, biosciences, medical sciences, and pharmaceutical sciences. The ability to probe and image small areas makes synchrotron radiation particularly suitable for the new field of nanoscience. The availability of synchrotron radiation has already made possible major scientific and technological advances.

The fundamental parameters that characterize the physical world (energy, momentum, position, and time) lend themselves to three broad categories of synchrotron experimental measurement techniques: spectroscopy, scattering, and imaging. Each of these techniques can be performed in a time-resolved mode by exploiting the short pulse lengths of synchrotron radiation.

Spectroscopy. Spectroscopy techniques are used to study the energies of particles that are emitted or absorbed by samples that are exposed to the light-source beam and are commonly used to determine the characteristics of chemical bonding and electron motion.

X-ray absorption spectroscopy (XANES and EXAFS). A plot of x-ray absorption versus photon energy (**Fig. 9**) shows a steep rise in the absorption at the absorption-edge energy, where the incident photon is just able to excite core electrons into empty states. X-ray absorption near-edge structure (XANES, also known as NEXAFS for near-edge x-ray absorption fine structure) is due to transitions into localized atomic levels which are modified strongly by the absorbing atom's surroundings. At higher energies, the observed oscillatory structure of the absorption, known as the extended x-ray absorption fine structure (EXAFS), is due to interference between the outgoing photoelectron waves and the electron waves backscattered from atoms adjacent to the absorbing atoms. This interference modulates the probability of exciting the electron. Because (1) XANES and EXAFS are significant modifications of the atomic absorption and (2) the absorption-edge energies of different atomic species are usually quite different, their analysis provides considerable information about the average atomic environment of each atomic species in complex, polyatomic solids, liquids, and gases.

The highly intense, continuous synchrotron radiation spectrum from multi-gigaelectronvolt storage rings is ideally suited for absorption measurements as a function of photon energy. Consequently, it has been used extensively for structural studies of amorphous materials, heterogeneous catalysts, and noncrystalline metalloproteins whose atomic arrangements are not ordinarily determinable by other structural techniques. These structural studies have led, in turn, to increased understanding of the properties of these materials. EXAFS patterns have also been measured in fractions of a second, so that time-resolved studies of structural changes may be anticipated.

Similar modulations of the x-ray fluorescence yield, the Auger electron yield, and the desorbed atom yield versus photon energy are caused by the same interference phenomenon. The fluorescence yield is used extensively to study the atomic environments of extremely dilute (less than 100 parts per million) constituents of complex systems such as alloys, semiconductors, toxic waste, and proteins in solution. The Auger electron yield is used for the determination of the atomic environments of atoms at the surfaces of materials. The desorbed atom yield is used to determine the surroundings of atoms adjacent to specific atoms on a surface. For example, it is used to determine the environments of those metal atoms adjacent to oxygen atoms on a partially

oxidized metal surface. The high-brightness beams achieved with insertion devices in second- and third-generation sources have led to x-ray absorption spectroscopy microscopy in which micrometer-size areas are examined. *See* AUGER EFFECT; DESORPTION; SPECTROSCOPY; SURFACE PHYSICS; X-RAY FLUORESCENCE ANALYSIS.

Photoemission spectroscopy. Energy and angular distributions of electrons ejected from solids by photons of varying incident energy provide basic information about band structure and core levels of both the surface and bulk of the sample. The extremely high intensity and tunability of synchrotron radiation in the ultraviolet and soft x-ray regions of the spectrum have made possible high-resolution studies which delineate the band structures of crystalline solids, as well as electronic states in highly correlated solids, such as high-temperature superconductors, in which band theory is inadequate. In addition, electrons of energy between 30 and 100 eV have such a large probability for interaction with other atoms that they have an escape depth of only about 0.5 nanometer in most solids. As a result, the detected electrons reveal the properties of surface electronic energy levels of clean materials and the extrinsic surface states of chemisorbed material, opening a route to the study of the phenomena of oxidation, corrosion, and catalysis. Such studies have led to major revisions of models of the metal–semiconductor interface which determines many properties of solid-state electron devices. Again, the tunability, when coupled with the high intensity of synchrotron radiation, makes possible the separate delineation of surface and bulk electronic states. High-brightness beams have led to photoelectron microscopy whereby electronic states in micrometer-size areas are examined. *See* BAND THEORY OF SOLIDS; SEMICONDUCTOR; SUPERCONDUCTIVITY; SURFACE PHYSICS.

In addition, synchrotron radiation facilitates photoemission studies in two ways: the high-vacuum environment of the storage ring minimizes contamination of clean surfaces prepared in place; and the sharply pulsed time structure of the radiation makes it possible to measure photoelectron energies by time of flight. *See* ELECTRON SPECTROSCOPY; PHOTOEMISSION.

X-ray fluorescence trace-element analysis. The high intensity, tunability, high collimation, and high linear polarization of synchrotron radiation make it valuable for trace-element analysis through x-ray fluorescence. Such trace-element analysis is used to determine very low concentrations of impurities introduced in semiconductor processing procedures, with the goal of further decreasing those concentrations (**Fig. 10**).

Scattering. Scattering or diffraction techniques make use of the patterns of light produced when x-rays are deflected by the closely spaced lattice of atoms in solids and are commonly used to determine the structures of crystals and large molecules such as proteins.

X-ray diffraction and scattering. Synchrotron radiation from multi-gigaelectronvolt storage rings is characterized by its high intensity, natural collimation, tunability, and unique polarization and time structure. These properties have contributed to its profound impact in the field of x-ray diffraction and scattering. The high intensity and natural collimation of the

Fig. 10. Worker checking silicon wafers for minute impurities at the Stanford Synchrotron Radiation Laboratory (SSRL). (*Photo by Peter Ginter*)

x-rays have made measurements of the atomic arrangements on surfaces and at interfaces of liquids and solids a routine technique. Detailed studies have been made of the structural changes that occur during two-dimensional phase transitions (melting and freezing) as well as changes in these processes as the system under study goes from a two-dimensional system to a three-dimensional system through the addition of individual atomic layers. *See* PHASE TRANSITIONS.

Anomalous (or resonant) scattering has been utilized in wide-angle diffraction experiments and small-angle x-ray scattering. Close to an atom's absorption edge, the atom's scattering cross section (atomic scattering factor) can vary rapidly as a function of incident x-ray energy. This anomalous scattering is used to alter the contribution of that atomic species to the overall scattering amplitude and thereby obtain additional information about the overall structure of the sample. The tunability of the highly monochromatized synchrotron radiation makes such experiments possible. *See* SCATTERING OF ELECTROMAGNETIC RADIATION.

The use of polarized x-rays makes possible studies of the magnetic properties of materials. The interaction of x-rays and the spin and orbital angular momentum of the electron is very weak, but in some cases it can be enhanced by many orders of magnitude by tuning the incident x-ray energy to an absorption edge of one of the constituent elements in the sample. This resonant scattering technique not only allows the bulk magnetic properties of materials to be studied by x-rays but also permits surface and interface magnetic properties to be explored. *See* ELECTRON SPIN; MAGNETISM.

The high brightness and flux of synchrotron radiation is routinely exploited to perform time-resolved scattering and diffraction experiments. Measurements on a millisecond time scale have been made to monitor the structure of muscles during contraction in an effort to understand the associated molecular rearrangements that occur during the contraction process. Even shorter time scales can be attained by using the pulsed nature of the source. Nanosecond time-resolved studies have been made of melting and recrystallization of semiconductors irradiated with short-pulse lasers, and the structural dynamics of macromolecules undergoing laser photolysis have been studied using synchrotron radiation. *See* LASER PHOTOCHEMISTRY.

All the experiments described so far deal with elastic scattering, in which the scattered x-ray has the same energy as the incident photon. Elastic scattering is not the only process that can occur in a sample. The incident x-ray can give up (or gain) some energy in the scattering process by exchanging energy with excitations in the sample. Energies associated with electronic excitations in samples are typically several hundred millielectronvolts, while vibrational excitations (for example, phonons in crystals) have energies of many millielectronvolts. One of the challenges of performing inelastic x-ray scattering experiments is to prepare monochromators and analyzers (to separate the inelastically scattered x-rays from elastically scattered ones) with energy resolutions, $\Delta E/E$, of a few times 10^{-7} (several millielectronvolts per 10^4 eV). The combination of very high collimation and high spectral brightness (number of photons per second per unit energy width, ΔE) of undulators on third-generation sources has allowed such experiments to be performed. Phonon dispersion curves of several materials have been mapped out. Using resonant techniques, the vibrations of specific atomic species can be isolated in complex systems and their partial densities of states can be measured. The partial densities of states related to the iron atoms in myoglobin and hemoglobin have been determined in this manner. *See* LATTICE VIBRATIONS; PHONON; X-RAY DIFFRACTION.

Structural molecular biology. Structural molecular biology is growing at an extraordinary pace, due in no small part to the impact of synchrotron radiation research. In addition to basic research, a significant contemporary use of macromolecular crystallography is so-called structure-based drug design, which enables pharmaceutical companies to develop potential drug candidates with the knowledge of where and how they bind to their targets in cellular structures (**Fig. 11**). Structural analysis of macromolecules (specifically proteins and nucleic acids) can, in many cases, be performed with considerably better resolution using synchrotron radiation as compared to laboratory x-ray sources. In particular, there has been progress in overcoming two major limitations of conventional x-ray diffraction research, associated with radiation damage and small crystal size. Frequently the limitation on the number of Bragg peaks whose intensity can be accurately measured is the x-ray damage of the crystal itself. However, it has been observed that a given dose of x-rays does

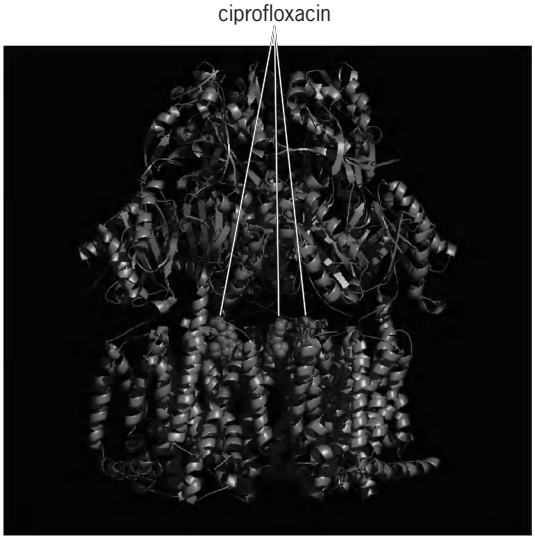


Fig. 11. Ciprofloxacin, an antibiotic used to treat a variety of bacterial infections including inhaled anthrax, bound in the large central cavity of *Escherichia coli* AcrB, a bacterial protein complex that repels a wide range of antibiotics. (*From Lawrence Berkeley Laboratory, Research News: Secrets of Drug Resistance Revealed, May 9, 2003*, http://www.lbl.gov/Science-Articles/Archive/PBD-drug-resistance.html)

less damage if applied over a short time; that is, more intense x-ray beams allow the data to be collected before the crystal is destroyed. As a result, synchrotron radiation makes possible the determination of many more measurable peaks before the crystal is damaged by the probing radiation. Crystals held at cryogenic temperatures have been found to be even more resistant to radiation. The combined effects of brief x-ray exposure and cryogenic temperatures allow many more macromolecular crystals to be candidates for crystal structural analysis. The quality of data from those crystals has improved significantly, with advanced analysis techniques producing resolution better than 0.1 nm. Structure of protein crystals whose unit cells have sides of over 100 nm are now being solved.

Anomalous scattering may be used to help solve the so-called phase problem in crystallography. Using the tunability of the synchrotron radiation, data sets can be collected at several wavelengths around the absorption edge of a strongly absorbing atom (an atom with a relatively high atomic number) contained within the macromolecule. This technique of multiwavelength anomalous dispersion (MAD) eliminates the necessity of growing crystals with three different isomorphic chemical forms [the multiple isomorphous replacement (MIR) technique], an important advantage since it is frequently impossible to grow crystals in three different forms. This technique was more fully realized when it was found possible to incorporate selenium in place of sulfur (selenomethionine in place of methionine) in samples grown in certain strains of *Escherichia coli*.

It is often difficult, if not impossible, to grow the large single crystals of macromolecules that were needed for structural analysis with laboratory x-ray sources. The high intensity of synchrotron radiation allows structures to be determined from crystals only tens of micrometers on a side.

Crystallographers are no longer confined to static observations of protein structures. The availability of extremely intense, multiwavelength, pulsed synchrotron x-ray sources has reduced exposure times enough to capture "movie sequences" of fundamental molecular processes. This approach has been applied to studies of myoglobin, the iron-based molecule responsible for oxygen transport in muscles. The entire photolysis, relaxation, and rebinding processes occur in less than 5 milliseconds at room temperature. Researchers are now extending this approach to several other light-sensitive signaling systems that are chemically and biologically diverse, and are developing new techniques that will enhance the time resolution from the nanosecond range, first to a few hundred picoseconds and perhaps ultimately to femtoseconds. *See* MOLECULAR BIOLOGY; X-RAY CRYSTALLOGRAPHY.

Lensless imaging. In a big step beyond the periodic objects studied by crystallography, it is possible to reconstruct images of nonperiodic samples using x-ray diffraction imaging (XDI) or "lensless imaging." Like crystallography, x-ray diffraction imaging is based on the analysis of diffraction patterns, but it uses iterative algorithms to extract the phase information needed to reconstruct the object and requires that the diffraction intensity be zero outside the object's boundary. The better this boundary is known, the faster the iterations converge to an accurate image. The boundary can be supplied by other imaging techniques or by more sophisticated iterative computation methods. Two-dimensional images of objects as complex as yeast cells with resolutions of a few tens of nanometers have been achieved, as have three-dimensional images of simpler, solid test objects, such as collections of gold balls. It is anticipated that a three-dimensional resolution of 10 nm will ultimately be possible for life science samples, where radiation damage is an issue, and 2 nm for solids.

Imaging. Imaging techniques use the light-source beam to obtain pictures with fine spatial resolution of the samples under study, and are used in diverse research areas such as cell biology, lithography, infrared microscopy, radiology, and x-ray tomography. There are two basic experimental methods, scanning and full-field imaging. In either case, the contrast in the images must have some physical cause. This contrast can be due to changes in absorption (amplitude contrast) or refractive index of the sample (phase contrast) caused by, for example, variable elemental or chemical composition or magnetic properties. The tunability of synchrotron radiation is absolutely essential for the creation of contrast mechanisms. Tomography can be combined with x-ray microscopy for full three-dimensional images, a technique that is especially useful in the life sciences in imaging cellular structures. Newer "lensless" imaging schemes based on reconstruction of scattering patterns made with coherent x-ray beams are now coming into use. Related to imaging is spatially resolved spectroscopy and diffraction, which allow the construction of images based on structural features other than density, such as an absorption feature, local lattice strain, and many others. Multidimensional data sets that contain, in the case of spectroscopy, a complete spectrum for each point (to the resolution of the technique being used) in the image can then be analyzed in multiple ways.

Owing to their somewhat different properties, soft and hard x-rays differ in their imaging applications. The wavelengths of soft x-ray photons (1–15 nm) are very well matched to the creation of nanoscopes capable of probing the interior structure of biological cells and inorganic mesoscopic systems. Problems addressed by soft x-ray imaging techniques include cell biology, nanomagnetism, environmental science, and soft matter and polymers. Hard x-ray imaging nondestructively visualizes samples, and frequently their internal or hidden components. It is applicable to nearly all fields of science from the life sciences to engineering and archeology. Electron microscopes will always have better spatial resolution, but they will be limited in the range of samples that they can study. The uniqueness of x-ray imaging resides in its deeper penetration, enabling the study, for example, of buried interfaces and wet biological samples.

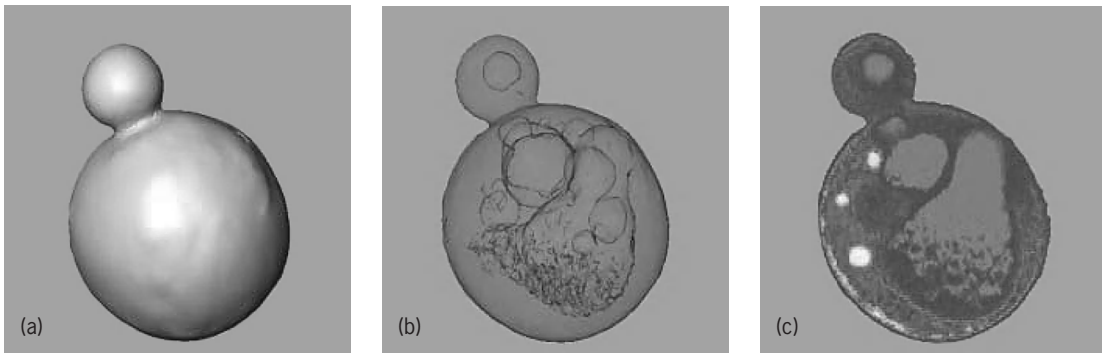


Fig. 12. Reconstructed data of a budding yeast, using different volume-analysis algorithms. (*a*) Opaque surface. (*b*) Transparent surface showing internal vesicles. (*c*) Volume-rendered thick-slice section. (*From C. A. Larabell and M. A. Le Gros, X-ray tomography generates 3-D reconstructions of the yeast, Saccharomyces cerevisiae, at 60-nm resolution, Mol. Biol. Cell, 15:957–962, 2004*)

Ampitude contrast. Literally from the moment they were discovered in 1895, x-rays have been used to image visibly opaque objects. This type of imaging, familiar from medical x-rays, is based on the contrast that arises from the attenuation of the amplitude of the incident x-ray wave by the different constituents of the sample. The focusing required for imaging in the x-ray region is difficult, but several advances in x-ray optics, coupled with the very high brightness of third-generation sources, have led to important advances in this area, and there are now several approaches in which x-ray beams with dimensions of 15 nm and higher can be produced for both scanning and full-field imaging techniques.

In scanning imaging, a very small illuminated spot is created on the sample using focusing devices. Fresnel zone plates and specially curved mirrors perform the same function for x-rays that lenses do for visible light. The image is then built up by "raster scanning" the sample through the illuminated spot. The species detected can be photons (transmitted or fluorescent) or photoelectrons. With the use of two zone plates (a condenser and an objective), it is possible to create a full-field image. In photoemission electron microscopy, a smallish spot on the sample is illuminated and the emitted photoelectrons are passed through an electron microscope column to produce a magnified full-field image. *See* DIFFRACTION; ELECTRON; MICROSCOPE.

Imaging methods can then be employed as probes, using any of the techniques previously discussed to carry out spatially resolved spectroscopy, microdiffraction, and microfluorescence analysis. Two-dimensional distributions of strains, chemical elements, and valence states of a particular element can be mapped out. These techniques allow the determination of the strains in individual grains in a metal, the elemental distributions of metals in cells treated with anticancer drugs, and the oxidation states in heavy metals that accumulate in the root systems of plants being considered for bioremedial applications. X-ray microscopes are fast becoming one of the most sought-after instruments at synchrotron radiation facilities. *See* X-RAY MICROSCOPE.

X-ray tomography. Imaging at multiple angles can also be used to reconstruct three-dimensional pictures of objects with submicrometer resolution using the same techniques as are used on larger scales for medical purposes in computerized tomography (CT) scans. In the life sciences, x-ray tomography is the first high-throughput imaging technology that generates images of whole, hydrated cells at better than 60-nm resolution (**Fig. 12**), thereby bridging the middle area between light (200 nm) and electron microscopy (0.3 nm). High-resolution images provide contrast between cellular structures and allow for discernment of individual structures. After data collection, tomographic techniques are used to reconstruct the original information into quantifiable three-dimensional views of the entire cell. Through the use of computer algorithms, the researchers then process the reconstructed data to create made-to-order images of whole cells and their internal structures. In the physical sciences, three-dimensional maps of the concentrations of individual atomic species in complex, polyatomic materials can be constructed. Applications range from the study of petroleum distributions in soils to the analysis of osteoporosis development and treatment. *See* COMPUTERIZED TOMOGRAPHY.

Phase contrast. When extremely bright (that is, partially coherent) x-ray beams are employed for the purposes of imaging, contrast mechanisms other than absorption may be used, in particular phase-contrast imaging. Phase-contrast imaging does not rely on attenuation of the beam for image contrast, but on the variation of the refractive index of the constituent components. The variation of refractive index results in a distortion of the phase fronts of a (partially) coherent incident beam as it traverses the sample. Since the x-rays are coherent, the direct and scattered waves can interfere with each other, resulting in a modulation of the amplitude, and hence intensity, of the x-ray beam. Phase-contrast imaging is particularly useful for imaging thin biological samples where the sample is composed primarily of low-atomic-number atoms, resulting in weak absorption contrast. **Figure 13** shows an image of an ant's head taken with phase-contrast imaging techniques.

Phase-contrast imaging is related to another type of imaging phenomenon in which the interference pattern from the combined incident and scattered

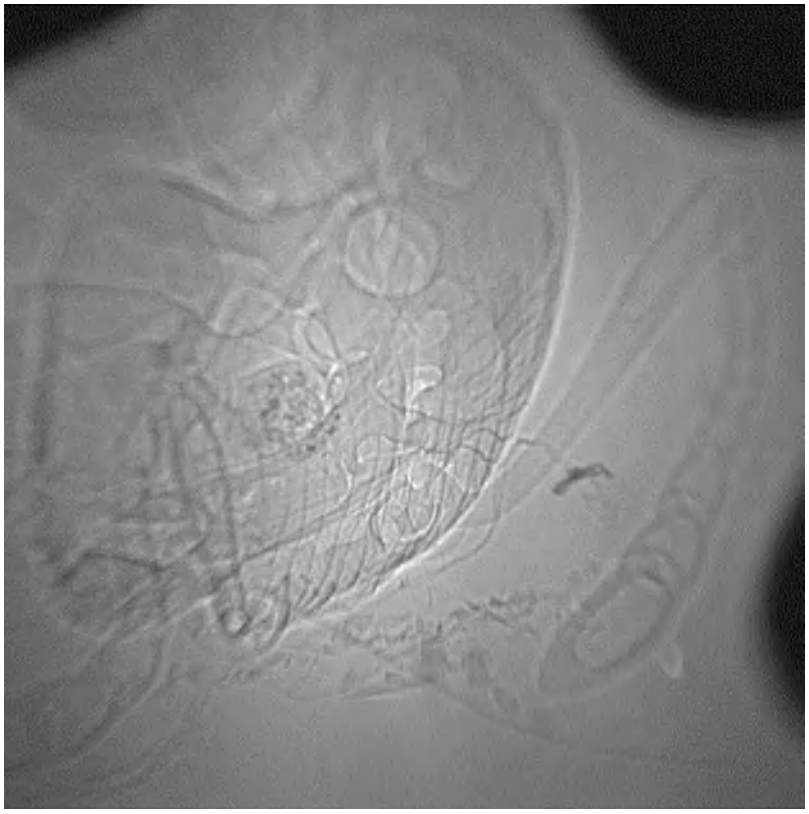


Fig. 13. Phase-contrast image of an ant's head, acquired at one of the undulator beam lines at the Advanced Photon Source, Argonne National Laboratory.

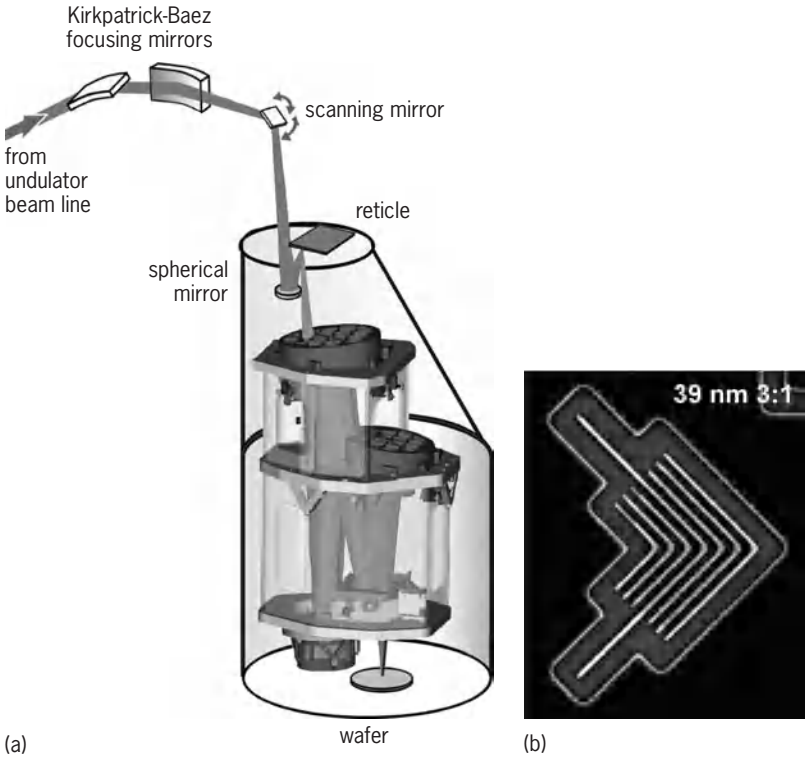


Fig. 14. Extreme ultraviolet (EUV) lithography. (*a*) Setup of equipment. The scanning mirror directs focused EUV light from the Kirkpatrick-Baez mirrors to the mask-carrying reticle with the desired degree of spatial coherence and illumination pattern. The optic images the EUV light reflected from the reticle onto a resist-covered silicon wafer. (*b*) Printed elbow test pattern with 39-nm-wide lines and 3:1 line-to-spacing ratio. (*From P.P. Naulleau et al., Static microfield printing at the Advanced Light Source with the ETS Set-2 optic, Proc. SPIE, 4688:64–71, 2002*)

waves is recorded, namely holography. Capturing a three-dimensional image with spatial resolution of a fraction of a nanometer is the ultimate goal of the x-ray imaging scientist. Third-generation sources are on the verge of achieving this capability, and the next generation of sources may capture three-dimensional images of cells and other extremely small objects with a single 100-Fs flash of x-rays from a free-electron laser. *See* HOLOGRAPHY; PHASE-CONTRAST MICROSCOPE.

Imaging of humans. Coronary arteries are difficult to image with normal radiography techniques because the heart and other surrounding tissue absorb much more radiation than do the arteries. In order to achieve contrast, it is necessary to introduce into the arteries a dye containing elements such as iodine that heavily absorb x-rays. In synchrotron radiation angiography, a much lower iodine concentration is needed, making it possible to use less hazardous catheterization procedures to introduce the iodine. The contrast is achieved by taking images in rapid succession at two different x-ray wavelengths, one just above and one just below the iodine absorption-edge wavelength. Subtraction of one image from the other yields an image of iodine-containing arteries. *See* MEDICAL IMAGING; RADIOGRAPHY.

In conventional mammograms, differences in tissue densities and composition show up as contrasting areas due to x-ray absorption, allowing doctors to see tumors or changes in tissue. However, differences between healthy and cancerous tissues are very small, and scattering of x-rays can lead to blurring and even lower contrast. A mammography technique called diffraction-enhanced imaging (DEI) uses ultrabright x-rays and provides a dramatic contrast between normal tissues and tumors. The DEI method uses a single-energy beam of x-rays instead of the broad-energy beam used in conventional imaging, and an analyzer crystal that can differentiate between x-rays that are traveling much less than one ten-thousandth of a degree apart, thereby reducing scatter, and helps visualize low-contrast areas that otherwise would be lost. In addition to mammography, potential applications of DEI include other low-contrast tissues and organs such as kidneys, and the nondestructive testing of materials.

X-ray lithography. A technique used in the art world for many centuries, lithography has been adopted, and adapted with phenomenal success, by high-technology industries. In microchip manufacturing, a silicon wafer is coated with a thin layer of photosensitive material called a resist. An image of a mask containing the desired pattern is projected onto the resist. The exposed (or unexposed) parts of the resist are etched away and, with further processing, the desired circuit is built up. The same basic process can be used in the manufacture of small mechanical components known as microelectromechanical structures (MEMS). In MEMS, features are not always as fine as in microchip circuits, but they must be sufficiently robust to withstand stresses due to fluid flow (in miniature chemical processing systems) or to mechanical motion (in gears, motors, and the like).

Work at synchrotron light sources focuses primarily on the exposures of the resists for MEMS, for optical devices such as zone plates, and for testing components and processes for microchip fabrication. The features that make synchrotron radiation attractive for this purpose are (1) high intensity, which provides for short exposures and therefore high throughput; (2) high collimation, which makes possible high resolution; and for MEMS (3) the ability to obtain high aspect ratios. In addition, hard x-rays (with energies greater than 12 keV) can penetrate deep into the photosensitive resist, making structures with high aspect ratios (deeper than they are wide). **Figure 14** shows an example of extreme ultraviolet lithography. Undulator light from a synchrotron is used to expose a test pattern onto a silicon wafer coated with resist, resulting in a test pattern with 39-nm lines. *See* INTEGRATED CIRCUITS; MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); MICROLITHOGRAPHY.

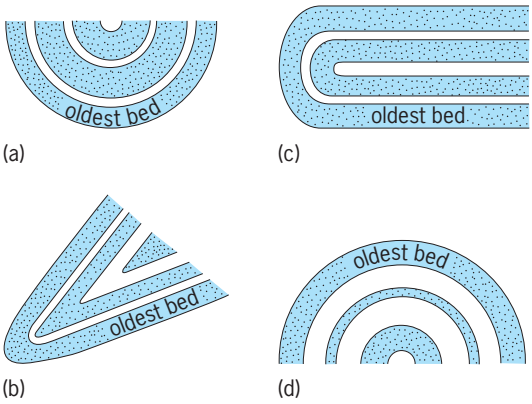
Alfred S. Schlachter; Arthur L Robinson; Arthur Bienenstock; Dennis Mills; Gopal Shenoy; Herman Winick

Bibliography. M. Altarelli, F. Schlachter, and J. Cross, Making ultrabright x-rays, *Sci. Amer.*, 279(6):66–73, December 1998; F. Ciocci (ed.), *Insertion Devices for Synchrotron Radiation and Free Electron Laser*, World Scientific, 2000; P. J. Duke, *Synchrotron Radiation: Production and Properties*, Oxford, 2000; P. Eisenberger, *Materials Science Using Synchrotron Radiation*, Elsevier, 2000; M. R. Howells, J. Kirz, and D. Sayre, X-ray microscopes, *Sci. Amer.*, 264(2):88–94, February 1991; E. E. Koch, T. Sasaki, and H. Winick (eds.), *Handbook on Synchrotron Radiation*, vols. 1–4, Elsevier, 1983–1991; T.-K. Sham (ed.), *Chemical Applications of Synchrotron Radiation*, World Scientific, 2002; H. Winick (ed.), *Synchrotron Radiation Sources: A Primer*, World Scientific, 1995.

Syncline

In its simplest form, a geologic structure marked by the folding of originally horizontal rock layers into a systematically curved, concave upward profile geometry (**illus.** *a*). A syncline is convex in the direction of the oldest beds in the folded sequence, concave in the direction of the youngest beds. Although typically upright, a syncline may be overturned, recumbent, or upside down (illus. *d*). Synclines occur in all sizes, from microscopic to regional. Profile forms may be curved smoothly (illus. *a*) to sharply angular (illus. *b*). Fold tightness of a syncline, as measured by the angle at which the limbs of the syncline join, may be so gentle that the fold is barely discernible, to so tight that the limbs are virtually parallel to one another (illus. *c*). The orientation of the axis of folding is horizontal to shallowly plunging, but synclines may plunge as steeply as vertical.

Synclines are products of the layer-parallel compression that arises commonly during mountain building. Stresses parallel to or at an inclined angle to rock layering can achieve shortening of the layers through buckling and bending. True bending is accommodated often by flexural slip, a mechanism in which the rock layers, like slick pages in a book or magazine, slip past one another along bedding-plane weaknesses. The final profile form of the fold reflects the mechanical properties of the rock sequence under the temperature-pressure conditions of folding, and the percentage of shortening required by the deformation. *See* ANTICLINE; FOLD AND FOLD SYSTEMS.



Varieties of synclines as seen in profile view. (*a*) Upright syncline with smoothly curved limbs. (*b*) Overturned, sharply angular syncline with planar limbs. (*c*) Recumbent, isoclinal syncline with parallel limbs. (*d*) Upside-down syncline, sometimes called an antiformal syncline.

George H. Davis

Bibliography. G. H. Davis, *Structural Geology*, 2d ed., 1996; S. Mitra and G. W. Fisher, *Structural Geology of Fold and Thrust Belts*, 1992; N. J. Price and J. W. Cosgrove, *Analysis of Geological Structures*, 1991.

Synthetic aperture radar (SAR)

Radar, airborne or satellite-borne, that uses special signal processing to produce high-resolution images of the surface of the Earth (or another object) while traversing a considerable flight path. The technique is somewhat like using an antenna as wide as the flight path traversed, that being the large "synthetic aperture," which would form a very narrow beam. Synthetic aperture radar is extremely valuable in both military and civil remote-sensing applications, providing surface mapping regardless of darkness or weather conditions that hamper other methods.

Basic theory. Resolution is the quality of separating multiple objects clearly. In radar imaging, fine resolution is desired in both the down-range and cross-range dimensions. In radar using pulses, down-range resolution is achieved by using broad-bandwidth pulses, the equivalent of very narrow pulses, allowing the radar to sense separate echoes from objects very closely spaced in range. This technique is called pulse compression; resolution of a few nanoseconds (for example, 5 ns = 5×10^{-9} s gives about 0.75 m or 2.5 ft resolution) is readily achieved in modern radar.

Cross-range resolution is much more difficult to

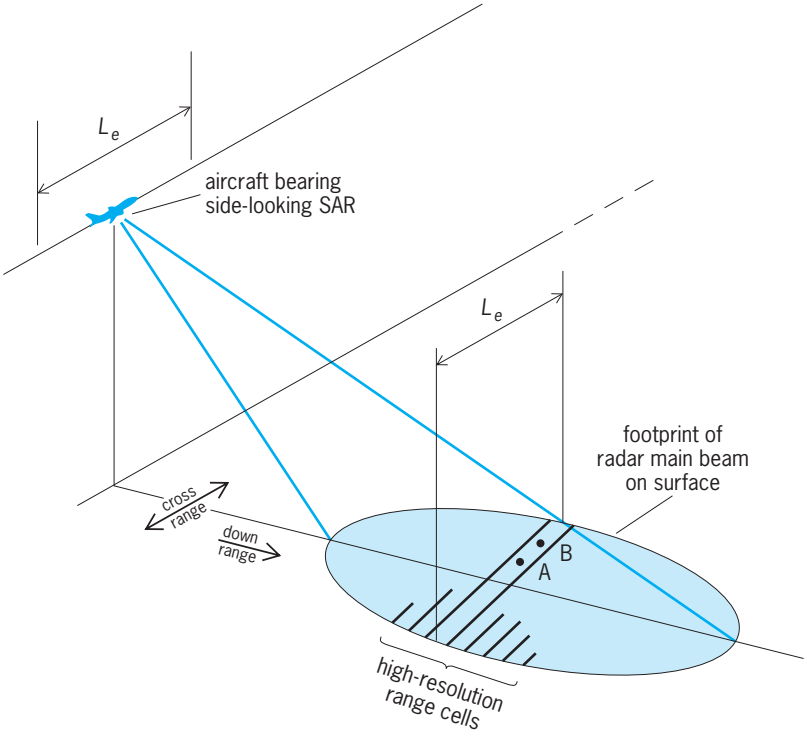


Fig. 1. The basic idea of synthetic aperture radar (SAR); a side-looking case is illustrated. Two example scatterers, A and B, are shown in the ground scene. L_e = maximum flight path length for effective SAR processing.

achieve. Generally, the width of the radar's main beam determines the cross-range, or lateral, resolution. For example, a 3° beam width resolves targets at a range of 185 km (100 nautical miles) only if they are separated laterally by more than 100 m (330 ft), not nearly enough resolution for quality imaging.

However, surface objects produce changing Doppler shifts as an airborne radar flies by. In side-looking radar (**Fig. 1**), even distant objects actually go from decreasing in range very slightly to increasing in range, producing a Doppler-time function (**Fig. 2**). If the radar can sustain high-quality Doppler processing for as long as the "footprint" of the beam illuminates the scene, these Doppler histories will reveal the lateral placement of objects. In fact, if such processing can be so sustained, the cross-range resolution possible is one-half the physical width of the actual antenna being used, a few feet perhaps. Furthermore, this resolution is independent of range, quite unlike angle-based lateral resolution in conventional radar. *See* DOPPLER EFFECT.

Many synthetic aperture radars use other than just a fixed side-looking beam. Spotlighting involves steering the beam to sustain illumination for a longer time or to illuminate a designated scene at some other angle. The principles remain unchanged: fine resolution in both down-range and cross-range dimensions (achieved by pulse compression and Doppler processing, respectively) permits imaging with picture cells (pixels) of remarkably fine resolution (**Fig. 3**). Many synthetic aperture radars today achieve pixels of less than 1 m (3.3 ft) square.

Focusing. Echoes from objects as the beam footprint passes over them have range-dependent rates of change of Doppler shift (Fig. 2), changing from positive to negative rapidly for objects close to the flight path and more slowly for objects at greater range. Conventional Doppler filtering, essential for clutter suppression in ordinary airborne radar, separates one fixed Doppler shift from others. Such processing is inadequate in synthetic aperture radar because of the need to compensate for the changing Doppler shift associated with the range of interest. Using such a reference in the signal processing results in good images being formed in swaths, regions relatively well focused around the reference range. Echoes from other ranges do not produce sharp images, unless processed separately with appropriate references. Synthetic aperture radars that form images in this range-dependent way are said to be focused.

Focusing must also compensate for variations in the radar's flight path that would introduce Doppler shifts not related to the object's position. While measurement of the flight-path deviations could be attempted (as with very sensitive instruments in the aircraft), automatic focusing is used now in most synthetic aperture radars. Autofocusing derives a reference signal (various techniques exist) from the scene itself that compensates for flight-path irregularities.

Waveform limitations. Radars determine Doppler by sensing the progressive phase change from pulse to pulse in a coherent dwell of many pulses. (Coherent means simply that both the amplitude and the phase of the returns are used in the processing.) Resolution in Doppler depends upon the length of time of the coherent processing. In synthetic aperture radar, that length of time is limited by the illumination time of the scene, clearly the width of the beam's footprint at the range of interest divided by the aircraft speed (again, for a fixed-beam side-looking radar). It would appear that if the antenna were arbitrarily small in width, the beam would be very broad indeed and the processing time would

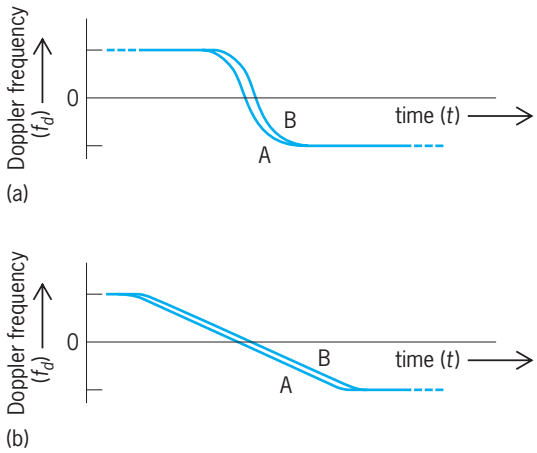


Fig. 2. The range-dependent Doppler-time function of surface objects (reference objects A and B of Fig. 1), as the synthetic aperture radar passes by, for (*a*) A and B close to the flight path and (*b*) A and B at greater range.

be great, giving very fine resolution in cross-range. While those relationships are true, a sampling problem arises related to the beam's footprint. The pulse repetition frequency (prf) must be high enough that no Doppler ambiguity occurs in the range of Doppler shifts being experienced across the beam's footprint. The wider the beam, the higher the prf must be. On the other hand, the beam's down-range footprint suggests that the prf must be low enough that no range ambiguity exists in that dimension. This latter limit on the prf can be eased somewhat by allowing poorly focused overlap (range-ambiguous echoing) from nearer and farther objects, perhaps with acceptable image degradation. *See* COHERENCE.

Continuing research in synthetic aperture radar processing addresses methods to overcome these waveform limitations: Forming multiple beams in cross-range and processing in separate beam-related channels may permit lower prfs; forming adaptive nulls in the down-range antenna pattern to mitigate ambiguous range return may permit higher ones. In any case, the bounding of prf imposed by the beam's footprint precludes use of an arbitrarily small antenna to achieve arbitrarily fine cross-range resolution.

Interferometric SAR. Modern synthetic aperture radar is able to form topological or three-dimensional images using interferometry. A synthetic aperture radar system using two radars, or at least two receiving antennas, spaced vertically by many wavelengths, can determine the relative height of each pixel quite accurately by interpreting the phase difference of the returns at the two antennas. In flat terrain, the down-range pattern of constructive and destructive interference seen at the two antennas would produce regularly spaced isophase contours in the interferogram covering the scene. Changes in terrain elevation distort these contours, and appropriate image processing can produce accurate topological data. The United States' Shuttle Imaging Radar (SIR) used this technique in the mid-1990s with a single radar making two passes over the same scene with different orbit altitudes, and in February 2000 using two antennas separated by 60 m (197 ft) for single-pass operation. *See* INTERFEROMETRY; SPACE SHUTTLE.

Polarimetric SAR. The polarization of an electromagnetic wave describes the spatial orientation of the electric field component. How an object reflects the radar signal in both the transmitted polarization and the orthogonal one is very significant in classifying that object. A polarimetric radar makes such measurements. Many synthetic aperture radars are polarimetric, permitting classification of the terrain (the kind of vegetation, distinctions between natural and artificial objects, and so forth). Adaptive systems can estimate the polarimetric state of uninteresting returns (the clutter) and reduce sensitivity to it, increasing the contrast between such background and pixels containing more interesting objects. Furthermore, if knowledge of the polarimetric character of a target of particular interest is available, sensitivity to its presence may be increased by polarimetric weighting techniques. Many synthetic aperture radars are multifrequency (operating simultaneously in many bands) and polarimetric. *See* POLARIZATION OF WAVES.

Fig. 3. Synthetic aperture radar image of an office building in Lincoln, Massachusetts. (*Lincoln Laboratory, Massachusetts Institute of Technology*)

Inverse SAR. Another technique for imaging targets is closely related to synthetic aperture radar, but is applicable to even fixed land-based radars. As an aircraft, for example, approaches a radar equipped with synthetic-aperture-radar-like Doppler processing in highly resolved range cells, the slight rotation of the target itself relative to the radar provides a meaningful Doppler spectrum in each range cell, indicative of the lateral position of the scatterers in that cell. Autofocusing is used to compensate for the translation of the target and for flight instabilities, leaving just the Doppler shift due to the target's orderly change of aspect. As in synthetic aperture radar, a few seconds of coherent processing is typically sufficient to achieve the Doppler resolution necessary for the desired cross-range resolution. At microwave wavelengths, only a few degrees of aspect change are necessary to produce discernible Dopplers. Since the radar is stationary and the target moving, this technique is called inverse synthetic aperture radar (ISAR). **Figure 4** shows images of an aircraft obtained in experiments with a C-band ISAR in China, where different methods of autofocusing were being explored. Ships, with their somewhat regular local rotations (rolling, pitching), have been imaged successfully with ISAR.

Other considerations. A limitation in image quality comes from the convenient but false assumption that

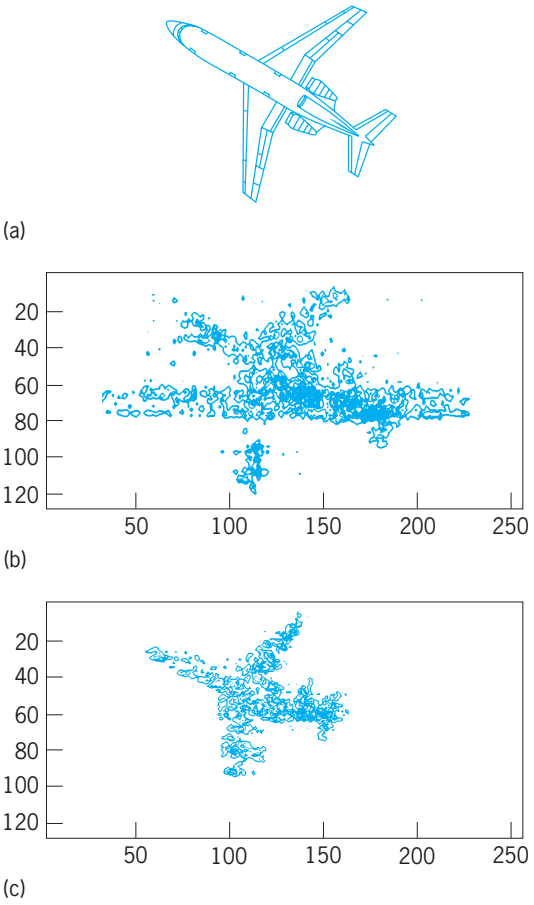


Fig. 4. Example of inverse synthetic aperture radar (ISAR). (*a*) Outline of YAK-42 aircraft. (*b*, *c*) ISAR images of the aircraft for two different focusing methods. (*Xidian University of China*)

the scatterers involved are all ideal point scatterers. Research continues toward techniques to accommodate better the return from structural elements actually found in real targets.

Another problem is that moving surface targets generally produce Doppler-time profiles different from those of stationary objects, causing their images to be misplaced and poorly focused unless compensation is made for the particular radial velocity involved. Such refinements are the subject of research.

Also being developed are techniques to access images efficiently, using various preview strategies. Having the ability to map continent-sized areas with submeter resolution introduces the problem of selecting what is of interest to image in a specific application.

SAR and ISAR radars must have good coherent processing capability, with high signal quality and phase coherence being maintained over seconds of time, much longer than that required of ordinary radar. The further inclusion of polarimetric, interferometric, and multiband capability adds to the development challenges in this fascinating radar application. *See* RADAR; REMOTE SENSING.

Robert T. Hill

Bibliography. J. C. Curlander and R. N. McDonough, *Synthetic Aperture Radar: Systems & Signal Processing*, John Wiley, 1991; S. A. Hovanessian, *Introduction to Synthetic Array and Imaging Radars*, Artech House, 1980; M. I. Skolnik (ed.), *Radar Handbook*, 2d ed., McGraw-Hill, 1990; D. R. Wehner, *High Resolution Radar*, 2d ed., Artech House, 1995.

Synthetic fuel

A gaseous, liquid, or solid fuel, also known as synfuel, that does not occur naturally. Synthetic fuel can be made from tar sand, coal, or oil shale. Included in the category are various fuel gases, such as substitute natural gas and synthesis gas. *See* COAL; OIL SAND; OIL SHALE.

Synthetic crude oil (syncrude) is a complex mixture of hydrocarbons, somewhat similar to petroleum, obtained from the bituminous material that is found in tar sand formations, from coal (liquefaction), from synthesis gas (a mixture of carbon monoxide and hydrogen), or from oil shale. Synthetic crude oil generally differs in composition from petroleum; for example, syncrude from coal usually contains more aromatic hydrocarbons than petroleum. Accordingly, the hydrogen-to-carbon atomic ratios vary widely (**Fig. 1**). Gaseous fuels can be produced from sources other than petroleum and natural gas. In addition, synthetic crude oil varies in composition and character, depending upon the source (**Table 1**). *See* PETROLEUM.

Synthetic crude oil from tar sand. Tar sand is sand saturated with a highly viscous crude hydrocarbon

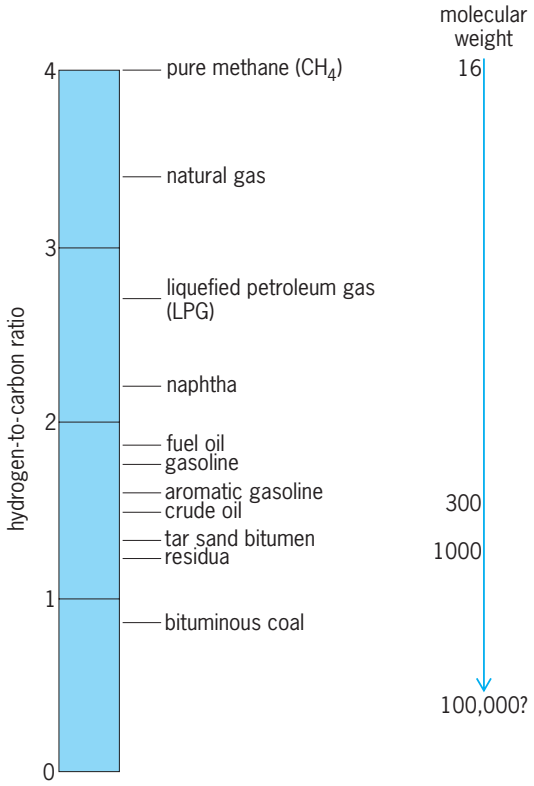


Fig. 1. Simplified comparison of the atomic hydrogen/carbon ratios and molecular weights of various fuels and fuel sources.

TABLE 1. Comparison of synthetic crude oil from various sources

Source	Tar sand bitumen	Coal	Oil shale
Gravity; °API	32	12–20	15–25
Viscosity,			
centistoke @ 100°F	6	15–25	<10
Composition, wt %			
Carbon	86	88	84
Hydrogen	13	10	11
Nitrogen	<0.1	0.3	2.0
Sulfur	<0.2	<0.1	2.0
Nickel, ppm	0	0	2
Vanadium, ppm	0	0	1
Carbon residue, wt %	0	5	5

that is not recoverable in its natural state by ordinary production methods, including enhanced oil (tertiary) recovery methods. Technically, tar sand should be called bituminous sand since the hydrocarbon is bitumen (soluble in carbon disulfide). High-grade tar sands contain about 18% by weight of bitumen that is almost equivalent in consistency to an atmospheric petroleum residuum and can be recovered by using technologies such as solvent extraction, thermal retorting, in situ combustion, and steam injection methods. In petroleum refining and in coal processing, the term "tar" is reserved to mean the cracked residue of a thermal cracking process. The term "oil sand" is used in allusion to the synthetic crude oil that can be manufactured from the bitumen. *See* BITUMEN; PETROLEUM.

A major source of synthetic crude oil is the tar sand deposits that occur in the Athabasca region of northeastern Alberta (Canada). The tar sand reserves in the United States are too deep for surface mining and require in situ treatment as the means of bitumen recovery. Some of the surface deposits are worked to produce asphalt for highway application and other minor uses. *See* ASPHALT AND ASPHALTITE.

Tar sand is a mixture of sand, water, and bitumen, and the sand component is predominantly quartz in the form of rounded or subangular particles, each of which is wet with a film of water. Surrounding the wetted sand grains and somewhat filling the void among them is a film of bitumen. The balance of the void volume is filled with connate water and sometimes a small volume of gas. *See* OIL FIELD WATERS.

In terms of bitumen separation and recovery, the hot-water process is, to date, the only successful commercial process used in North America. The process exploits the linear and the nonlinear variation of bitumen density and water density, respectively, with temperature so that the bitumen that is heavier than water at room temperature becomes lighter than water at 80°C (180°F).

In the hot-water extraction process, the tar sand is introduced into a conditioning drum in which it is heated and mixed with water to form a pulp. The pulp enters the separation cell through a central feed well and distributor. The bulk of the sand settles in the cell and is removed from the bottom, but the majority of the bitumen floats to the surface and is removed as froth. The froth from the hot-water process may be mixed with a hydrocarbon diluent, such as coker naphtha, and centrifuged.

The quality of tar sand bitumen is poor compared to that of conventional crude oil and heavy oil (**Table 2**) and has a lower proportion of distillable constituents (**Table 3**). The high carbon residue of bitumen dictates that considerable amounts of coke will be produced during thermal refining. Upgrading and refining bitumen requires a different approach to that used for upgrading heavy oil. In addition, the distance that the bitumen must be shipped to the refinery and in what form as well as product quality must be taken into account when designing a bitumen refinery. *See* COKE.

To upgrade tar and bitumen to synthetic crude oil, coking is the process of choice, and bitumen is

TABLE 2. Comparison of synthetic crude oil from bitumen with conventional petroleum and with heavy oil

Property	Athabasca bitumen (Canada)	Cold Lake heavy oil (Canada)	Lloydminster heavy oil (Canada)	Conventional crude oil (Canada)
Gravity, °API	8	12	14	35
Viscosity				
Centipoise @ 100°F (38°C)	500,000	2,000	500	10
Pour point, °F	50		5	0
Elemental analysis, wt %				
Carbon	83	84	83	86
Hydrogen	10.6	11	12	13.5
Sulfur	4.8	4.4	3.6	0.1
Nitrogen	0.4	0.4	0.4	0.2
Oxygen	1	0.2	1	0.2
Fractional composition, wt %				
Asphaltenes	19	12	12	5
Resins	32	28	17	10
Aromatics	30	35	24	25
Saturates	19	25	47	60
Metals, ppm				
Vanadium	250	190	100	10
Nickel	100	70	40	5
Carbon residue, wt %	14	11	10	5
Heating value, Btu/lb	17,500	18,000	18,200	19,500

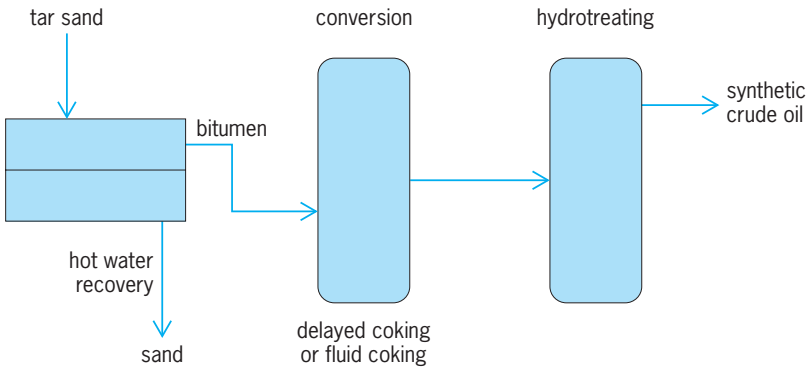


Fig. 2. Production of synthetic crude oil from tar and bitumen.

currently converted commercially by delayed coking and by fluid coking (**Fig. 2**). In each case, the charge is converted to distillate oils, coke, and light gases. The coke fraction and product gases can be used for plant fuel. The coker distillate is a partially upgraded material in itself and is a suitable feed for hydrodesulfurization to produce a low-sulfur synthetic crude oil. *See* COKING (PETROLEUM).

Delayed coking is a semibatch process in which feed bitumen is heated before being fed to coking drums that provide sufficient residence time for the cracking reactions to occur. Bitumen conversion to liquids is on the order of 75% by volume, with fluid coking giving a generally higher yield of liquids compared to delayed coking. The remainder appears as coke (approximately 15% by weight) and gases.

Fluid coking is a continuous process employing two vessels with fluid coke. It provides a better yield of overhead (distillate) products than delayed coking. Feed oil flows to the reactor vessel, where cracking and formation of coke occur; coke is combusted in the burner. Fluid transfer lines between these vessels provide the coke circulation necessary for heat balance. The proportion of coke burned is just sufficient to satisfy heat losses and provide the heat for the cracking reactions. *See* CRACKING.

In the fluid coking process, whole bitumen (or topped bitumen) is preheated and sprayed into the reactor, where it is thermally cracked in the fluidized coke bed at temperatures of 510–540°C (950–1000°F) to produce light products and coke. The coke is deposited on the fluidized coke particles, while the light products pass overhead to a scrubbing section in which any high-boiling products are condensed and recombined with the reactor fresh feed. The uncondensed scrubber overhead passes into a fractionator in which liquid products of suitable boiling ranges for downstream hydrotreating are withdrawn. Cracked reactor gases contain butanes and lower-molecular-weight hydrocarbon gases that pass overhead to a gas recovery section. The propane material ultimately flows to the refinery gas system, and the condensed butane and butenes may (subject to vapor pressure limitations) be combined with the synthetic crude. The heat necessary to vaporize the feed and to provide the heat of reaction is supplied by hot coke that is circulated back to the reactor from the coke heater. Excess coke that has formed from the fresh feed and deposited on hot circulating coke in the fluidized reactor bed is withdrawn (after steam stripping) from the bottom of the reactor. *See* FLUIDIZED-BED COMBUSTION.

Sulfur is distributed throughout the boiling range of the delayed coker distillate, as with distillates from direct coking. Nitrogen is more heavily concentrated in the higher-boiling fractions but is present in most of the distillate fractions. Raw coker naphtha contains significant quantities of olefins and diolefins that must be saturated by downstream hydrotreating. The gas oil has a high aromatic content typical of coker gas oils. *See* NITROGEN; SULFUR.

The primary liquid product from the coking process is then hydrotreated (secondary conversion or refining) to remove sulfur and nitrogen (as

TABLE 3. Distillation data for bitumen and conventional crude oil

Cut point		Cumulative wt % distilled		
°C	°F	Athabasca bitumen (Canada)	P. R. Spring bitumen (U.S.)	Leduc conventional crude oil (Canada)
200	390	3	1	35
225	435	5	2	40
250	480	7	3	45
275	525	9	4	51
300	570	14	5	
325	615	26	7	
350	660	18	8	
375	705	22	10	
400	750	26	13	
425	795	29	16	
450	840	33	20	
475	885	37	23	
500	930	40	25	
525	975	43	29	
538	1000	45	35	
538+	1000+	55	65	

hydrogen sulfide and ammonia, respectively) and to hydrogenate the unsaturated sites exposed by the conversion process. It may be necessary to use separate hydrotreaters for light distillates and medium-to-heavy fractions; for example, the heavier fractions require higher hydrogen partial pressures and higher operating temperatures to achieve the desired degree of sulfur and nitrogen removal. Commercial applications have therefore been based on the separate treatment of two or three distillate fractions at the appropriate severity to achieve the required product quality and process efficiency. *See* HYDROCRACKING; PETROLEUM PROCESSING AND REFINING.

Hydrotreating is generally carried out in downflow reactors containing a fixed bed of cobalt-molybdate catalysts. The reactor effluents are stripped of the produced hydrogen sulfide and ammonia. Any light ends are sent to the fuel gas system, and the liquid products are recombined to form synthetic crude oil. *See* CATALYSIS.

Finishing and stabilization (hydrodesulfurization and saturation) of the liquid products is achieved by hydrotreating two or three separate liquid streams. This is necessary because of the variation in conditions and catalysts required for treatment of a naphtha fraction relative to the conditions required for treatment of gas oil. It is more efficient to treat the liquid product streams separately and then to blend the finished liquids to a synthetic crude oil. In order to take advantage of optimum operating conditions for various distillate fractions, the coker distillate is treated as separate fractions, such as naphtha, kerosene, and gas oil or naphtha and mixed gas oils. The hydrotreated fractions are combined to form synthetic crude oil that is then shipped by pipeline to a refinery. The upgraded or synthetic crude oil has properties that are quite different from the original feedstock (**Table 4**) and are closer to the properties of a conventional high-API-gravity crude oil (Tables 2 and 3). *See* PETROLEUM PRODUCTS.

TABLE 4. Properties of synthetic crude oil from Athabasca bitumen

Property	Bitumen	Synthetic crude oil
Gravity, °API	8.0	32.0
Sulfur, wt %	4.8	0.2
Nitrogen, wt %	0.4	0.1
Viscosity, centipoise @ 100°F	500,000	10
Distillation profile, wt %		
0°C, 30°F	0	5
30°C, 85°F	0	30
220°C, 430°F	3	60
345°C, 650°F	17	90
550°C, 1020°F	45	100

Synthetic fuel from coal. There are several routes (**Fig. 3**) by which synthetic fuels can be prepared from coal. Gasification can yield clean gases for combustion or synthesis gas that has a controlled ratio of hydrogen to carbon monoxide. Catalytic conversion of synthesis gas to liquids (indirect liquefaction) can be carried out in fixed- and fluidized-bed reactors and in dilute-phase systems. The processes also require the addition of hydrogen to upgrade coal into finished products to accompany the molecular weight reduction that occurs during liquids production. *See* COAL.

Coal liquefaction. The chemical objectives of coal liquefaction are (1) to reduce the effect of weak bonds and thus separate fairly large units of coal structure into smaller units, (2) to bring about the

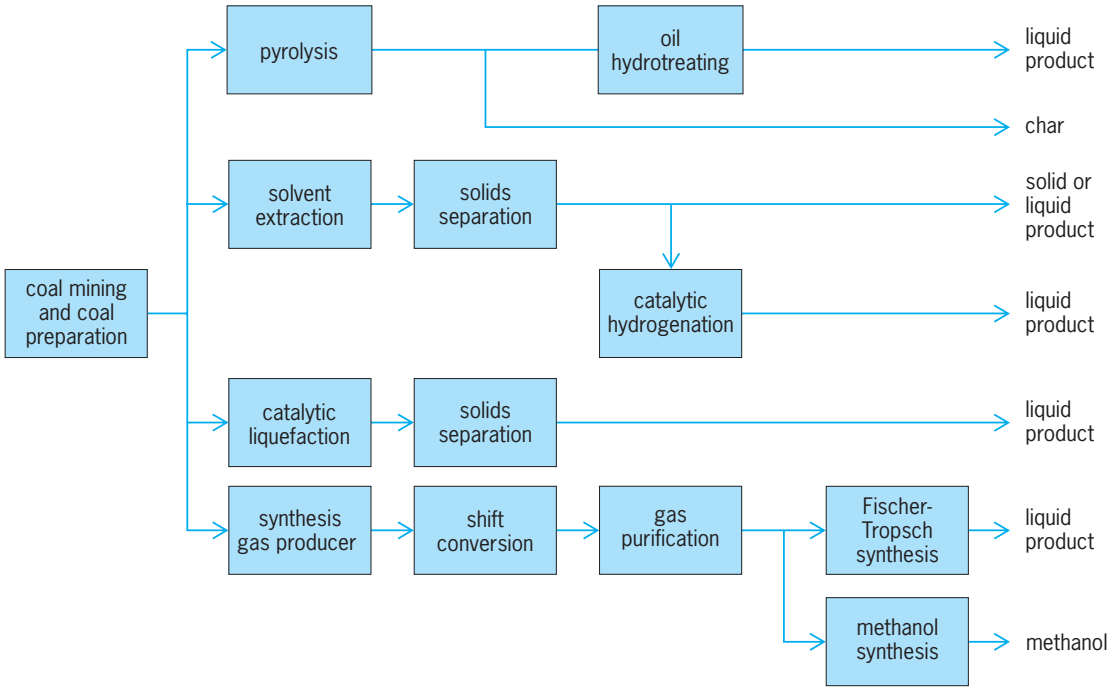


Fig. 3. Simplified schematic of the production of synthetic crude oil and syngas from coal.

decomposition of chemical bonds within the coal to form smaller fragments, and (3) to increase the hydrogen-to-carbon atomic ratio to produce a low-sulfur, ash-free liquid product that is comparable to crude oil, gasoline, or even heavy oil and bitumen.

Coal liquefaction is accomplished by four methods: (1) direct catalytic hydrogenation, (2) solvent extraction, (3) pyrolysis, and (4) indirect catalytic hydrogenation (of carbon monoxide). For example, hydrogenation of coal can be achieved by mixing coal and the catalyst with a coal-derived recycle oil (an oil that is recycled at some point in the process); and the slurry is pumped into a high-pressure system where hydrogen is present and operating conditions are on the order of 400–480°C (752–896°F) and 1000–8000 psi (6.9–55 MPa). Catalysts containing iron, molybdenum, cobalt, nickel, and tungsten are effective for such a process. *See* COAL LIQUEFACTION.

The concept of catalytic liquefaction uses a catalyst to add hydrogen to the coal. These processes usually require a liquid medium with the catalyst dispersed throughout, or may even employ a fixed-bed reactor. On the other hand, the catalyst may also be dispersed with the coal, whereupon the combined coal-catalyst system can be injected into the reactor. *See* HYDROGENATION.

In the catalytic liquefaction processes, coal is brought into contact with a catalyst in the presence of hydrogen. Many processes of this type have the advantage of eliminating the need for a hydrogen donor solvent (and the subsequent rehydrogenation of the spent solvent), but there is still the need for an adequate supply of hydrogen. The nature of the process also virtually guarantees that the catalyst will be deactivated by the mineral matter in the coal as well as by coke lay-down during the process. *See* HYDROGEN.

In order to achieve the direct hydrogenation of the coal, the catalyst and the coal must be in intimate contact, but if this is not the case, process inefficiency is the general rule. There has been the tendency of late to achieve coal-catalyst contact by the use of a hydrogen donor solvent. Alternatively, a catalyst with a sufficiently high vapor pressure may be used so that catalyst deposition on the coal surface is achieved under process conditions.

The major features of catalytic liquefaction processes are (l) rapid heating to temperatures of the order of 450–600°C (840–1110°F), (2) short residence times, and (3) reactor effluent quenching. Instead of using a slurry-feed system, the coal is entrained in a rapidly moving stream of hydrogen and has a residence time of (usually) less than 20 s in the reactor at 500°C (930°F) and 2000 psi (13.8 MPa).

In solvent extraction processes, coal is mixed with a solvent that is capable of effecting the transfer of hydrogen from the solvent to the coal or even from gaseous hydrogen to the coal) at temperatures up to 500°C (930°F) and pressures up to 5000 psi (34.5 MPa). Processing by solvent extraction solubilizes and disperses coal in a solvent that transfers some of its hydrogen to the coal. Ash and insoluble coal are separated from the liquid product to recover recycle oil and product oil. The carbonaceous residue is reacted with steam in a gasifier to produce the hydrogen needed. Hydrogenation of the solvent is also practiced to improve the product quality. *See* SOLVENT EXTRACTION.

Solvent extraction processes generally use liquids (often derived from the feed coal) as donor solvents that are capable of donating hydrogen to the system under the conditions of the reaction. The overall result is an increase (relative to pyrolysis processes) in the amount of coal that is converted to lower-molecular-weight, that is, soluble, products. Reaction temperatures usually have an upper limit of 510°C (950°F); hydrogen may be supplied (under pressure) during the process, or it may be introduced by hydrogenation of the solvent prior to the process. Such hydrogen may be produced from any unreacted coal, from feed coal, or from by-product gases.

High-temperature solvent extraction processes of coal have been developed in three process configurations: (1) extraction in the absence of hydrogen but using a recycle solvent that has been hydrogenated in a separate process stage; (2) extraction in the presence of hydrogen with a recycle solvent that has not been previously hydrogenated; and (3) extraction in the presence of hydrogen using a hydrogenated recycle solvent.

The solvent extraction concept can also be achieved under milder conditions, but the product may be high-nitrogen solid or heavy oil with only low yields of light oils and gases. Severe conditions are more effective for sulfur and nitrogen removal to produce a "lighter" liquid product that is more amenable to downstream processing.

A novel approach to the solvent extraction process uses bitumen or heavy oil as the process solvent. In fact, the coprocessing of coal with a variety of petroleum-based feedstocks (such as heavy oils) has received attention. The concept is a variation of the solvent extraction process for coal liquefaction. Whether the coprocessing option is a means of producing more liquids or whether the coal should act as a scavenger for the metals and nitrogen species in the petroleum material is dependent upon the process conditions. What is certain is that special effort should be made to ensure the compatibility of feedstocks and products. Incompatibility can, at any stage of the liquefaction operation, lead to expensive shutdowns as well as to (in respect of the heavier feedstocks) the onset of coke formation.

Pyrolysis processes involve heating coal to temperatures in excess of 400°C (750°F), which results in the conversion of the coal to gases, liquids, and char (carbon). The char is hydrogen-deficient, thereby enabling inter- or intramolecular hydrogen transfer processes to be operative, resulting in relatively hydrogen-rich gases and liquids. Unfortunately, the char produced often amounts to more than 45% by weight of the feed coal and, therefore, such processes have often been considered to be an uneconomical or inefficient use of coal. *See* PYROLYSIS.

Pyrolysis (or carbonization) is perhaps the oldest technique for obtaining liquid products directly from

coal and involves heating the coal in the absence of air (or oxygen) to produce heavy and light oils, gases, and char. In the presence of hydrogen, the process is called hydrocarbonization. The composition and relative amounts of the products formed are dependent on the process parameters such as heating rate, pressure, coal type, coal (and product) residence time, coal particle size, and reactor configuration. A major disadvantage of this type of process is the large yields of char, reducing the yield of liquid products.

On the other hand, pyrolysis and hydrocarbonization processes are less complex than liquid-phase hydrogenation processes. The operating pressures for pyrolysis processes are usually less than 100 psi (690 kPa; more often between 5 and 25 psi or 34–172 kPa), but hydrocarbonization processes require hydrogen pressures of the order of 300–1000 psi (2.1–6.9 MPa). In both processes, the operating temperature can be as high as 600°C (1110°F).

Three types of pyrolysis reactors are of interest: mechanically agitated entrained-flow reactors and fluidized-bed reactors. The agitated reactor may be quite complex, but the entrained-flow reactor has the advantage of either down-flow or up-flow operation and can provide short residence times. In addition, the coal can be heated rapidly, leading to higher yields of liquid (and gaseous) products that may well exceed the volatile matter content of the coal as determined by the appropriate test. The short residence time also allows a high throughput of coal and the potential for small reactors.

Coal gasification. The other category of producing synthetic fuel (gas and liquid) from coal invokes the concept of the coal gasification followed by conversion of the gaseous products to liquid fuel. This is often referred to as indirect liquefaction of coal insofar as the coal is not converted directly into liquid products. *See* COAL GASIFICATION.

The indirect liquefaction of coal involves a two-stage conversion operation in which coal is first converted (by reaction with steam and oxygen) to a gaseous mixture that is composed primarily of carbon monoxide and hydrogen (syngas, or synthesis gas). The gas stream is subsequently purified (to remove sulfur, nitrogen, and any particulate matter), after which it is catalytically converted to a mixture of liquid hydrocarbon products. In addition to the production of fuel gas, coal may be gasified to prepare synthesis gas (carbon monoxide and hydrogen), which may be used to produce a variety of products, including ammonia, methanol, and liquid hydrocarbon fuels.

The synthesis of hydrocarbons from carbon monoxide and hydrogen (synthesis gas) is a procedure (the Fischer-Tropsch synthesis) for the indirect liquefaction of coal. This process is the only coal liquefaction scheme currently in use on a relatively large commercial scale; South Africa is currently using the Fischer-Tropsch process on a commercial scale. Notably, Germany produced roughly 156 million barrels of synthetic petroleum annually using the Fischer-Tropsch process during World War II. *See* FISCHER-TROPSCH PROCESS.

Coal is converted to gaseous products at temperatures in excess of 800°C (1470°F), and at moderate pressures, to produce synthesis gas:

$$C_{coal} + H_2O \rightarrow CO + H_2$$

The purified gas has a heating value of 325 Btu/ft^3 (12,135 kJ/m^3) if oxygen is used or about 100 Btu/ft^3 (3725 kJ/m^3) if air is used. In some processes, light hydrocarbons are formed by cracking of higher-molecular-weight compounds, and the heating value can be higher.

There are several processes systems for coal gasification, including the commercially available Lurgi, Koppers-Totzek, or Winkler processes. Many coal gasification units are in operation throughout the world to produce synthesis gas or fuel gas. Modern gasification systems prepare synthesis gas from coal or oil for methanol, ammonia, and oxo-alcohol production. The gasification may also be attained by means of gasification of the coal in place (underground gasification or in situ gasification). The exothermic nature of the process and the decrease in the total gas volume in going from reactants to products suggest the most suitable experimental conditions to use in order to maximize product yields. The process should be favored by high pressure and relatively low reaction temperature. In practice, the Fischer-Tropsch reaction is carried out at temperatures on the order of 200–350°C (390–660°F) and at pressures of 75–4000 psi (0.5–28 MPa); the hydrogen/carbon monoxide ratio is usually approximately 2.2:1 or 2.5:1. *See* COAL MINING.

Since up to three volumes of hydrogen may be required to achieve the next stage of the liquids production, the synthesis gas must then be converted (by means of the water-gas shift reaction) to the desired level of hydrogen:

$$CO + H_2O \rightarrow CO_2 + H_2$$

Next, the gaseous mix is purified and converted to a wide variety of hydrocarbons:

$$CO + (2n + 1)H_2 \rightarrow C_nH_{2n+2} + n\,H_2O$$

This method of the catalytic hydrogenation of carbon monoxide is also a method of producing synthetic liquid fuel from coal. Catalysts can be prepared from iron, nickel, cobalt, ruthenium, and zinc either in a metal salt or contained on a support. Each catalyst gives a different product distribution that is also a function of the method of preparation and pretreatment. Primary products are normally methane and higher-molecular-weight hydrocarbons, alcohols, and organic acids. Operating conditions for the Fischer-Tropsch synthesis are usually in the range of 300–500 psi (2–3.5 MPa) and 200–400°C (392–752°F). These reactions result primarily in low- and medium-boiling aliphatic compounds; present commercial objectives are focused on the conditions that result in the production of *n*-hydrocarbons as well as olefinic and oxygenated materials.

Although the efficiency of a two-stage process as compared with a single-stage liquefaction process has often been cited as a disadvantage of the indirect liquefaction of coal, the advantages of the indirect approach are that (1) the gasification stage can tolerate a higher degree of impurities (such as mineral matter) in the coal; (2) the feed to the liquefaction stage can be maintained at the desired hydrogen level regardless of the nature and quality of the feed coal; (3) a "clean" product is produced; (4) the method is adaptable to the use of the products from the underground gasification of coal; and (5) there is the added advantage that hydrogen from a variety of other hydrogen production methods may be used, thereby increasing the overall production of liquids from a given amount of coal.

The underground gasification of coal in place is also an option for coal conversion that does not require a mining operation. It is suitable for difficult-to-mine coal and seams that are below an economical thickness. The concept was patented in 1909 in the United States. In this process, air, oxygen, or a mixture of steam and air or of steam and oxygen is introduced into the coal seam through boreholes or shafts. The burned coal produces a gaseous product that is recovered through properly spaced boreholes or shafts.

Underground gasification affords the recovery of energy from unminable coal, greater recovery of the coal seam, and negligible impact on the health and safety of miners. In some systems, miners are needed only for the preparation of shafts and galleries. There are no spoil banks, slack piles, or acid mine drainage problems. On the other hand, problems can occur with control of the combustion, roof support, bed permeability, ground-water control, leakage to adjacent strata, and monitoring of the underground events.

Synthetic fuel from oil shale. Shale oil (synthetic crude oil from oil shale) is produced by the thermal processing of oil shale at high temperatures in a variety of retorts or even in situ (that is, without being mined). The ingredient in the shale that is decomposed, kerogen, is an insoluble material containing, in addition to carbon and hydrogen, considerable amounts of nitrogen, oxygen, and sulfur. The resources of oil shale in the United States are substantial, easily dwarfing the resources of petroleum. The conversion of oil shale to shale oil has received commercial setbacks due to fluctuations in the price of petroleum that have rendered many of the technologies uneconomical. *See* KEROGEN.

Crude shale oil, sometimes called retort oil, is the liquid oil condensed from the effluent in oil shale retorting (**Fig. 4**). Crude shale oil typically contains appreciable amounts of water and solids, and tends to form sediments. As a result, it must be upgraded to a synthetic crude oil before being suitable for pipelining or substitution for petroleum crude as a refinery feedstock. However, shale oils are sufficiently different from petroleum that processing shale oil presents some unusual problems.

Shale oils, especially those from the Green River

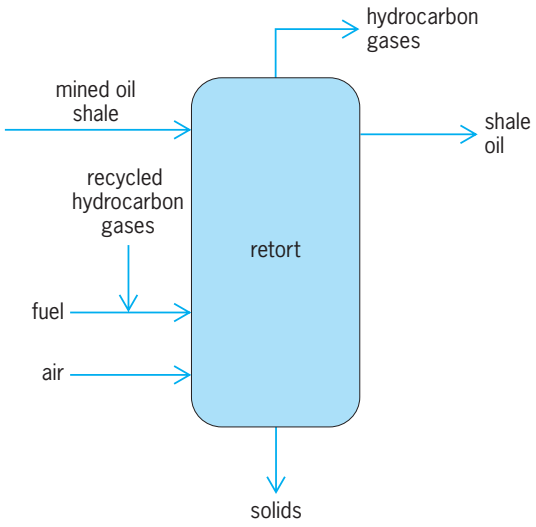


Fig. 4. Production of synthetic crude oil from oil shale.

in southern Wyoming, have particularly high nitrogen contents, typically 1.7–2.2 wt % compared to approximately 0.2–0.3 wt % for conventional (Table 2) petroleum. In many other shale oils (including those from eastern United States oil shale), nitrogen contents are lower than in the Green River shale oils but still higher than those typical of petroleum. Because retorted shale oils are produced by a thermal cracking process, olefin (—CH═CH—) and diolefin [—CH═CH—$(CH_2)_n$—CH═CH—, where $n \geq 0$] contents are high. It is the presence of these olefins and diolefins, in conjunction with high nitrogen contents, which gives crude shale oils their characteristic instability toward sediment formation. The sulfur content of shale oils varies widely, but is generally lower than that of high-sulfur petroleum crude oil and tar sand bitumen.

The Green River shale oils contain appreciable amounts of aromatics, polar aromatics, and pentane-insolubles. Oxygen contents are higher than those typically found in petroleum but lower than those of crude coal liquids. Crude shale oil also contains appreciable amounts of soluble arsenic, iron, and nickel that cannot be removed by filtration. *See* ARSENIC; IRON; NICKEL.

Upgrading, or partial refining, to improve the properties of a crude shale oil may be carried out with different objectives, depending on the intended use for the product: (1) stabilization to produce a product that can be transported to a distant refinery by pipeline; (2) more complete upgrading to produce a premium refinery feedstock with low nitrogen, low sulfur, and essentially no residuum; (3) upgrading to produce chemical feedstock streams; and (4) complete refining of the crude shale oil or selected fractions to produce finished end products (such as gasoline, diesel, jet fuel).

The arsenic and iron in shale oil would poison and foul the supported catalysts used in hydrotreating. Because these materials are soluble, they cannot be removed by filtration. Several methods have been used specifically to remove arsenic and iron. Other

methods involve hydrotreating; these also lower sulfur, olefin, and diolefin contents and thereby make the upgraded product less prone to form gum. Shale oil is also rich in high-molecular-weight, waxy paraffinic material. Thermal cracking lowers molecular weight but yields straight-chain products of low octane number. Fluid catalytic cracking not only lowers molecular weight but also causes isomerization to produce branched products with higher octane numbers. As a result, the "cat cracked" shale naphtha is a more desirable feedstock for hydrotreating to make gasoline blend stock than is the naphtha from thermal cracking or coking of shale oil. *See* HYDROCRACKING; ISOMERIZATION.

James G. Speight

Bibliography. J. G. Speight (ed.), *The Chemistry and Technology of Petroleum*, 3d ed., Marcel Dekker, 1999; J. G. Speight (ed.), *The Desulfurization of Heavy Oils and Residua*, 2d ed., Marcel Dekker, 2000; J. G. Speight (ed.), *Fuel Science and Technology Handbook*, Marcel Dekker, 1990; J. G. Speight, *Gas Processing: Environmental Aspects and Methods*, Butterworth-Heinemann, Oxford, England, 1993.

Syphilis

A sexually transmitted infection of humans caused by *Treponema pallidum* ssp. *pallidum*, a corkscrew-shaped motile bacterium (spirochete). Due to its narrow width, *T. pallidum* cannot be seen by light microscopy but can be observed with staining procedures (silver stain or immunofluorescence) and with dark-field, phase-contrast, or electron microscopy. *Treponema pallidum* cannot be continuously cultured in vitro by standard bacteriological methods. For antigen production and experimental studies, the organism is propagated by inoculation of laboratory rabbits. Limited multiplication of *T. pallidum* has been obtained in a tissue culture system at a temperature of 91–95°F (33–35°C) in an atmosphere containing 1–5% oxygen and in media containing reducing agents. The organism is very sensitive to environmental conditions and to physical and chemical agents. The complete genome sequence of the *T. pallidum* Nichols strain (1.13×10^6 base pairs) has been determined. Analysis of the nucleotide sequence of the small, circular treponemal chromosome indicates that *T. pallidum* lacks the genetic information for many of the metabolic activities found in other bacteria. Thus, this spirochete is dependent upon the host for most of its nutritional requirements. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); ELECTRON MICROSCOPE.

Syphilis is usually transmitted through direct sexual contact with active lesions and can also be transmitted by contact with infected blood and tissues. If untreated, syphilis progresses through various stages (primary, secondary, latent, and tertiary). Infection begins as an ulcer (chancre) and may eventually involve the cardiovascular and central nervous systems, bones, and joints. Congenital syphilis results from maternal transmission of *T. pallidum* across the placenta to the fetus. *See* SEXUALLY TRANSMITTED DISEASES.

Epidemiology. *Treponema pallidum* is an obligate parasite of humans and does not have a reservoir in animals or the environment. Syphilis has a worldwide distribution. Its incidence varies widely according to geographical location, socioeconomic status, and age group. Although syphilis is controlled in most developed countries, it remains a public health problem in many developing countries. Epidemics of syphilis recently occurred in the former Soviet Union due to conditions that have limited identification and treatment of infected individuals. Approximately 6000 cases of primary and secondary syphilis were reported in the United States in 2000. This represents a dramatic decrease from the post–World War II peak of 51,060 cases of primary and secondary syphilis reported in 1990. The number of reported cases of congenital syphilis also declined from 4438 cases in 1991 to approximately 300 cases in 2000. Various studies have shown that syphilis is a risk factor for infection with the human immunodeficiency virus (HIV) since syphilitic lesions may act as portals of entry for the virus. There is little natural immunity to syphilis infection or reinfection. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS).

Clinical manifestations. Syphilis has been called "the great imitator" because its diverse clinical manifestations can mimic many infectious and noninfectious diseases. *Treponema pallidum* enters the body through microscopic abrasions of skin or mucosal membranes, colonizes host cell surfaces, and begins to multiply extracellularly at the site of infection. Within a few hours, significant numbers of treponemes depart the local site and are carried to the regional lymph nodes. The organisms spread throughout the body via the circulation. The incubation period of syphilis ranges from 10 to 90 days. Clinical signs and symptoms of the primary stage include a painless chancre at the site of infection (genital, oral, or anal) and enlargement of regional lymph nodes. The chancre and lymph nodes contain treponemes, and the disease is communicable. Although the untreated chancre heals in a few weeks, presumably due to the host immune response, infectious treponemes are still present in the individual.

The secondary or disseminated stage of syphilis occurs within a few weeks of the healing of the primary lesion(s). It is characterized by low-grade fever, headache, generalized enlargement of lymph nodes, malaise, loss of appetite (anorexia), and a hyperpigmented skin rash that appears on the palms and soles and eventually spreads to other areas. Superficial sores (mucous patches) may occur in the mouth, vagina, or anus, and wartlike lesions (condylomata) may form in moist areas of the body. Severe hepatitis and meningitis are occasionally observed. Although secondary lesions disappear within a few weeks, the untreated person remains infectious.

The latent stage of syphilis represents conversion from acute to chronic disease. Relapses of secondary

syphilis may occur early in latency. After the first year of infection, the individual usually becomes clinically asymptomatic. However, treponemes continue to persist in infected tissues. Although transmission of syphilis from an infected pregnant woman to her fetus can occur during the latent stage, the absence of exposed lesions reduces the probability of transmission to a sexual partner. Approximately two-thirds of persons with untreated syphilis will remain in the latent stage for the remainder of their lives.

Tertiary (late) syphilis occurs in the remaining one-third of untreated cases within months to decades after the initiation of latency. Treponemes invade the central nervous system, cardiovascular system, eye, skin, bones, or internal organs where they produce damage due to a host delayed-type hypersensitivity reaction to treponemal antigens. Approximately 80% of syphilis mortalities are caused by cardiovascular involvement, while 20% are from neurologic involvement. Destructive, soft, tumorlike lesions (gummas) may develop in skin, bone, and other tissues. During the tertiary stage, sexual transmission of syphilis does not occur and congenital transmission is rare.

Congenital syphilis results from transplacental transmission of *T. pallidum* from an infected pregnant woman to her fetus. The extent of damage to the fetus depends on the disease stage at the time of transmission. Overwhelming infection of the fetus may result in miscarriage or stillbirth. Infants born with congenital syphilis may have skin lesions, mucoid nasal discharge (snuffles), long-bone malformation, anemia, or involvement of the central nervous system. Clinical manifestations of congenital syphilis may be delayed until puberty or adulthood. Signs of late congenital syphilis may include blindness, deafness, tooth malformation, bone involvement, soft lesions, neurosyphilis, and, rarely, cardiovascular disease. Late congenital syphilis is not infectious.

Diagnosis. The diagnosis of syphilis is aided by direct microscopic examination of specimens obtained from genital and skin lesions and the detection of antibodies that develop during infection. Microscopic examination of specimens from oral and anal lesions is not reliable due to the presence of spirochetal species that are members of the normal flora. Nontreponemal screening tests such as the Rapid Plasma Reagin (RPR) or the Venereal Disease Research Laboratory (VDRL) detect serum antibodies formed by the patient in response to cardiolipin, a phospholipid. These tests can yield false-negative reactions in early syphilis due to their limited sensitivity, and false-positive reactions can occur with patients who are pregnant or have autoimmune diseases. Treponemal tests such as the fluorescent treponemal antibody absorbed (FTA-ABS) test and the *Treponema pallidum* particle agglutination test are used to confirm reactive screening test (TPPA) results.

Recombinant-based *T. pallidum* proteins have been recently used in Europe as antigens in enzyme-linked immunosorbent assays (ELISA) and Western blot assays for the serodiagnosis of syphilis. These newer assays have advantages over certain serologic tests for syphilis, such as the FTA-ABS test and the TPPA test, since they do not require the use of native *T. pallidum* antigens and are compatible with automation. *See* ANTIBODY; ANTIGEN; FLUORESCENCE MICROSCOPE; SEROLOGY.

Treatment. Parenteral penicillin G is the preferred antibiotic for treatment of all stages of syphilis. Alternative antibiotics for syphilis treatment include erythromycin and tetracycline. However, tretracycline is contraindicated for pregnant women and young children. Although penicillin resistance has not emerged in *T. pallidum*, erythromycin resistance has been documented in a clinical isolate, and erythromycin treatment failures have been reported. *See* ANTIBIOTIC.

Prevention and control. The prevention and control of syphilis are dependent upon case reporting to health authorities, contact tracing, adequate treatment of cases and contacts, education of high-risk populations, use of condoms, prenatal screening of pregnant women, and screening of blood donors. There is currently no vaccine to prevent syphilis. However, it is anticipated that information obtained from the *T. pallidum* genome sequence will lead to further improvements in diagnostic tests for syphilis and to the eventual development of a vaccine that would prevent infection. Certain proteins encoded by the multigene *T. pallidum tpr* family are under investigation as potential vaccine candidates. *See* PUBLIC HEALTH.

Lola V. Stamm

Bibliography. C. M. Fraser et al., Complete genome sequence of *Treponema pallidum*, the syphilis spirochete, *Science*, 281:375–387, 1998; K. K. Holmes et al. (eds.), *Sexually Transmitted Diseases*, 3d ed., McGraw-Hill, New York, 1999; P. R. Murray et al. (eds.), *Manual of Clinical Microbiology*, 7th ed., ASM Press, Washington, DC, 1999; *1998 Guidelines for Treatment of Sexually Transmitted Diseases*, MMWR 47 (RR-1):28–48; L. V. Stamm and H. L. Bergen, A point mutation associated with bacterial macrolide resistance is present in both 23S rRNA genes of an erythromycin-resistant, *Treponema pallidum* clinical isolate, *Antimicrob. Agents Chemother.*, 44:806–807, 2000.

Systellommatophora

An order (or superorder) in the gastropod mollusc an subclass Pulmonata containing three families of sluglike mollusks that lack any trace of a shell, have separate external male and female orifices, lack a mantle cavity, have a posterior anus and excretory pore, and bear eyes on the tops of two contractile, but not retractile, tentacles. The Rathouisiidae include about 30 species of carnivorous land slugs and range from South China and India to Queensland, Australia, in very wet tropical forests. They have slender, often brightly colored bodies. The Veronicellidae comprise about 200 species of slugs that feed on decaying plant matter. Their distribution is now

pantropical (including Florida and south Texas), owing to accidental horticultural introductions. They are minor garden and crop pests in several subtropical and tropical areas. They have long flattened bodies with usually dull colors and are very abundant nocturnally in garden and lawn areas. The Onchidiidae include perhaps 250 species of intertidal to high-subtidal dwellers, reaching greatest abundance in the rocky shore areas of the tropics. They range from a few inches to nearly 4 in. (100 mm) in length. Many species have dorsal eyespots or branchial plumes present, and can be confused with opisthobranch snails. *See* PULMONATA.

G. Alan Solem

System design evaluation

A comprehensive and life-cycle assessment of the effectiveness and cost of competing system designs, in order to choose the best candidate. System design evaluation is essential within the systems engineering process. It should be embedded appropriately within the process and then pursued continuously as system design and development progress. This article focuses on (1) the placement of design evaluation within the systems engineering process, (2) the evaluation factors that underpin system design evaluation, and (3) the system selection decision.

Evaluation within systems engineering. The systems engineering process is supported by a morphology for linking technology to customer needs (**Fig. 1**). It is composed of three major activities: synthesis, analysis, and evaluation.

Synthesis. System synthesis is accomplished by combining top-down and bottom-up design activities. The focus should be on customer requirements, normally expressed in functional terms (Fig. 1, block 0). Key elements are the design team (block 1) interacting with traditional and computer-based tools for design synthesis (block 2). Existing components, parts, and subsystems (block 5) are integrated to synthesize candidate systems. The product is a candidate system design that should be evaluated against the need and compared with other candidates.

Analysis. Analysis of the candidate system design is a necessary but not sufficient ingredient in system design evaluation. It involves estimating and predicting design dependent parameter (DDP) values (Fig. 1, block 3), forecasting design independent parameter (DIP) values (block 5), and maintaining physical and economic databases. System analysis and operations research is a step on the way to system design evaluation, but adaptation of the models and techniques is required.

Evaluation. Evaluation of each candidate system design (Fig. 1, block 4) is accomplished after receiving design dependent parameter values for the candidate from block 3. It is the specific values for design dependent parameters that differentiate (or instance) candidate systems. Design independent parameter values from block 5 are externalities: They apply across all candidate systems. Each candidate is optimized in block 4 before being subjected to the design decision schema (block 6). It is here that the best candidate is sought (based on the customer's subjective evaluation). *See* SYSTEMS ENGINEERING.

Evaluation factors. To be comprehensive, system design evaluation must encompass both system effectiveness and life-cycle cost. To be rigorous, evaluation should be pursued with the aid of models and

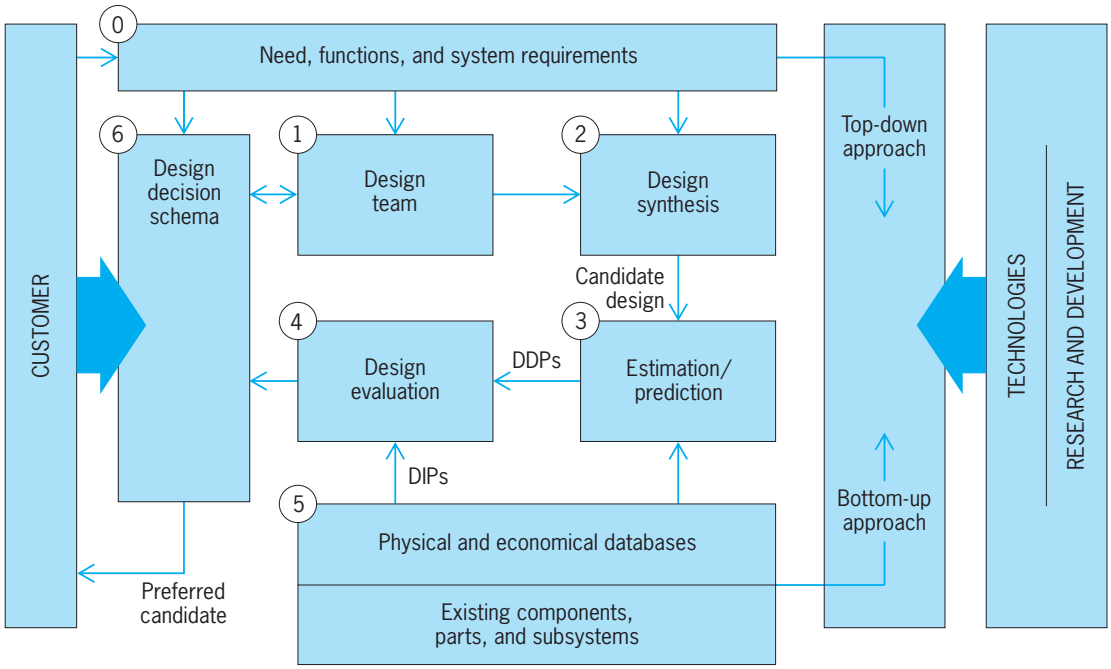


Fig. 1. Systems engineering schematic.

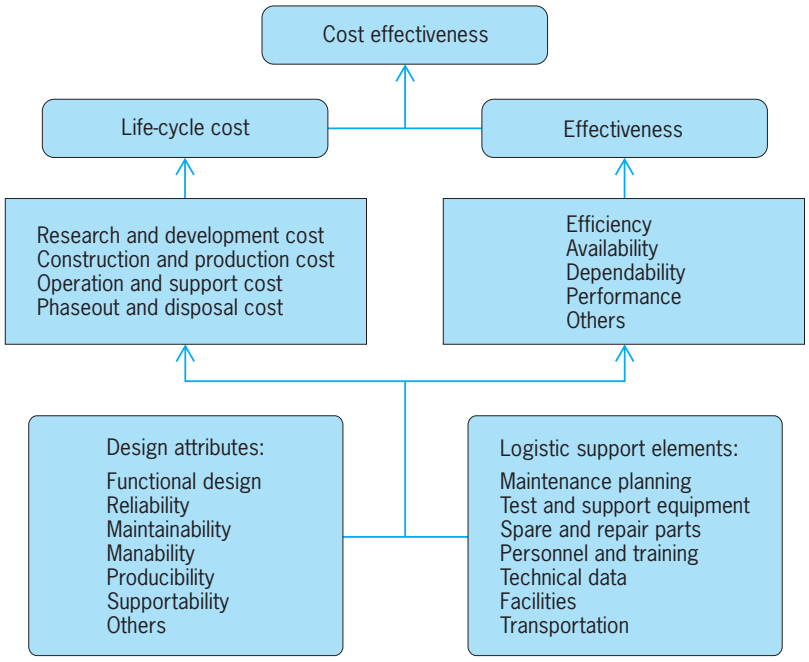


Fig. 2. Factors to be evaluated.

simulation. Even when expert opinion must be used in place of formal analytical approaches, a modeling structure may offer guidance. *See* MODEL THEORY; SIMULATION.

Cost and effectiveness. Life-cycle cost arises from the four life-cycle phases: research and development; construction and production; operation and support; phaseout and disposal (**Fig. 2**). Effectiveness is a multidimensional measure based on mission fulfillment in terms of a stated need. Mission fulfillment may be expressed by one or more criteria (effectiveness measures), depending on the type of system and the objectives it seeks to satisfy. Some common effectiveness measures are shown on the right side of Fig. 2.

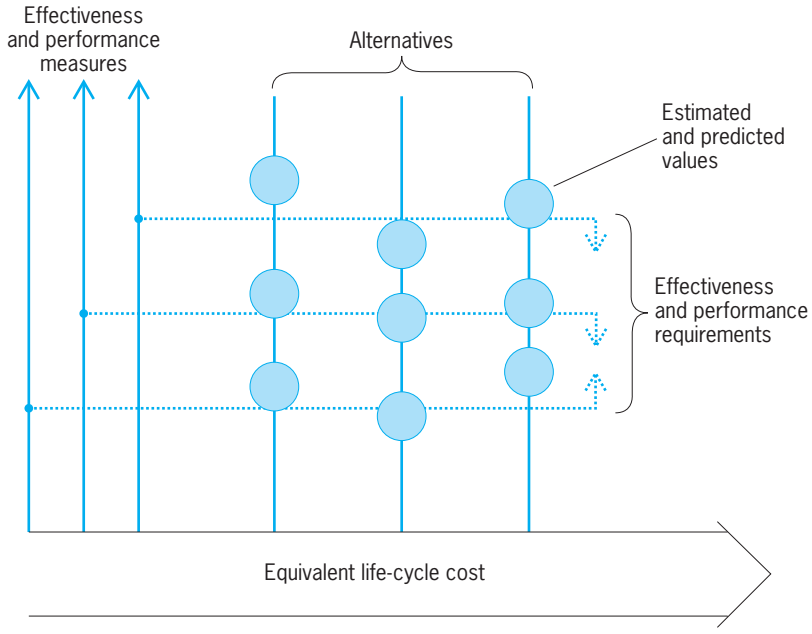


Fig. 3. Design evaluaton display.

Design dependent parameters are design characteristics inherent in the product or system (for example, reliability, maintainability, and disposability). They define the design space and are subject to manipulation by the design team during the process of seeking the best design. The objective is to consider and balance all applicable cost and effectiveness factors.

Evaluation mathematics. The evaluation process incorporates a paradigm based on normative models from the domains of operations research and systems analysis. This paradigm involves the identification and incorporation of design dependent parameters (Fig. 1, block 3) to mathematically link design characteristics with operational outcomes. Optimization and tradeoff decisions are facilitated during the systems engineering process with the aid of a design evaluation function (block 4) and a design evaluation display (block 6).

The design dependent parameter approach utilizes the design evaluation function, expressed as

$$E = f(X; Y_d, Y_i)$$

where E = a life-cycle complete evaluation measure (usually equivalent life-cycle cost incorporating the time value of money and inflation factors); X = design variables (such as number of deployed units, armor thickness, retirement age, number of repair channels, or pier spacing); Y_d = design dependent parameters (for example, design life, reliability, maintainability, weight, capacity, and disposability); Y_i = design independent parameters (for example, cost of money, labor rates, material cost per unit, energy cost per unit, and shortage cost penalty).

The design evaluation function, with its design dependent and design independent parameters, facilitates design optimization. It provides the mathematical basis for clarifying the true difference between alternatives (a design-based choice among a set of design dependent parameters, Y_d) and optimizing a given system (a search-based choice among a set of design variables, X).

System selection decision. The factors used to evaluate candidate system design must be aggregated to make them apparent. Accordingly, an evaluation process must be followed which combines evaluation factors and decision making. The candidate system design that complies best with the customer's requirement is the preferred choice and may be recommended for implementation.

System design evaluation must recognize and incorporate multiple criteria if it is to be viable as part of the systems engineering process. Multiple criteria considerations arise when both economic and noneconomic elements are present in the evaluation. In these common situations, design evaluation is facilitated by the use of a design evaluation display (DED), exhibiting both cost and effectiveness measures (**Fig. 3**). Life-cycle cost and one or

more effectiveness measures may be displayed simultaneously as an aid in decision making. Effectiveness requirements, or thresholds, are shown in fig. 3. These are useful to the decision maker in subjectively assessing the degree to which each candidate system satisfies effectiveness criteria. Life-cycle cost is an objective measure. The goal is to select the alternative with the lowest life-cycle cost that satisfies the effectiveness measures to an acceptable degree.

Formal trade-off between life-cycle cost and effectiveness measures would require the application of preference functions and utility theory. This is sometimes done, but in practice a simple weighting scheme is often used. In many applications it may be sufficient to have the decision maker or team decide by visually inspecting a design evaluation display.

The commitment to technology, system configuration, performance, and life-cycle cost is particularly important during the early stages of system design. A large information gap exists between this commitment and the system-specific knowledge available during conceptual and preliminary design. Evaluation approaches utilizing modeling and indirect experimentation may be used to help narrow this gap to reduce development risk. *See* DECISION SUPPORT SYSTEM; OPERATIONS RESEARCH; OPTIMIZATION; SYSTEMS ANALYSIS.

Wolter J. Fabrycky

Bibliography. B. S. Blanchard and W. J. Fabrycky, *Systems Engineering and Analysis*, 3d ed., 1998; C. W. Churchman, R. L. Ackoff, and E. L. Arnoff, *Introduction to Operations Research*, 1957; W. J. Fabrycky and B. S. Blanchard, *Life-Cycle Cost and Economic Analysis*, 1991; G. A. Hazelrigg, *Systems Engineering*, 1993; A. P. Sage and W.B. Rouse, *Handbook of Systems Engineering and Management*, 1999.

System families

Large systems that are formed from a variety of component systems: custom systems, which are newly engineered from the "ground up;" existing commercial-off-the-shelf (COTS) systems that are subsequently tailored for a particular application; and existing or legacy systems. Such related terms as systems of systems (SOS), federations of systems (FOS), federated systems of systems (F-SOS), and coalitions of systems (COS) are often used to characterize these systems. These appellations capture important realities brought about by the fact that modern systems are not monolithic. Rather, they have five characteristics, initially summarized by Mark Maier, that make one of the system family designations appropriate:

1. *Operational independence of the individual systems*. A system of systems is composed of systems that are independent and useful in their own right. If a system of systems is disassembled, the constituent systems are capable of independently performing useful operations by themselves and independently of one another.

2. *Managerial independence of the systems.* The component systems generally operate independently in order to achieve the technological, human, and organizational purposes of the individual unit that operates the system. These component systems are generally individually acquired, serve an independently useful purpose, and often maintain a continuing operational existence that is quite independent of the larger system of systems.

3. *Geographic distribution*. Geographic dispersion of the constituent systems in a system of systems is often quite large. Often, these constituent systems can readily exchange only information and knowledge with one another, and not substantial quantities of physical mass or energy.

4. *Emergent behavior*. The system of systems performs functions and carries out purposes that do not reside uniquely in any of the constituent systems. These behaviors arise as a consequence of the formation of the entire system of systems and are not the behavior of any constituent system. The principal purposes supporting engineering of these individual systems and the composite system of systems are fulfilled by these emergent behaviors.

5. *Evolutionary and adaptive development*. A system of systems is never fully formed or complete. Development of these systems is evolutionary and adaptive over time, and structures, functions, and purposes are added, removed, and modified as experience of the community with the individual systems and the composite system grows and evolves.

Unfortunately, there is no universally accepted definition of these system families, and at present there are no definitive criteria that distinguish a system of systems from other systems. In a formal sense, almost anything could be regarded as a system of systems. A personal computer is a system. However, the monitor, microprocessor, disk drive, and random-access memory are also systems. Thus, should a personal or a mainframe computer be called a system of systems? Is the Internet a system of systems? Instinctively, they are quite different. The big difference in these two examples is that the former system may be massive in size but monolithic in purpose, while the latter is capable of supporting the myriad communications and commerce purposes of utilizing organizations and humans. There is also a distinction based on operational and managerial independence of the system components and in evolution and emergence possibilities. *See* INTERNET; MICROCOMPUTER.

Federations and coalitions of systems. Often, appropriate missions exist for relatively large systems of systems in which there is a very limited amount of centralized command-and-control authority. Instead, a coalition of partners has decentralized power and authority and potentially differing perspectives of situations. It is useful to term such a system a "federation of systems" and sometimes a "coalition of systems." The participation of the federation or

coalition of partners is based upon collaboration and coordination to meet the needs of the federation or coalition.

Innovation and change. Support for innovation and change of all types is a desirable characteristic of these system families. Innovation includes both technological innovation and organizational and human conceptual innovation. Accomplishing this requires continuous learning, a reasonable tolerance for errors, and experimental processes to accomplish both the needed learning and the needed change. The systems fielded in order to obtain these capabilities will not be monolithic structures in terms of either operations or acquisition. Rather, they will be systems of systems, coalitions of systems, or federations of systems that are integrated in accordance with appropriate architectural constructs in order to achieve the evolutionary, adaptive, and emergent cooperative effects that will be required to achieve human and organizational purposes and to take advantage of rapid changes in technology. They can potentially accommodate system life-cycle change, in which the life cycle associated with use of a system family evolves over time; system purpose change, in which the focus in use of the system emerges and evolves over time; and environment change, in terms of alterations in the external context supporting differing organizational and human information and knowledge needs, as well as in the technological products that comprise constituent systems.

Autonomy, heterogeneity, and dispersion. One can contrast system families and identify relationships between conventional systems, systems of systems, and federations or coalitions of systems with regard to three characteristics, as initially noted by A. J. Krygiel: autonomy, heterogeneity, and dispersion. A federation of systems, federated system of systems, or coalition of systems will generally have greater values of these characteristics than a (nonfederated) system of systems.

Many conventional systems are built for special purposes, as a mixture of commercial-off-the-shelf systems and custom developments of hardware and software. These constituents are generally provided by multiple contractors who are used to supporting a specific customer base and working under the leadership of a single vertical program management structure. For best operation, these systems should be managed as a system of systems, federation of systems, or coalition of systems.

A system of systems generally has achieved integration of the constituent systems across communities of contractors, and sometimes across multiple customer bases, and is generally managed by more horizontally organized program management structures, such as integrated product and process development (IPPD) teams. When the IPPD team effort is well coordinated, the team is generally well able to deal with conflict issues that arise due to business, political, and other potentially competing interests.

Federations of systems face the same dilemmas identified for systems of systems but are generally much more heterogeneous along transcultural and transnational socio-political dimensions, are often managed in an autonomous manner without great central authority and direction such that they satisfy the objectives and purpose of an individual unit in the federation, and often accommodate a much greater geographic dispersion of organizational units and systems. Thus, the delimiters between systems of systems and federations of systems or coalitions of systems, while generally subjective, are nonetheless principled.

The notions of autonomy, heterogeneity, and dispersion are not independent of one another. Increasing geographic dispersion will usually lead to greater autonomy and consequently will also increase heterogeneity. The Internet is perhaps the best example of a system that began under the aegis of a single sponsor, the U.S. Department of Defense, and has grown to become a federation of systems.

Federalism. This approach, initially studied by C. Handy as "new federalism," addresses the necessary considerations for the structuring of loosely coupled organizations in order to help them adapt to changes in the information and knowledge age. The application of federalist political principles to the management of systems of systems and federations of systems is an appealing way of obtaining a systems engineering ecology—in other words, a sustainable systems engineering approach that possesses adaptation, evolution, and emergence characteristics analogous to those in a natural ecology. This is so because many contemporary organizational alliances, including those for engineering and development, take the form of virtual organizations or virtual teams rather than the classical physical organizations and teams. The concept of federalism is particularly appropriate since it offers a well-recognized way to deal with the systems engineering management paradoxes of power and control such that the desired systems ecological balance is obtained. Generally, this is accomplished by making things big by keeping them small, a goal which, in turn, is accomplished by expanding the domain of an enterprise by instantiating multiple quasiautonomous units as opposed to acquiring mass by aggregation around a centralized command-and-control authority base; encouraging autonomy but within appropriate bounds set by process and architecture standards; and combining variety with shared purpose, and individuality with partnerships at national and global levels.

Federalism is based on five principles: subsidiarity, interdependence, a uniform and standardized way of doing business, separation of powers, and dual citizenship. Subsidiarity, the most important principle, suggests that power belongs to the lowest possible point within the engineering team of a federation of systems. Interdependence, or pluralism, requires that the autonomous development units or teams of a development federation of a federation of systems stick together because they need one another as much as they need management leadership and leadership authority. Having a uniform and

standardized way of doing business ensures interdependence within federated engineering organizations of a system of systems or federation of systems through agreement on basic rules of conduct, common traditions of communicating, and common units of measurement of progress and quality. Separation of powers requires that management, monitoring, and governing aspects of engineering programs and projects of federations of systems be viewed as separate functions to be accomplished by separate bodies whose membership may overlap. Finally, dual citizenship requires that every individual is a "citizen" in two communities: the local development group, professional group, or union, and the overall program of the federation of systems.

Systems engineering approaches. These systems-of-systems and federations-of-systems concepts have numerous implications for systems engineering and management.

Grand design approach. Contemporary organizations often treat the engineering of systems of systems or federations of systems with systems engineering protocols that are, at best, suitable only for monolithic systems. The archetype of such ill-advised protocols is the "grand design" life cycle, which is based on the waterfall model that came into prominence around 1970. A large number of problems have been encountered with grand design efforts to engineer a system. Today, the classic waterfall approach is suggested only in those rare cases where user and system-level requirements are crystal clear and unlikely to change at all during or after engineering the system, and where funding for the grand design is essentially guaranteed. This is rarely the case for major systems, especially those that are software-intensive, and would be the rarest of all cases for a system of systems or federation of systems. Changing user and organizational needs and changing technologies virtually guarantee that major systems cannot be developed using the grand design approach.

Incremental and evolutionary approaches. Two leading alternatives to the grand design approach for the engineering of systems were initially termed incremental and evolutionary, although the term "evolutionary" is now generally used to characterize both of these. In incremental development, the system is delivered in preplanned phases or increments, in which each delivered module is functionally useful. The overall system capability improves with the addition of successive modules. The desired system capability is planned to change from the beginning, as the result of "build N" being augmented and enhanced through the phased increment of "build N+1." This approach enables a well-functioning implementation to be delivered and fielded within a relatively short time and augmented through additional builds. It also allows time for system users to thoroughly implement and evaluate an initial system with limited functionality compared to the ultimately desired system. Generally, the notion of preplanning of future builds is strong in incremental development. As experience with the system at build N is gained, requirements changes for module N+1 may be more easily incorporated into this, and subsequent, builds.

Evolutionary life-cycle development is similar in approach to its incremental complement; however, future changes are not necessarily preplanned. This approach recognizes that it is impossible to initially predict and set forth engineering plans for the exact nature of these changes. The system is engineered at build N+1 through reengineering the system that existed at build N. Thus, a new functional system is delivered at each build, rather than obtaining build N+1 from build N by adding a new module. The enhancements to be made to obtain a future system are not determined in advance, as in the case of incremental builds. Evolutionary development approaches can be very effective in cases where user requirements are expected to shift dramatically over time, and where emerging and innovative technologies allow for major future improvements. They are especially useful for the engineering of unprecedented systems that involve substantial risk and allow potentially enhanced risk management. Evolutionary development may help program managers adjust to changing requirements and funding priority shifts over time, since new functionality introductions can be advanced or delayed in order to accommodate user requirements and funding changes. Open, flexible, and adaptable system architecture is central to the notion of evolutionary and emergent development. These are major elements in the contemporary U.S. Department of Defense Initiatives in evolutionary acquisition. *See* DISTRIBUTED SYSTEMS (CONTROL SYSTEMS); LARGE SYSTEMS CONTROL THEORY; REENGINEERING; RISK ASSESSMENT AND MANAGEMENT; SYSTEMS ARCHITECTURE; SYSTEMS ENGINEERING; SYSTEMS INTEGRATION.

Andrew P. Sage

Bibliography. P. G. Carlock and R. E. Fenton, Systems of systems (SoS) enterprise systems engineering for information-intensive organizations, *Sys. Eng.*, 4(4):242–261, 2001; P. Chen and J. Clothier, Advancing systems engineering for systems-of-systems challenges, *Sys. Eng.*, 6(3):170–183, 2003; C. Handy, Balancing corporate power: A new Federalist Paper, *Harvard Bus. Rev.*, 70(6):59–72, November-December 1992; C. Handy, Trust and the virtual organization, *Harvard Bus. Rev.*, 73(3):8–15, May/June 1995; A. J. Krygiel, *Behind the Wizard's Curtain: An Integration Environment for a System of Systems*, CCRP Publication Series, Vienna, VA, 1999; M. W. Maier, Architecting principles for systems-of-systems, *Sys. Eng.*, 1(4):267–284, 1998; J. Morganwalp and A. P. Sage, A system of systems focused enterprise architecture framework and an associated architecture development process, *Inform. Knowl. Sys. Manag.*, 3(4):87–105, 2003; A. P. Sage and C. D. Cuppan, On the systems engineering and management of systems of systems and federations of systems, *Inform. Knowl. Sys. Manag.*, 2(4):325–345, 2001; A. P. Sage and W. B. Rouse (eds.), *Handbook of Systems Engineering and Management*, Wiley, New York, 1999.

Systematics

The comparative analysis of living and fossil species, including their discovery, description, evolutionary relationships to other species, and patterns of geographic distribution. Because systematics involves crucial activities of describing and understanding the history of life, it is the primary science of biological diversity.

Systematics can be divided into four major fields. Taxonomy, often equated with systematics, is the discipline concerned with the discovery, description, and classification of organism groups, termed taxa (singular, taxon). Classification is the clustering of species into a hierarchical arrangement according to some criterion, usually an understanding of their relationships to other species. Phylogenetic analysis, an increasingly important aspect of systematics, is the discovery of the historical, evolutionary relationships among species; this pattern of relationships is termed a phylogeny. The fourth component of systematics is biogeography, the study of species' geographic distributions. Historical biogeography examines how species' distributions have changed over time in relationship to the history of landforms, ocean basins, and climate, as well as how those changes have contributed to the evolution of biotas (groups of species living together in communities and ecosystems).

Taxonomy. The habitats and ecosystems of the natural world are extremely complex in terms of their ecological and taxonomic diversity. From those environments, systematists have discovered and described approximately 1.4 million species, of which about 950,000 are insects, 250,000 are plants, and 45,000 are vertebrates. Estimates of the number of species left to be discovered range from 3 million to nearly 100 million. A major reason for this uncertainty involves the difficulty of sampling over the entire surface of the Earth many different kinds of organisms, especially very small organisms. Documenting biodiversity is among the most challenging scientific problems. For example, an enormous effort is needed to describe and classify these species, one that far outstrips available taxonomic expertise. Moreover, the lack of a global electronic information system that can store information about those species which have already been described as well as those that will be discovered in the future also hinders progress.

A fundamental problem underlying all systematic inquiry, and a major concern for taxonomic practice, is the nature of species. Debates about how species should be defined and how they might be recognized are among the most long-standing and intense in all of biology. Many authorities have proposed solutions to the species problem and offered a variety of definitions, but basically two primary criteria have been used to define species. By far the most commonly applied criterion is diagnosability: one or more populations that are recognizably distinct from other such populations and that cannot be subdivided into other diagnosable units constitute a species. The notion of diagnosability means that once comparisons have been made to establish patterns of variability within and among different populations, the systematist finds that some populations can be characterized as distinct because the organisms share one or more characters not found in other populations. These characters may be morphological, behavioral, physiological, or biochemical. This view of species has been the one most commonly used by systematists, especially those studying insects and other small organisms, which make up the majority of the world's species.

A second view of species, adopted frequently within vertebrate systematics, is the biological species concept: species are those populations or groups of populations that are reproductively isolated from other such populations. Thus, if an ancestral population is subdivided by some extrinsic ecological or geological barrier into two geographically isolated populations, given sufficient time one or both will likely differentiate, perhaps to the extent that if the barriers were to disappear and the populations came back into contact they would be incapable of interbreeding or of producing fertile offspring. This scenario is the classic allopatric speciation model, which is often closely linked to the biological species concept. *See* SPECIES CONCEPT.

In principle, application of the biological species concept requires that the putative species geographically overlap (that is, be sympatric with) its close relatives so that reproductive isolation can be assessed. In nature, however, many populations remain geographically separated (allopatric) from one another, and in these situations followers of the biological species concept attempt to infer the presence of reproductive isolation by evaluating the degree of morphological, behavioral, or ecological differences among the populations. The more differentiated the populations, it is thought, the greater the likelihood that they will not be capable of interbreeding upon becoming sympatric.

It is important to understand that species designations are hypotheses regardless of which definition might be used. In contrast to individual organisms, a species cannot be directly observed, and therefore new information may lead a systematist to reevaluate species limits.

For over 200 years naturalists have sought to document Earth's species diversity by mounting expeditions. These inventories have helped build the major natural history collections that contain worldwide some 2 billion specimens. Once specimens are collected and brought back to the museum or herbarium for study, they are cataloged and identified as to species. If they cannot be assigned to any known species, they are described as new and a type specimen is designated. The type serves as the reference specimen for the new name, and future workers will compare the type with newly collected specimens to see whether any of those are new. This comparison leads to revisionary monographs, which summarize all the species of a group, their characters, distributions, and relationships. Monographs

constitute a major vehicle for studies of phylogenetic relationships and for a summary classification. *See* TAXONOMY.

Phylogeny. Organisms have been evolving and diversifying on Earth for several billion years, and a major task of systematics is the generation of hypotheses about the historical pattern of this evolution, that is, about phylogeny. Determining phylogenetic relationships is essentially a discovery process in which comparisons are made among species, similarities and differences are assessed, and a hierarchical pattern of relationships is hypothesized based on these patterns of similarity.

For many years there were debates over how similarities were to be interpreted during the process of phylogenetic inference. One group of systematists advocated an approach called phenetics, in which overall similarity was quantified and relationships were then based on those measures. Another group proposed a method termed cladistics, in which observed similarities are partitioned into primitive or derived similarity. Only derived similarities—that is, characters modified from more primitive conditions—could then be used as evidence of relationships. For example, vertebrates commonly called reptiles were placed in their own formal classificatory group, Reptilia. Once comparisons among reptiles and other vertebrate groups were made, however, it was discovered that the characters of the Reptilia (such as scales, absence of hair or feathers) were primitive and that other, derived characters linked some reptiles with mammals and some with birds. Thus, theropod dinosaurs are known to be more closely related to birds than to many other so-called reptiles.

Phylogenetic analysis is of major importance to biology because an understanding of relationships provides the basis for making powerful predictions about the characters of organisms that have yet to be studied. Phylogenetic relationships, because they establish a historical hierarchy, constitute the most logical framework for synthesizing all that is known about species and groups of species. The picture of animal phylogeny is very incomplete, but the use of new characters that are revealed by technological advances such as electron microscopy and deoxyribonucleic acid (DNA) sequence contribute to the solution of many phylogenetic puzzles. *See* PHYLOGENY.

Classification. Biological classification is the grouping together of species on the basis of one or more criteria and then naming those groups. Present-day classification schemes are founded on the work of the eighteenth-century Swedish botanist, Carolus Linnaeus. He proposed a hierarchical series of taxonomic categories, such as kingdom, phylum, class, order, family, genus, and species. Thus, the tiger has the generic and specific name *Panthera tigris*, is a member of the family Felidae, which is in the order Carnivora, and carnivores and other mammals are in the class Mammalia.

Debates over which criteria are most appropriate for establishing hierarchical relationships have been closely tied to debates over methods of assessing phylogenetic relationships. Again, some have advocated constructing the hierarchy on the relative degree of overall similarity (phenetics), but the importance of basing classifications on the pattern of phylogenetic relationships is clearly recognized.

The Linnaean scheme of subordinate categorical ranking is intrinsically hierarchical, and so too is the historical pattern of life's evolution. To the extent that a classification scheme mirrors the hierarchical pattern of the best-corroborated hypotheses about group relationships, the groups of the classification can be said to be natural. Constructing classifications that reflect the genealogical relationships of taxa is critical for making the findings of systematics maximally useful within basic and applied biology. Natural classifications are highly predictive. For example, a newly discovered species of cat belonging to the genus *Felis* would be predicted to have characters in common with other species of the genus, some that are common with the other species in the family Felidae, and so on to include the higher taxonomic levels of Carnivora and Mammalia. *See* CLASSIFICATION, BIOLOGICAL.

Biogeography. A major component of systematic research involves documenting the distributions of species—both contemporaneous and fossil—and understanding why groups of species have such distributions and how the distributions have changed over time. The distributions of most species are very imperfectly known, mainly because they have been determined from a small number of specimens collected from a few sites. Even in well-known, highly visible groups such as birds, distributional limits are often uncertain. This uncertainty underscores the importance of an extensive inventory effort around the globe. Information about distributions is crucial for many aspects of basic and applied biology, from understanding how and why groups diversify over space and time, to identifying the geographic sources of agricultural pests, diseases, or other introduced exotics.

Distribution patterns of species can be explained by the interaction of a number of processes that take place over different spatiotemporal scales. Long-distance dispersal involves the idiosyncratic movement of relatively few individuals across a geographic or ecologic barrier; population dispersion involves the movement of many individuals of a population into new areas that are ecologically compatible; and vicariance is a process whereby a widespread population is divided into smaller components by the origin of a geographic or ecological barrier. *See* POPULATION DISPERSAL; POPULATION DISPERSION.

Prior to the application of phylogenetic methods in systematics, long-distance dispersal was often considered to be the primary mechanism explaining species distributions. This view grew out of a belief that continents, islands, and ocean basins were stable over time. With the realization that continents had shifted their positions, systematists began invoking dispersion of many groups from one region to another over land connections in order to

explain similarities in biotas. It also became clear to workers attempting to explain disjunct distributions among closely related groups that these geographic patterns could be explained by vicariance of a once widespread ancestral population. This line of thinking has led to the discipline of vicariance biogeography, which entails a search for congruent historical patterns in distributions across different groups of organisms and then seeks to explain that congruence by postulating a common pattern of vicariance.

Evidence suggests that the geographic history of species and biotas, whether terrestrial or marine, is intimately tied to the history of the areas in which they reside. When changes in climate or Earth history reduce or eliminate geographic or ecological barriers, many species of a biota expand their ranges together, and thus the biota becomes more cosmopolitan. The subsequent origin of barriers partitions this widespread biota into smaller, isolated areas within which populations differentiate to create areas of endemism. *See* BIOGEOGRAPHY.

Importance of systematics. Humans use tens of thousands of species in their daily lives for food, medicine, clothing, shelter, and many other aspects of commerce. At the same time, more than 40,000 mi^2 (100,000 km^2) of habitat are being lost each year. In order to conserve biodiversity and use what remains in a sustainable manner, resource managers must have access to reliable scientific information about species. Included in this information is knowledge about species systematics.

Systematists provide answers to four fundamental questions that are required by resource managers and users of biodiversity: What species is it? What are its characteristics? What are its relationships to other species? Where is it distributed? Correct identifications may be life-and-death issues or may have major economic implications: for example, distinguishing between pathogenic and nonpathogenic organisms, or between an indigenous, benign species and an introduced pest. Likewise, knowing the characteristics of organisms is fundamental for developing new genetic resources or designing a biocontrol program for an exotic agricultural pest. Predictive classifications, made possible by an understanding of relationships, help direct the search for new genetic resources to improve agricultural productivity, or lead to the discovery of new pharmacologically active compounds in close relatives of medicinally useful plants.

Finally, systematic data and interpretations underlie progress in all of biology. An understanding of relationships, in particular, is fundamental for interpreting comparative data across different kinds of organisms, whether those data be morphological, physiological, or biochemical. Joel Cracraft

Bibliography. D. R. Brooks and D. A. McLennan, *Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology*, 1991; N. Eldredge and J. Cracraft, *Phylogenetic Patterns and the Evolutionary Process*, 1980; P. L. Forey et al., *Cladistics: A Practical Course in Systematics*, 1992; G. J. Nelson and N. I. Platnick, *Systematics and Biogeography: Cladistics and Vicariance*, 1981; E. O. Wiley, *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*, 1981.

Systems analysis

The application of mathematical methods to the study of complex human physical systems. A system is an arrangement or collection of objects that operate together for a common purpose. The objects may include machines (mechanical, electronic, or robotic), humans (individuals, organizations, or societal groups), and physical and biological entities. Everything excluded from a system is considered to be part of the system's environment. A system functions within its environment. Examples of systems include the solar system, a regional ecosystem, a nation's highway system, a corporation's production system, an area's hospital system, and a missile's guidance system. A system is analyzed so as to better understand the relationships and interactions between the objects that compose it and, where possible, to develop and test strategies for managing the system and for improving its outcomes.

Systems of physical processes, such as the weather system, an oil refinery plant, or an energy generation system, are often studied under the methodological headings of systems theory, systems science, or systems engineering. The procedures that constitute these methodologies have contributed greatly to the ability to organize and control very large and complex physical systems, with space exploration via crewless satellites being a prime example. These successes have caused investigators to attempt to apply systems engineering concepts to a wide range of systems that are characterized by having a very strong human component. In particular, a major thrust in systems research is directed toward the study of problems that arise in public policy (at all levels of government), and business and industrial decision making. The term "systems analysis" is reserved for the study of systems that include the human element and behavioral relationships between the system's human element and its physical and mechanical components, if any. Examples of public policy systems are the federal government's welfare system, a state's criminal justice system, a county's educational system, a city's public safety system, and an area's waste management system. Examples of industrial systems are a manufacturer's production distribution system and an oil company's exploration, production, refining, and marketing system. Examples with physical environmental components are the atmospheric system (in which the concerns are with air pollution and the greenhouse effect) and a water supply system (in which water quality and quantity are the attributes of concern). The direct transfer of systems engineering concepts to the study of a system in which the human element must be considered is restricted by limitations in the ability to comprehend and quantify human interactions. (Operations research, a related field of study, is directed toward the analysis of

components of such systems. Public policy analysis is the term used for a system study of a governmental problem area.) *See* DECISION THEORY; OPERATIONS RESEARCH; SYSTEMS ENGINEERING.

Process. Systems comprise interrelated objects, with the objects having a number of measurable attributes. A mathematical model of a system attempts to quantify the attributes and to relate the objects mathematically. The resultant model can then be used to study how the real-world system would behave as initial conditions, attribute values, and relationships are varied systematically. Thus, the mathematical model serves as a controlled means of studying the system in a descriptive sense. Most importantly, if the purpose of the systems analysis is to aid in the design of a new system or the modification of an ongoing system, the mathematical model serves as an experimental tool that can prescribe a particular alternative solution. Here it is assumed that some criterion—an objective (for example, cost) or other measure of effectiveness (for example, police response time)—can be stated that enables the analysis to differentiate between possible alternative designs. The process is often confounded by the fact that there are usually many criteria that impact the final decision. To a limited extent, multicriteria methods are available to address such situations. *See* MODEL THEORY.

The systems analysis process is an iterative one that cycles repeatedly through the following interrelated and somewhat indistinct phases: (1) problem statement, in which the system is defined in terms of its environment, goals, objectives, constraints, criteria, actors (decision makers, participants in the system, impacted constituency), and other objects and their attributes; (2) alternative designs, in which solutions are identified; (3) mathematical formulation, in which a mathematical description of the system is developed, tested, and validated; (4) evaluation of alternatives, in which the mathematical model is used to evaluate and rank the possible alternative designs by means of the criteria; and (5) selection and implementation of the most preferred solution. The process includes feedback loops in which the outcomes of each phase are reconsidered based on the analyses and outcomes of the other phases. For example, during the implementation phase, constraints may be uncovered that hinder the solution's implementation and thus cause the mathematical model to be reformulated. The analysis process continues until there is evidence that the mathematical structure is suitable; that is, it has enough validity to yield answers that are of value to the system designers or the decision maker.

Mathematical methods and examples. The mathematical foundations of systems analysis derive from engineering and its use of differential equations as applied in optimal control theory; from computer science and its theory of automata; and from applied mathematics (especially operations research), which includes optimization, statistical analysis, simulation, and the calculus of variations. The analytical procedures used in systems analysis encompass the full range of scientific and computational methodology. Systems analyses have been applied in such diverse scientific fields as biology, engineering, transportation, ecology, and energy, and in social areas such as welfare, housing, and health. *See* AUTOMATA THEORY; CALCULUS OF VARIATIONS; DIFFERENTIAL EQUATION; OPTIMIZATION; SIMULATION; STATISTICS.

Typical systems analyses and their mathematical bases include the following: (1) Energy flow in a marine ecosystem was modeled as a set of differential equations under steady-state assumptions. (2) A regional blood bank improved its operations after a study based on stochastic analyses of its supply and demand, and creation of a mathematical optimization model of its distributional activities. (3) The Georges Bank haddock fishery system used a combination of biological fish population models and econometric market models to study the financial operation of a fishing system over time. (4) The future structure of the United States energy system was studied with the aid of an integrated econometric linear-programming-based model of the total United States energy production and distribution system. (5) The economic management of an Alaskan salmon fishery was analyzed by a computer-based simulation. (6) Proposed changes to the U.S. Department of Agriculture's food stamp program were analyzed by a model based upon microanalytic simulation in which the actions of the population in general were studied by a representative sample of households. (7) The appropriate number and location of firehouses in Wilmington, Delaware, were determined by using travel-time analyses and optimization-based location models. (8) The entire Dutch water management system was analyzed to provide a basis for a new national water management policy. *See* LINEAR PROGRAMMING; STOCHASTIC PROCESS; SYSTEMS ECOLOGY.

Status. The form and substance of systems analysis have been codified and applied with much success in a number of important areas. The technique is, on the whole, an accepted one, but there are concerns about its application and its ability to serve the needs of society. The incorrect application of systems analysis as originally developed and practiced in the Department of Defense led to some critical and costly failures. This was due mainly to the fact that the engineering basis of systems analysis was applied to social systems, in which the relationships (political and bureaucratic) between the objects (people) were not well understood and the values of the parameters were not accurately known. These difficulties still exist, but current systems analyses now take them into account. As systems analysis has evolved, its practitioners have learned to differentiate between different types of systems.

Hard and soft systems. As originally developed, systems analysis studies have been applied to those areas that are "hard" in that they are well defined and well structured in terms of objectives and feasible alternative systems (for example, blood-bank design, and integrated production and inventory processes). The aim of hard systems analysis is to select

the best feasible alternative. In contrast, soft systems are concerned with problem areas that involve ill-defined and unstructured situations, especially those that have strong political, social, and human components. These generally involve public and private organizations (for example, design of a welfare system, and structure and impact of a corporate mission statement). The objectives of soft systems and the means to accomplish them are problematical and, in fact, a systemic view of the problem area is not assumed. The aim of soft systems analysis is to find a plan of action that accommodates the different interests of its human actors.

Large-scale systems. There is also need for further study of large-scale systems, which by definition are most complex. It is important to find ways to describe mathematically the systems that represent the totality of an industrial organization, the pollution concerns of a country and a continent, or the worldwide agricultural system. These are multicriteria problems with the solutions conflicting across criteria, individuals, and countries. The possibility that such systems may be studied in a computer-based laboratory is very promising. But this challenge must be approached cautiously, with the awareness that the methods and models employed are only abstractions to be used with due consideration of the goals of the individual and society. *See* LARGE SYSTEMS CONTROL THEORY; LINEAR SYSTEM ANALYSIS. Saul I. Gass

Bibliography. P. Checkland, *Systems Thinking, Systems Practice*, 1981; P. Checkland and J. Scholes, *Soft Systems Methodology in Action*, 1981; C. W. Churchman, *The Design of Inquiring Systems*, 1971; C. W. Churchman, *The Systems Approach*, 1968; P. Edwards, *Systems Analysis and Design*, 1993; K. E. Kendall and J. E. Kendall, *Systems Analysis and Design*, 3d ed., 1994; I. R. Hoos, *Systems Analysis in Public Policy*, 1972; M. D. Mesarovic and Y. Takahara, *Abstract Systems Theory*, 1989; H. J. Miser and E. S. Quade (eds.), *Handbook of Systems Analysis*, vol. 1: *Overview of Uses*, *Procedures*, *Applications*, *and Practice*, 1995, vol. 2: *Craft Issues and Procedural Choices*, *and Practices*, 1988, vol. 3: *Cases*, 1996; J. Moder and S. E. Elmaghraby, *Handbook of Operations Research*, 1978; E. S. Quade, *Analysis for Public Decisions*, 3d ed., 1989, L. A. Zadeh and C. A. Desoer, *Linear Systems Theory*, 1963, reprint 1979.

Systems architecture

The discipline that combines system elements which, working together, create unique structural and behavioral capabilities that none could produce alone. The degree to which well-designed systems-level architectures are critical to the success of large-scale projects—or the lack thereof to failure—has been dramatically demonstrated in recent years. The explosion of technological opportunities and customer demands has driven up the size, complexity, costs, and investment risks of such projects to levels feasible for only major companies and governments. Without sound systems architectures, these projects lack the firm foundation and robust structure on which to build. This conclusion holds whether the system is a satellite surveillance system, a crewed spacecraft, a stealth aircraft, a commercial airliner, a line of personal computers, or an advanced microprocessor.

Growth of large systems. One of the earliest major system developments began in the early 1900s with the change from 100-customer telephone companies to the Bell System, a million times larger and capable of providing quality services. Air service progressed from airplanes that could barely carry a few mailbags to DC-3's, the global jet air transportation system with its vast array of local and international airlines and airports. Simple 200-lb (100-kg) satellites were replaced by multiton spacecraft exploring the entire solar system, providing global communications and navigation, and helping to keep world peace through surveillance. The combination of global transportation and global communication forced regional markets to become global in less than a decade. Software programs, supported by 100,000-element microprocessors and million-line codes, revolutionized many fields of endeavor. Crewed space programs developed into major national enterprises. Nuclear power systems became major contributors to energy production in many countries. Programs of this size and cost rapidly became matters of social and political debate, necessarily complicating system design, engineering management, production planning, and operational constraints. *See* SYSTEMS ENGINEERING.

Concept of an architecture. In the 1950s, the technical literature began reflecting the importance of the architecture of communication, computer, aerospace, and industrial systems. The word "architecture" is commonly used to describe the underlying structure of networks, command-and-control systems, spacecraft, and computer hardware and software [By implication, if an architecture is not strong enough, it and the system it supports will collapse.] *See* COMPUTER ARCHITECTURE.

It was a small step to recognize the philosophical similarities between these architectures and those begun millennia ago and carried forth by successive civilizations. The field of classical architecture provides principles, insights, and approaches of proven value for the more recent engineering disciplines. Not the least of these is the concept of an architecture. Only the context has been changed.

Complexity and its consequences. A recent addition to these historical principles is the concept of systems, first propounded for feedback control applications in the late 1940s. Systems are collections of dissimilar elements which collectively produce results not achievable by the elements separately. Their added value comes from the relationships or interfaces among the elements. (For example, open-loop and closed-loop architectures perform very differently.) But this value comes at a price: a complexity potentially too great to be handled by standard rules or rational analysis alone. Complexity

comes from many sources: uncertainties in technological change; geopolitical upheaval; environmental concerns; public perceptions of cost and risk; an opponent's weapons of attack or defense; and uncertainties in funding and political support. Similar complexities had been treated by earlier architects, but the principles of systems extended the scope to more disciplines and technologies. *See* CONTROL SYSTEMS.

As projects became ever more complex and multidisciplinary, new structures were needed for projects to succeed. It was no longer sufficient simply to use known elements, even in new ways. A skyscraper is not built of wood and stone. A stealth aircraft is not a converted airliner. Distributed workstations are not sliced-up mainframes. Object-oriented software architectures are neither modified communications protocols nor table-driven interpreters.

Nor could analytic techniques be used to find optimal solutions. Indeed, given the disparate perspectives of different customers, suppliers, and government agencies, unique optimal solutions generally would not exist. Instead, many possibilities might be good enough, with the choice dependent more on ancillary constraints or on the criteria for success than on detailed analysis. Nor could subsystems be locally optimized; the system as a whole might suffer.

Conceptual phases. As increasingly complex systems were built and used, it became clear that success or failure had been determined very early in their projects. The industry adage that "all the serious mistakes (in software) are made in the first day" may be an overstatement, but in the early phases all the critical assumptions, constraints, choices, and priorities are made that will determine the end result. Unfortunately, no one knows in the beginning just what the final performance, cost, and schedule will be. But the decisions which will determine them will have been made. The later phases, assuming they are carried out well, contribute much less.

Systems-level architecture. It is no coincidence that the systems-level architecture is laid out in the early phases. Systems-level architecture specifies how system-level functions and requirements are gathered together in related groups. It indicates how the subsystems are partitioned, what the relationships between the subsystems are, what communication exists between the subsystems, and what parameters are critical. It makes possible the setting of specifications, the analysis of alternatives at the subsystem level, the beginnings of detailed cost modeling, and the outlines of a procurement strategy.

Design with incomplete information. Information critical for rational architectural decision making is seldom available at the systems-level architecture stage. The practice of systems architecture (architecting), then, must be more an art than a science, more dependent on insights and intuition than on quantitative data.

It is simply unrealistic to postulate that complex systems designs begin with iron-clad requirements which flow down to specifications, and hence to an optimum system. For example, there rarely is enough information early in the design stage for the client to decide on the relative priority of the requirements without having some idea of what the end system might be. Instead, provisional requirements and alternative system concepts have to be iterated until a satisfactory match is produced.

Joint participation. Unavoidably, successful systems architecting in the conceptual phase becomes a joint process in which both client and architect participate heavily. In the ideal situation, the client makes the value judgments (What is good enough? What is affordable? When is it needed?), and the architect makes the technical decisions (What can be built? How can it be done? How long will it take?). Failure of either party to make timely decisions has all the subsequent deleterious consequences that might be expected: performance deficiencies, cost and schedule overruns, emergency redesigns, software patches and restarts, disappointed sponsors and users, recriminations and even lawsuits.

Conceptual model. Systems-level architecture begins with a conceptual model, a top-level abstraction which attempts to discard features deemed not essential at the system level (for example, the color of paint on an aircraft). Physically, the model can be a small-scale display built out of balsa and paper, a block diagram of an advanced radar, a computer simulation of a spacecraft, or a bubble diagram of a software program. Such a model is an essential tool of communication between client, architect, and builder, each viewing it from a different perspective. As the system comes into being, the model is progressively refined. Secondary features are brought into play; subsystems are fleshed out; subsystem components critical to the system as a whole are identified and tracked (for example, an untried technology).

Acceptance criteria. As development nears completion, a hard truth becomes apparent. What is actually being built is determined by the acceptance criteria—not the desired functions, not the conceptual model, but what will pass the specified acceptance tests.

The builder's interest is in being paid for delivery of a system that has passed its final test, no more and no less. The buyer's interest is in receiving a system as originally conceived. The key to resolving this conflict is the early incorporation of acceptance criteria into the conceptual model of the systems-level architecture.

This failure to foresee and resolve such situations in the conceptual stage can easily lead to severe disappointment, recriminations, financial loss to the builder, delays in operation for the buyer, and costly redesign. Systems that have not been partitioned to minimize communications between the subsystems will fail when first assembled, seriously complicating system-level testing and diagnosis. Untestable or unpassable acceptance criteria will provoke argument and dissension. Systems that have not been designed to be tested will prove untestable. Systems that do

not have designed-in quality will face difficult trials as efforts are made to test them. System designs that do not solve the users' real problems will be rejected.

Recent developments. The most important addition to the field of systems architecture since 1990 is software systems architecture. Not only does it contribute greatly to the theory and practice of architecting, but it comes right at the time when "smart," or software-intensive, systems are displacing simpler, hardware-only predecessors. Smart systems are far more capable, less expensive to produce, and faster to reach the marketplace, and consequently are much more profitable than their predecessors. They are also critically dependent not only on their own software architectures but on the architecture of the systems they support. *See* SOFTWARE ENGINEERING.

Eberhardt Rechtin

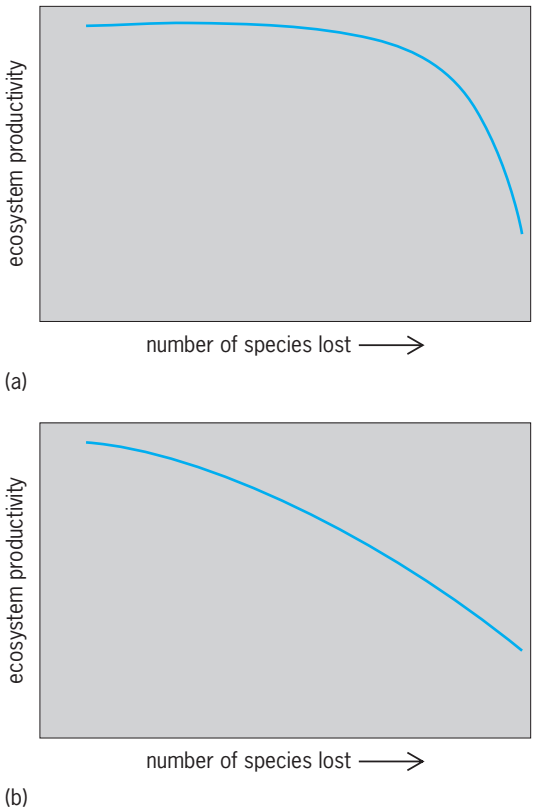
Bibliography. M. W. Maier and E. Rechtin, *The Art of System Architecting*, 2d ed., CRC Press, 2000; G. Nadler, *The Planning and Design Approach*, John Wiley, 1981; E. Rechtin, *Systems Architecting: Creating and Building Complex Systems*, Prentice Hall, 1991; E. Rechtin, *Systems Architecting of Organizations: Why Eagles Can't Fly*, CRC Press, 2000; M. Shaw and D. Garlan, *Software Architecture: Perspectives on an Emerging Discipline*, Prentice Hall, 1996.

Systems ecology

The analysis of how ecosystem function is determined by the components of an ecosystem and how those components cycle, retain, or exchange energy and nutrients. Systems ecology typically involves the application of computer models that track the flow of energy and materials and predict the responses of systems to perturbations that range from fires to climate change to species extinctions. Systems ecology is closely related to mathematical ecology, with the major difference stemming from systems ecology's focus on energy and nutrient flow and its borrowing of ideas from engineering. *See* ECOLOGY; MATHEMATICAL ECOLOGY.

Historical overview. Systems ecology emerged as a mathematical framework for posing questions about how ecosystems work, shortly after the concept of ecosystems gained momentum. R. H. Lindeman launched systems ecology with his pioneering idea of depicting ecological systems, such as a lake, as compartments of feeding groups (producers, herbivores, carnivores) or trophic levels through which energy flowed. As ecologists began to track the flow of energy, they also began to track the flow of nutrients and started asking questions about efficiency, stability, and production, all of which bring to mind theories from engineering and the industrial design of production systems. At first systems ecology was in conflict with evolutionary ecology because systems ecologists often discussed how systems evolved toward particular properties (more efficiency and more stability), whereas evolutionary biologists argued that there was no grand design of ecosystems and that evolution was the result of natural selection acting on individuals. Ecosystems do not evolve. Systems ecology no longer discusses the evolution of ecosystem attributes, but it still focuses on emergent properties such as stability and resilience. *See* ECOLOGICAL COMMUNITIES; ECOLOGICAL ENERGETICS; ECOSYSTEM.

Ecosystem theory. Ecosystem theory draws heavily on applied mathematics and engineering studies to ask questions about system performance. For example, when designing a manufacturing or control system, systems engineers have recognized the value of redundancy, so that if one component fails the entire system does not fail. This seemingly abstract idea becomes practical when investigating the value of protecting biodiversity. If it is believed that many species are redundant, then an engineering perspective would suggest that losing species might not produce any noticeable changes in ecosystem functioning over a short time scale. Only after too many redundant species were lost would a pronounced change in system performance be noticed (**illus.** *a*). Alternatively, if species were not redundant, then each loss of a species might be expected to decrement ecosystem performance, resulting in a linear decline in productivity or stability as the number of species declined (illus. *b*). The scientific



Decrease in system productivity with loss of species. (*a*) Under the assumption that most species are redundant. Initial losses do not matter because the remaining species are able to carry out the ecosystem function necessary for production. (*b*) Under the assumption that there are no redundant species. In this case, each species that is lost causes some decrement in system productivity.

challenge is to determine how ecosystems are constructed—are they clusters of functional groups, with many redundant species in each functional group, or are they collections of truly unique species, with each species performing a unique function? *See* BIODIVERSITY; SYSTEMS ENGINEERING; THEORETICAL ECOLOGY.

Large-scale studies and models. Systems ecology is an especially valuable approach for investigating systems so large and complicated that experiments are impossible, and even observations of the entire system are impractical. In these overwhelming settings, the only approach is to break down the research into measurements of components and then assemble a system model that pieces together all components. The world's largest ecosystem is the North Pacific Ocean gyre, which covers 20 million square kilometers of the Earth's surface. This huge ecosystem cannot be experimentally manipulated. It is even hard to measure aspects of this huge oceanic ecosystem. Nonetheless, by constructing food-web models that track the flow of energy and carbon from photosynthetic plankton to herbivores to higher trophic levels, systems ecologists have collaborated with oceanographers to discover an important pathway for carbon, which has implications for feedback loops between biological processes and global climate change. *See* FOOD WEB; PACIFIC OCEAN.

Specifically, it appears that there is a major downward flux of carbon into the bottom of the Pacific Ocean, where it becomes the basis of a bacterial community. The amount of carbon that is pumped downward is so substantial that variations in this pump could alter the amount of carbon dioxide in the atmosphere and, hence, the rate of global warming. The story is complicated because phytoplankton production, which "primes the pump," varies as a result of ocean circulation patterns and nutrient limitation. Thus there is a very complex system in which global climate change may alter patterns of ocean circulation and primary production, which in turn alters the carbon sink represented by the ocean, which then feeds back to influence the magnitude of carbon dioxide increases in the atmosphere and global climate change. These cycles, pumps, and feedback loops are not understood well enough to make firm predictions; however, it is clear that the tools of systems ecology are necessary for investigating such complex dynamics, in which no simple experiment or suite of observations is likely to produce unequivocal insight. *See* CARBON DIOXIDE; GLOBAL CLIMATE CHANGE; PHYTOPLANKTON.

Addressing environmental problems. An important contribution of ecosystem science is the recognition that there are critical ecosystem services such as cleansing of water, recycling of waste materials, production of food and fiber, and mitigation of pestilence and plagues. The United Nations and nonprofit organizations are combining to fund global assessments of these ecosystem services, with the idea of providing key information to policymakers regarding possible threats to environmental well-being. Systems ecology becomes an important tool in interpreting data about changing ecosystems and possible loss of ecosystem services, because it asks whether crossing certain thresholds can so fundamentally change a system that recovery is impossible. An example of such a threshold involves semiarid grazing systems in which the vegetation is composed of grasses and woody shrubs. Cattle eat grasses but not the woody plants. In the absence of cattle, many of these savannah systems are dominated by grasses with a scattering of shrubs. As the number of cattle on the land increases, the density of grass declines. At some point the cattle grazing pressure can be so severe that the grasses effectively disappear and the landscape becomes a barren shrubland. At that point, even if the cattle are taken off the land it may never recover, because the system has moved to a new equilibrium in which the shrubs are able to outcompete and prevent the grasses from returning. This system has been modeled, and the implications of this model for resource management are clear: overgrazing of certain systems may not be reversible and, hence, must be managed carefully to maintain the land's capacity for food production. *See* ECOLOGY, APPLIED.

Scaling laws. Variability changes as a function of the spatial scale and temporal scale over which a process is observed. The scaling of variability is important because variability drives ecological dynamics and influences human resource extraction. In addition, if scientists fail to appreciate the sensitivity of scale measurements, the results obtained can be seriously misleading. A result obtained at the scale of a square meter may disappear if measured at a scale of thousands of square kilometers. Spatial scale also commonly determines how tight the feedback loops are expected to be, with implications for environmental policy. For example, it is easy to get neighborhood groups to adopt environmentally rational zoning laws because the feedback loops between zoning laws and local greenspace, parks, and quality of life are tight. However, it is much harder to get the public excited about modifying their individual patterns of carbon consumption in ways that would reduce global warming, because the feedback loop between local behavior and total global carbon budget is very loose; the translation from local action to global effects is almost invisible when it comes to carbon fluxes.

Linking individual ecology, population ecology, and ecosystem dynamics. Systems ecology is one of the few theoretical tools that can simultaneously examine a system from the level of individuals all the way up to the level of ecosystem dynamics. An excellent example of this type of system analysis entails a detailed model of Florida's Big Cypress Natural Preserve, in which the trophic exchanges (measured as grams of carbon per square meter per year) among 68 principal taxa are inserted in a network model depicting all feeding relationships. The taxa in this cypress swamp ranged from phytoplankton to vines, cypress trees, snails, frogs and salamanders, woodpeckers, vultures, and all the way up the food chain to the Florida panther and American alligator. Without a model, the first guess would be that

any increase in a predator population would naturally decrease the population of its prey. However, by quantifying the details of carbon exchange among all of these interacting components, biologists have found that in some cases an increase in a predator population actually also increases its prey population.

A notable example of the positive effects of an increase in predators on prey populations involves the alligator and frogs and salamanders in the Cypress swamp. Alligators certainly eat large numbers of frogs and salamanders, but they also eat snakes, which are especially voracious predators of frogs and salamanders. The net result is that the more alligators in a region, the fewer snakes and the more frogs and salamanders. Software packages exist that can determine when these sorts of counterintuitive results are likely to emerge from complicated networks of interacting species. By using such a network system model, it is possible to predict the consequences of removing any species in an ecosystem. When such a network model is used to broadly explore the role of alligators, it becomes clear that alligators are a keystone species for the Cypress swamp—this means that any large change in alligator abundance would dramatically alter the Cypress swamp ecosystem. *See* POPULATION ECOLOGY.

Human impact. No ecosystem is immune to human impact. In fact, for many ecosystems, human activities are the dominant drivers of change. This means that studying ecosystems without also studying human systems (such as commerce, agriculture, and living patterns) is a sterile and disconnected academic pursuit. For that reason, systems ecology is now combining with the social sciences to model human and natural systems in a coherent framework. The principles and theory are similar—attention is paid to redundancy, attributes that might confer stability, and feedback loops. Models that couple patterns of human land use to biological properties of systems are a good example of how social science might be infused into systems ecology. Land-use change is the most important pressure altering biodiversity; land-use change also alters regional climate, hydrology, agricultural productivity, and even human exposure to diseases. Patterns of taxation, zoning laws, and highway systems interact with soil fertility and vegetation types to shape land-use changes. Consequently, the dynamics of ecosystems may be governed as much by zoning laws as by predator-prey relationships and grazing. But there is more to the melding of social sciences and systems ecology than simply noting that human systems alter ecological dynamics. The basic productivity of ecosystems may have constrained the type of human cultures that have evolved, and it prohibits the development of wealth if lands are so degraded that food production is impossible. Thus, ecological dynamics also alter human systems in profound ways. *See* BIOLOGICAL PRODUCTIVITY.

P. M. Kareiva

Bibliography. T. Allen and T. Starr, *Hierarchy*, University of Chicago Press, 1982; R. Kitching, *Systems Ecology: An Introduction to Ecological Modeling*, University of Queensland Press, 1983; S. Levin, *Fragile Dominion*, Perseus Publishing, Cambridge, MA, 1999; H. T. Odum, *Systems Ecology*, Wiley Interscience, New York, 1983; C. Walters, *Adaptive Management of Renewable Resources*, McGraw-Hill, New York, 1986.

Systems engineering

A management technology involving the interactions of science, an organization, and its environment as well as the information and knowledge bases that support each (**Fig. 1**). The functional definition of technology, as a fundamental human activity, is the organization, application, and delivery of scientific knowledge for the presumed enhancement of society. Thus, technology inherently involves a purposeful human extension of one or more natural processes. Management involves the interaction of the organization with the environment. Associated with this, to make an effective management technology, is the information and knowledge that enables understanding and action to effect change.

Characteristics. The purpose of systems engineering is to support organizations that desire improved performance. This improvement is generally obtained through the definition, development, and deployment of technological products, services, or processes that support functional objectives and fulfill needs. Thus, systems engineering is inherently associated with user organizations and humans in fulfillment of their objectives. The engineering of a system also involves the interaction with humans and organizations that are responsible for the physical implementation of systems. Systems engineers generally play an important role as brokers of information and knowledge in working with user enterprises and with the implementation specialists who accomplish the actual realization of physical systems. **Figure 2** illustrates these conceptual interactions. The enterprise will have various functional needs; the conceptual design parts of a systems engineering effort are concerned with expressing these needs in the form of a functional architecture. Systems engineers are also concerned with translation of this functional architecture into a physical architecture that describes the logical breakdown of the system into subsystems. Each subsystem should be as independent as possible and should be such that integrating them after implementation is straightforward. This physical architecture, or logical design description, of the system is next translated into an implementation architecture that provides guidance for implementation contractors in bringing about the various subsystems. This is also represented in Fig. 2.

Systems engineering has triple bases: a physical (natural) science basis, an organizational and social science basis, and an information science and knowledge basis. The natural science basis involves primarily matter and energy processing. The organizational and social science basis involves human, behavioral, economic, and enterprise concerns. The

information science and knowledge basis is more difficult to cope with, in many ways, than the other bases, since knowledge is not a truly fundamental quantity. In part, this derives from the structure and organization inherent in the natural sciences and in the organizational and social sciences. It also results from the uses to which information is to be put, and the experiential familiarity of information holders with the task at hand and the environment into which the task is embedded such as to enable interpretation of information, within an appropriate context, as knowledge. Thus, the presence of information in Fig. 1 is especially important.

There are at least three major drivers of new technologies: (1) The natural sciences must provide new discoveries that can be converted into technological innovations. (2) There must be a marketplace need for technological innovations. (3) There must be a need to ensure sustainable development.

These three suggest consideration of a technical system, an enterprise system, and a knowledge system. Management of the environment for each is needed. Systems engineers often act as brokers of knowledge in enterprises that have needs for support and implementation specialists able to deliver innovative technologies and services that provide this support. **Figure 3** illustrates these interrelations. It indicates that systems engineering knowledge comprises (1) knowledge perspectives, which represent the view that is held relative to future directions in the technological area under consideration; (2) knowledge principles, which generally represent formal problem-solving approaches to knowledge, generally employed in new situations or unstructured environments; and (3) knowledge practices, which represent the accumulated wisdom and experiences that have led to the development of standard operating policies for well-structured problems.

These three components act together and are associated with learning to enable continual improvement in performance over time. On the basis of the appropriate use of these knowledge types, systems engineers are able to accomplish the technological system design and management system design that lead to an innovative product or service. Knowledge perspectives enable the forecasting of the need for innovation. Innovation results when new knowledge principles are applied to produce new and different products and services, and associated knowledge practices, that fulfill a societal need.

Systems engineering may also be defined as management technology to assist and support policy making, planning, decision making, and associated resource allocation or action deployment. It accomplishes this by quantitative and qualitative formulation, analysis, and interpretation of the impacts of action alternatives upon the needs perspectives, the institutional perspectives, and the value perspectives of clients to a systems engineering study. In fact, all of systems engineering can be thought of as consisting of formulation, analysis, and interpretation activities. Practitioners may exercise these in a formal sense or in an "as if" or experientially based intuitive

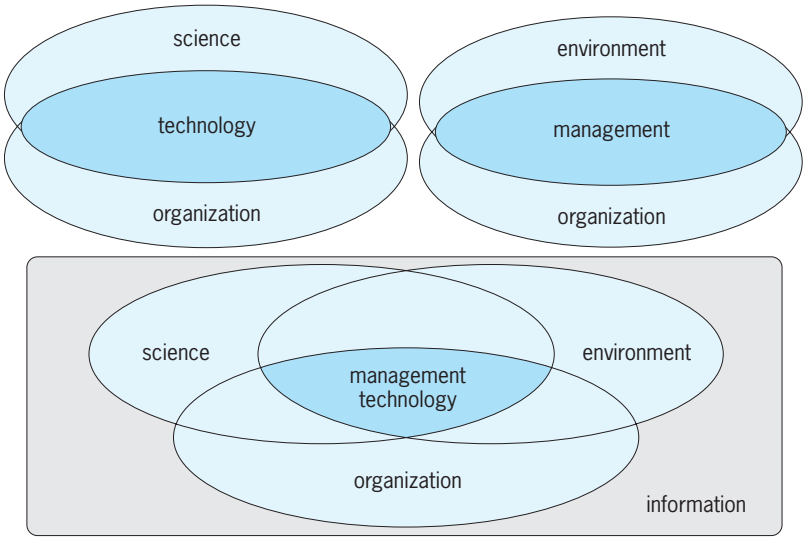


Fig. 1. Systems engineering as a management technology.

sense. Each essential phase of a systems engineering effort—definition, development, and deployment—is associated with formulation, analysis, and interpretation. These enable systems engineers to define the needs for a system, develop the system, and deploy it in an operational setting and provide for maintenance over time. These are the components of a framework for systems engineering (**Fig. 4**).

In order to resolve large-scale and complex problems or manage large systems, systems engineers must be able to deal with contemporary issues that involve:

1. Many considerations and interrelations
2. Many different and perhaps controversial value judgments
3. Knowledge from several disciplines
4. Risks and uncertainties involving future events which are difficult to predict
5. Fragmented decision-making structures
6. Needs perspectives and value perspectives as well as technology perspectives
7. Resolution of issues at the level of institutions and values as well as the level of symptoms

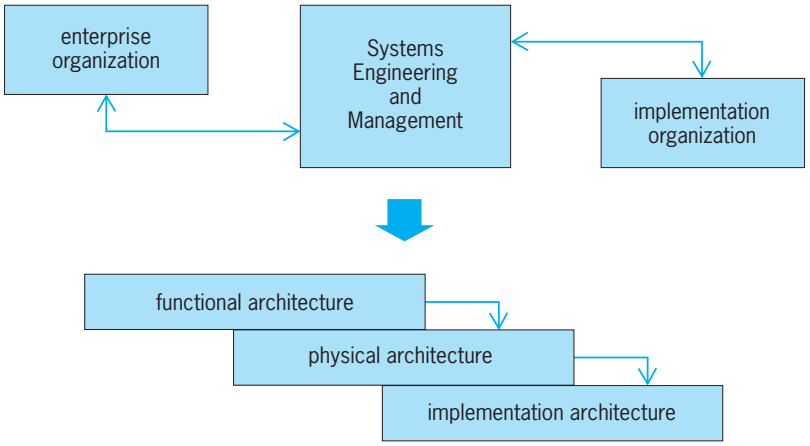


Fig. 2. Systems engineering as a broker of knowledge to enable specification of system architectures.

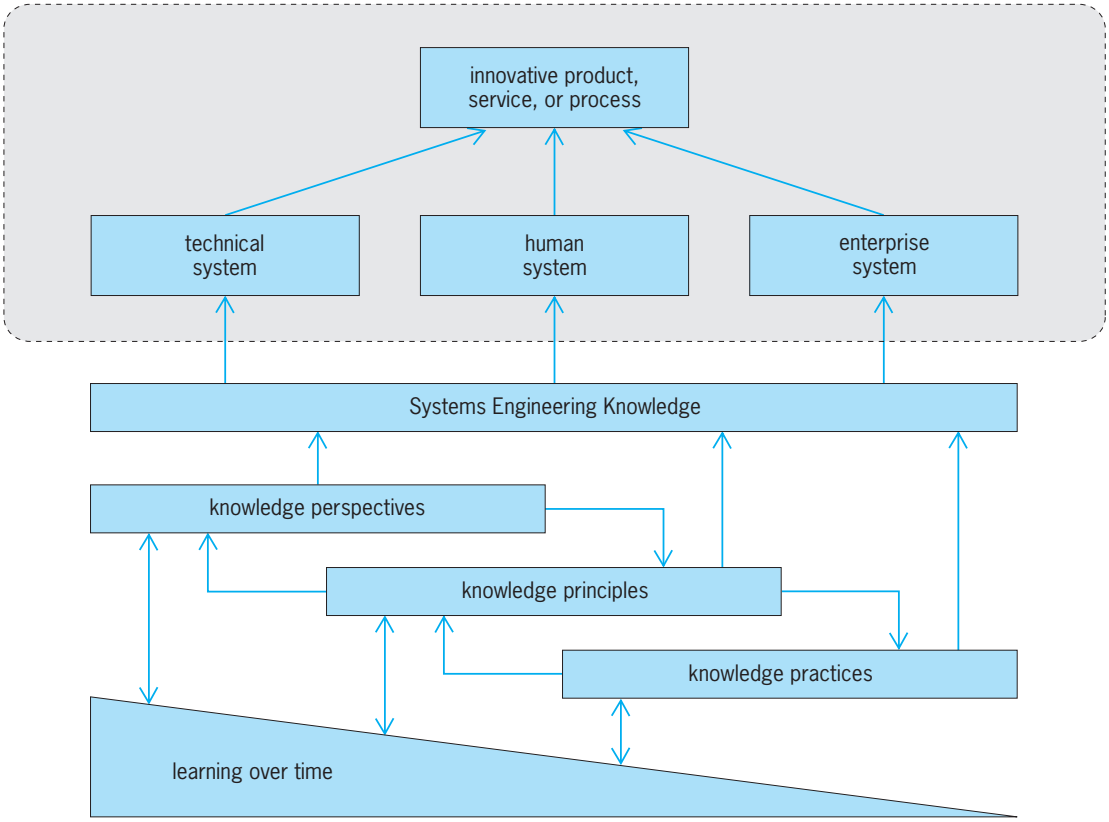


Fig. 3. Systems engineering knowledge and results of its effective use.

In addition, there are potential pitfalls to be avoided:

1. Overreliance upon specific methods or technologies advocated by some group
2. Consideration of perceived problems and issues only at the level of symptoms, and the development and deployment of "solutions" that address only symptoms
3. Failure to develop and apply appropriate lifecycle processes that allow formulation, analysis, and interpretation of impacts in terms of institutional and value considerations
4. Failure to involve the client
5. Failure to consider the effects of poor information processing
6. Failure to identify a sufficiently robust set of options for development
7. Failure to make and properly utilize reactive, interactive, and proactive measurements to guide the systems engineering efforts
8. Failure to identify risks associated with system costs and benefits, or effectiveness
9. Failure to properly design the system for effective user interaction
10. Failure to consider implications of strategies adopted in one of the three life cycles [research, design, testing, and evaluation (RDT&E); acquisition; and planning and marketing] on the other two life cycles
11. Failure to address quality issues comprehensively through all phases of all life cycles
12. Failure to properly integrate a new system with heritage or legacy systems
13. Failure to obtain top-level enterprise management support for the effort
14. Failure to consider sustainability issues

Emergence. Throughout history, the development of more sophisticated tools has invariably been associated with a decrease in our dependence on human physical energy. Generally, this is accomplished by control of nonhuman sources of energy in an automated fashion. The industrial revolution represented a major thrust in this direction.

In most cases, a new tool or machine makes it possible to perform a familiar task in a new and different way, typically with enhanced efficiency and effectiveness and sometimes with increased explicability. Sometimes a new tool makes it possible to do something entirely new.

Profound societal changes have often been brought about by new tools. In the 1850s about 70% of the labor force in the United States was employed in agriculture. Now, less than 3% is so employed, but is able to produce sufficient food for the entire country, generally with large surpluses. Occasionally, tools have produced undesired side effects. Pollution due to chemical plants and potential depletion of fossil fuel sources are examples. There is the potential for even worse side effects, such as following operational errors in nuclear power plants.

Concerns associated with the design of tools for efficient and effective use have always been addressed,

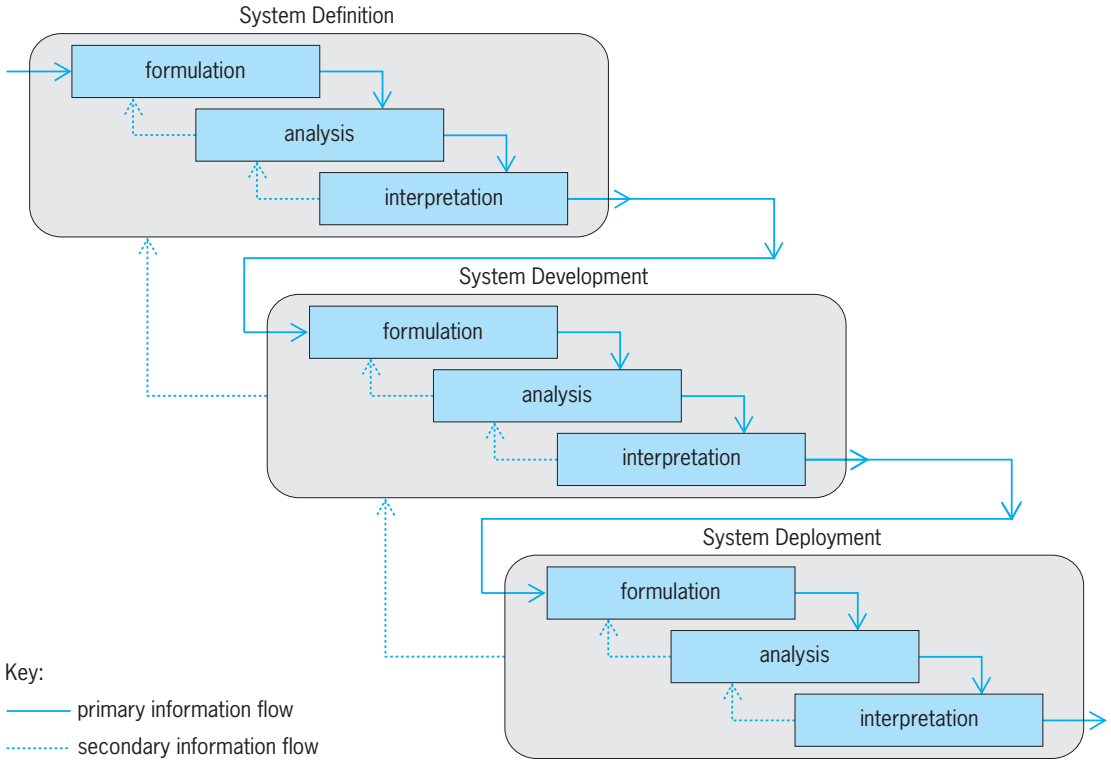


Fig. 4. A systems engineering framework comprising three phases and three steps per phase.

but often on an implicit and trial-and-error basis. In the past, when tool designers were also tool users, the designs were often good initially or soon evolved into good designs. But when physical tools, machines, and systems became so complex that they required a team of designers, a host of new problems emerged. To cope, a number of methodologies associated with systems engineering have evolved. Through these, it has been possible to decompose large design issues into smaller component ones to design the subsystems, and then to collect these subsystems into a complete system. However, simply connecting together the individual subsystems often does not result in an effective and efficient system. This has led to the realization that systems integration engineering and systems management are necessary throughout an entire system life cycle. Thus, contemporary systems engineering focuses on tools, methods, and metrics, as well as on the engineering of life-cycle processes that enable appropriate use of these tools to produce trustworthy systems. There is also a focus on systems management to enable the wise determination of appropriate processes. *See* SYSTEMS INTEGRATION.

Computers, information, and knowledge technologies. Computers are fundamentally different from the usual combination of motors, gears, pulleys, and other physical components that have assisted humans in performing physical tasks. While a computer can assist in performing functions associated with physical tasks, such as computing the optimal trajectory for an aircraft to take between two locations with minimum energy consumption and cost, it can also assist humans in cognitive tasks such as planning, resource allocation, and decision making. The availability of digital computers, and associated communications and networking capabilities, has led to the information technology revolution. In turn, this has led to an increased focus on the use of knowledge and intellectual capital as a major driver of productivity in the developed nations.

The digital computer is an information machine, a knowledge machine, and a cognitive support machine. It has led to the growth of a new engineering field which involves information and knowledge management. This professional area is concerned with efforts whose structure, function, and purpose are associated with the acquisition, representation, storage, transmission, and use of data that are of value as information, and the ability to associate this information with context so as to enable it to become knowledge. This has led to a fundamental change in the way that systems engineering is accomplished. *See* COMPUTER; COMPUTER-AIDED DESIGN AND MANUFACTURING; COMPUTER-AIDED ENGINEERING; COMPUTER-INTEGRATED MANUFACTURING; DATA COMMUNICATIONS; DATABASE MANAGEMENT SYSTEM; DIGITAL COMPUTER.

Requirements. To achieve a high degree of functionality, the requirements of a system must be appropriately defined. The system is often an assemblage of subsystems that must be integrated into existing systems. Systems engineers need to develop a system concept, or architecture, that results in a design that is capable of being efficiently and effectively produced, used, maintained, retrofitted, and modified throughout an extended life cycle. The system life cycle begins with the need for conceptualization

and identification, and proceeds through specification of system requirements and architectural structures, production or manufacturing, to ultimate system installation and evaluation, to ultimate operational implementation.

Many difficulties are often associated with the production of functional reliable, and trustworthy systems of large scale:

1. Large systems are expensive.
2. System capability is less than expected.
3. System deliveries are quite late.
4. Large system cost overruns occur.
5. Large system maintenance is complex and error-prone.
6. Large system documentation is inadequate.
7. Large systems are cumbersome to use, and system design for human interaction is generally lacking.
8. Individual subsystems cannot be integrated.
9. Large systems cannot be transitioned to a new environment or modified to meet evolving needs.
10. Large system performance is unreliable.
11. Large systems do not perform according to specifications.
12. System requirements do not adequately capture user needs.

A plethora of potential difficulties have led to these observations, including inconsistent, incomplete, and otherwise imperfect system requirements specifications; system requirements that do not provide for change as user needs evolve; and poorly defined management structures for product design and delivery. These same studies generally show that the major problems associated with the production of trustworthy systems have more to do with the organization and management of complexity than with the underlying technologies.

Defunctions. Systems engineering may be viewed as the design, production, and maintenance of functional, reliable, and trustworthy systems within cost and time constraints. There is an embedded hierarchy of performance levels for systems engineering efforts. Discussion of systems engineering is assisted by definitions of the discipline in terms of structure, function, and purpose.

Structure. The structural definition of systems engineering states that practitioners are concerned with a framework for problem resolution that, from a formal perspective at least, consists of three steps: issue formulation, issue analysis, and issue interpretation. Regardless of the characterization of the design process, and regardless of the type of process or system that is being designed, all characterizations necessarily involve (1) formulation of the issue, in which the needs and objectives of a client group are identified, and potentially acceptable alternatives or options are identified or generated; (2) analysis of the alternatives, in which the impacts of the identified design options are evaluated; and (3) interpretation and selection, in which options are compared by evaluating the impacts of the alternatives. The needs and objectives of the client group are used as a basis for evaluation. The most acceptable alternative is selected for implementation or further study in the systems engineering effort.

Function. The functional definition of systems engineering states that practitioners will be concerned with the various tools and techniques that enable us to design systems. Often, these will be systems science and operations research tools. It also tells us that systems engineers will be concerned with a combination of these tools to enable the establishment of an appropriate life cycle in engineering a system. Finally, the definition states that systems engineers will establish an appropriate setting. The term "systems management" will be used to refer to the cognitive tasks in producing a useful process from a systems methodology and design study, combining systems science and operations research to resolve issues and to engineer systems.

Purpose. The purpose of systems engineering is the organization and management of information and knowledge to assist clients who desire to develop policies for management, direction, control, and regulation. These activities relate to forecasting, planning, development, production, and operation of total systems to maintain overall integrity as related to trustworthiness in performance and reliability. The model of the steps of the fine structure of the systems process, shown in Fig. 4, is based upon this conceptualization.

Systems design model. Figure 4 illustrates a formal rational model of how design is accomplished. There is the need for much iteration between steps when it is discovered that improvements in an earlier step are needed in order to obtain a quality result at a later step. Also, this description does not emphasize the key role of information throughout all of systems engineering. This framework is not a complete view of the realities of systems engineering. It is correct in an "as if" manner. It is a morphological box that consists of a number of phases and steps. There is also a third-dimensional variable, effort level, and others could be identified as well.

A number of questions may be posed with respect to formulation, analysis, and interpretation that indicate the role of values in every portion of a systems engineering effort. Issue formulation questions are:

1. What is the problem? the needs? the constraints? the alterables?
2. How do the client and the analyst bound the issue?
3. What objectives are to be fulfilled?
4. What alternative options are appropriate?
5. How are the alternatives described?
6. What alternative state of nature scenarios are relevant to the issue?

Analysis questions of importance are:

7. How are pertinent state variables selected?
8. How is the issue formulation disaggregated for analysis?
9. What generic outcomes or impacts are relevant?
10. How are outcomes and impacts described across various societal sectors?
11. How are uncertainties described?

12. How are ambiguities and other information perfections described?

13. How are questions of planning period and planning horizon dealt with?

Interpretation concerns with respect to value influence are:

14. How are values and attributes disaggregated and structured?

15. Do value and attribute structuring and associated formal elicitation augment or replace experience and intuition?

16. How are flawed judgment heuristics and cognitive information processing biases dealt with?

17. Are value perspectives altered by the phase of the systems engineering effort being undertaken?

Finally, how is total issue resolution time divided between formulation, analysis, and interpretation? This is important because the allocation of resources to various systems engineering activities reflects the value perspectives of the analyst and the client. All of this has strong implications for a set of guidelines for professional practice of systems engineering.

There must also be an awareness of appropriate methods, tools, metrics, and techniques of systems science and operations research; and appropriate human judgment at the cognitive process level (systems management). Figure 4 and **Figure 5** show representations of the complete, simplified, systems engineering process described here. We will consider the three primary phases (definition, development, deployment) and three fundamental steps within each (formulation, analysis, interpretation).

Systems engineering efforts are concerned with the technical direction and management of systems design, production, and maintenance. The management technology of systems engineering aims to ensure that correct systems are engineered, which requires considerable emphasis on the front end of the systems life cycle. It also requires attention to ensure not only that the engineered system satisfies the technological specifications (verification) but that it performs so as to satisfy user needs (validation).

To support these ends, there must be emphasis on the accurate definition of a system, what it should do, and how people should interact with it before the system is produced and implemented. In turn, this requires emphasis upon conformance to system requirements specifications, and the development of standards to ensure compatibility and integratibility of system products. Such areas as documentation and communication are important. Thus, the need for the technical direction and management technology efforts that constitute systems engineering is apparent.

Life-cycle methodology. Figure 5 illustrates a typical sequence of seven phases for a systems engineering effort. These may be thought of as being sequenced in an iterative, "once-through," manner. In actual practice, the phases must be sequenced in an incremental or evolutionary manner so as to produce various versions of a product. Within each phase it is necessary to have a number of steps as represented in Fig. 4. Another representation is the systems engineering process framework, as shown in **Fig. 6**, comprising 21 steps, three for each of the seven phases. In practice, many more than 21 activity steps are associated with a systems engineering effort. One dichotomy of systems engineering practice is that, especially for large efforts, it is desirable to consider each of these activity boxes as

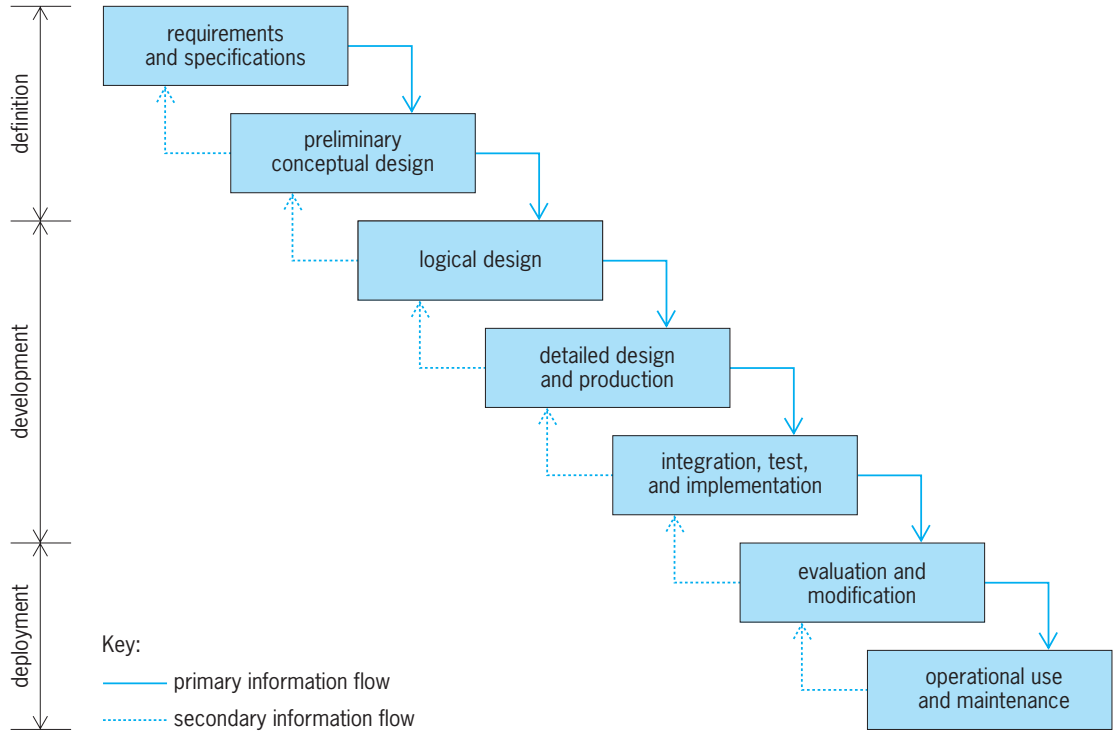


Fig. 5. One of several possible life-cycle models for systems engineering.

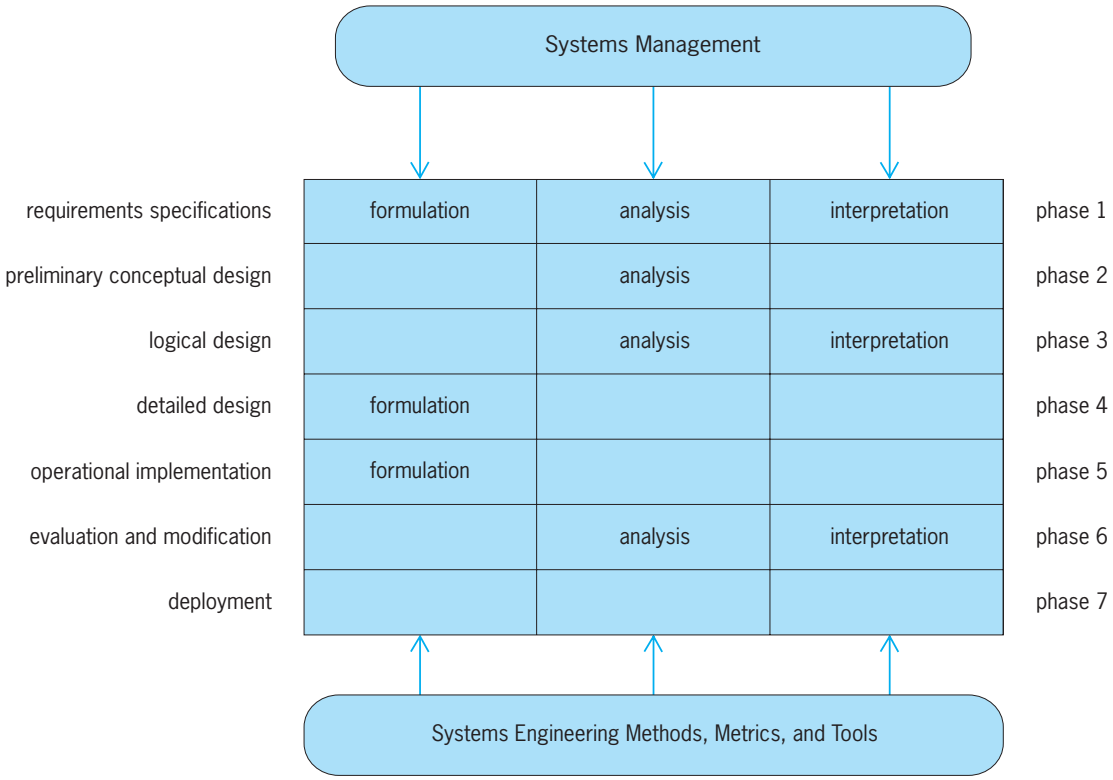


Fig. 6. An alternative representation of a two-dimensional systems engineering framework.

if it can be conducted in a manner dependent only on results from the previous activities. However, the whole is generally greater than the sum of the parts, and this independence rarely occurs. This leads to the need to implement a number of systems project management approaches, such as configuration management and control, that attempt to assure the needed integration throughout the systems engineering process. *See* ENGINEERING DESIGN; SOFTWARE ENGINEERING.

Life-cycle phases. In Fig. 5, the requirements specification phase of systems engineering has as its goal the identification of client needs, activities, and objectives to be achieved by implementation of the resulting design as a product, process, or system. Initially, the user requirements are identified, and these are translated to functional requirements, or specifications for the system. The effort in this phase should result in the description of preliminary conceptual design considerations that are appropriate for the next phase. It is necessary to translate operational deployment needs into requirements specifications in order that these needs be addressed by the system design efforts. Thus, the information concerning the requirements specifications is affected by each of the other phases of the systems engineering process.

From the requirements specifications phase, a clear definition of design issues should result so that it becomes possible to decide whether to undertake preliminary conceptual design. If the requirements specifications effort indicates that client needs can be accommodated in a satisfactory manner, then documentation is typically prepared concerning specifications for the preliminary conceptual design phase. Initial specifications for the following three phases are typically also prepared, and a concept design team is selected to implement the next phase.

Preliminary conceptual design typically includes or results in the functional architecture of the system. This leads to logical design and architectural specification of the physical architecture for the ultimate system (product, service, or process). Arriving at the functional architecture of the system may lead to the development of a prototype that is responsive to the requirements specifications. Preliminary concept design according to the requirements specifications should be obtained. Rapid prototyping of the functional architecture of the conceptual design is desirable for many applications. *See* PROTOTYPE.

The desired product of the next phase, logical design, in Fig. 5 is a set of detailed architectural specifications for a physical system architecture that should result in a useful product, process, or system. There should be a sufficient degree of user confidence that a useful product will result from detailed design, in terms of realization of an implementation architecture, or the entire effort should be redone or possibly abandoned. Another product of this phase is a refined set of specifications for the integration of the resulting system, both in terms of subsystems and in terms of legacy systems. In this third phase of effort, these specifications are translated into detailed representations in logical form, in terms of a physical architecture, such that detailed system development may occur.

A product, service, or process (system) is produced in the fourth phase of design. This is not the final design, but the result of implementation of the physical architecture and the functional architecture

resulting from the last two phases. Establishment of an implementation architecture is needed in this phase. In the fifth phase, the implementation architecture is realized in the form of a system in an operational setting. The actual details of implementation may be the responsibility of the systems engineering organization or of another organization so charged.

Evaluation of the implementation—in terms of the resulting product, process, or system—is achieved in the sixth phase of this process. Preliminary evaluation criteria are obtained as a part of requirements specifications and modified during the following two phases. The evaluation effort must be adapted to other phases such that it becomes an integral part of the overall process. Generally, the critical issues for evaluation are adaptations of the elements in the requirements specifications phase. A set of evaluation test requirements and tests are evolved from the objectives and needs determined in requirements specifications. These should be such that each objective measure and critical evaluation issue component can be measured with at least one evaluation test instrument. *See* SYSTEM DESIGN EVALUATION.

If it is determined, perhaps through an operational evaluation, that the system cannot meet user needs, the effort reverts to an earlier phase and effort continues. An important by-product of evaluation is determination of ultimate performance limitations for an operationally realizable system. Often, operational evaluation is the only realistic way to establish meaningful information concerning functional effectiveness of the result of a systems engineering effort. Successful evaluation is dependent upon explicit development of a plan for evaluation before initiation of the evaluation effort.

The last phase of the systems engineering effort concerns final acceptance and operational deployment. The process could continue until such extended life-cycle phases as maintenance, retrofit, and reengineering are identified. Alternatively, the operational deployment phase can be considered to include these activities. *See* REENGINEERING.

Systems management. Many ingredients are associated with the development of trustworthy systems, including:

1. Systems engineering processes, including process development life cycle and configuration management
2. Process risk, operational-level quality assurance and evaluation, and product risk and development standards
3. Metrics for quality assurance, and process and product evaluation
4. Metrics for cost estimation, and product cost and operational effectiveness evaluation
5. Strategic quality assurance and management, or total quality management
6. Organizational cultures, leadership, and process maturity
7. Reengineering at the levels of systems management, organizational processes and product lines, and product

A number of related issues concern enterprise management and systems integration, economic systems analysis, cognitive ergonomics, and system assessment and evaluation.

One of the first efforts in systems management is to identify process life cycle for production of a trustworthy system. The precise life cycle that is followed will depend on client needs. It will also depend on such environmental factors as the presence of existing subsystems into which a new system must be integrated, and the presence of existing software modules that may be retrofitted and reused as part of the new system. This need for system integration brings about a host of systems management issues and, in many cases, legal issues that are much larger in scale than those associated with program development. Similarly, the development of appropriate system-level architectures is very important. Efficiency and effectiveness in systems architecting have great influence on the ease with which systems can be integrated and maintained and, therefore, of the extent to which an operational system is viewed as trustworthy and of high quality. *See* SYSTEMS ARCHITECTURE.

Following the identification of an appropriate systemic process development life cycle, configuration management plans are identified. This step involves using the life cycle and defining a specific development process for the set of tasks at hand. Metrics are needed for effectiveness. Examples are the metrics of cost analysis, or cost estimation, for systems engineering such as cost and economic estimation for software and information technology–based systems; and of effectiveness analysis or estimation of software productivity indices. The development of a product is coupled with the process needs. These metrics must form a part of a management approach for process, and ultimately product, improvement if substantial progress is to be made in such important areas as software testing and evaluation. *See* ACTIVITY-BASED COSTING; SOFTWARE TESTING AND INSPECTION.

Conclusion. Much contemporary thought concerning innovation, productivity, and quality can be cast into a systems engineering framework. This framework can be valuably applied to systems engineering in general and information technology and software engineering in particular. The information technology revolution provides the necessary tool base that, together with knowledge management–enabled systems engineering and systems management, allows the needed process-level improvements for the development of systems of all types. The large number of ingredients necessary to accomplish needed change fit well within a systems engineering framework. Systems engineering constructs are useful not just for managing big systems engineering projects according to requirements, but for creative management of the organization itself. *See* INFORMATION SYSTEMS ENGINEERING; LARGE SYSTEMS CONTROL THEORY; QUALITY CONTROL; SYSTEMS ANALYSIS.

Andrew P. Sage

Bibliography. J. N. Martin, *Systems Engineering Guidebook: A Process for Developing Systems and Products*, CRC Press, Boca Raton, FL, 1997; E. Rechtin and M. W. Maier, *The Art of Systems*

Architecting, CRC Press, Boca Raton, FL, 1997; A. P. Sage, *Systems Engineering*, John Wiley, New York, 1992; A. P. Sage, *Systems Management for Information Technology and Software Engineering*, John Wiley, New York, 1995; A. P. Sage and W. B. Rouse (eds.), *Handbook of Systems Engineering and Management*, John Wiley, New York, 1999.

Systems integration

A professional activity that utilizes processes and procedures from systems engineering, systems management, and product development for the purpose of developing large-scale complex systems. These complex systems involve hardware and software and may be based on existing or legacy systems coupled with totally new requirements to add significant functionality through integration of new systems or subsystems. Systems integration generally involves combining products of several contractors to produce the working system. Systems integration applications range from creation of complex inventory tracking systems to designing flight simulation models and reengineering large logistics systems.

Life-cycle activities. Application of systems integration processes and procedures generally is in accordance with a set of life-cycle phases for systems engineering. Minimally, these systems engineering life-cycle phases are requirements definition, development and integration of the system, and deployment together with associated operations and maintenance. These three phases include such necessary efforts as feasibility analysis, program and project plans, logical and physical design, design compatibility and interoperability tests, reviews and evaluations, and graceful system retirement.

Systems integration activities require in-depth knowledge of pertinent methodologies, experience in the implementation and management of large-scale integration programs, and the ability to combine and utilize the results of existing system details and constraints on new requirements. A primary need is to work with existing systems and resources that may have been in place for years and have undergone a number of previous modifications to determine how the integration program should be managed to assure that the needs of the user are met. This will enable informed insights concerning such subjects as maintenance of the existing systems in the face of implementation of additional system requirements, and the transition of the existing system to the new one over a significant interval of time.

Systems integration relies on a number of activities that are interactive and iterative. The systems integration life cycle generally envisions the utilization of computer-assisted systems (or software) engineering (CASE) tools to develop, analyze, and implement large-scale systems integration programs. The purposes for CASE tools are to enhance productivity, improve quality of the product, and increase the functionality of the product in operation. *See* SOFTWARE ENGINEERING.

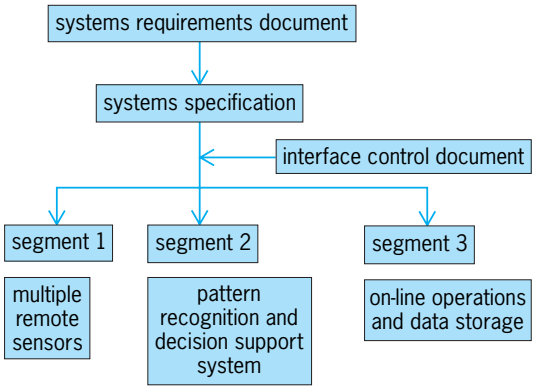


Fig. 1. Systems integration activity, showing the architecture for the system.

As an example, a prime contractor might be responsible for the development of a new data-fusion system involving an existing pattern-recognition system, modification of data storage and on-line operations, and improvement of decision support (**Fig. 1**). The associated activities might proceed as follows. Requirements specifications and feasibility analyses for the system are generally the responsibility of the prime contractor. Traceability matrices and an audit trail are established to track and account for all aspects of the new and improved system. System management parameters are fixed, and project planning is completed. Logical and physical designs are developed. Subcontract specifications and requests for proposal are prepared for the new hardware and software to be procured. Interface design and compatibility and interoperability tests are developed. Subcontractors are selected and are monitored by the prime contractor for contract compliance, quality control, and on-time and within-budget delivery. As component parts and subsystems are delivered, systems integration is accomplished by the prime contractor, generally with the assistance of subcontractors. Interface design, hardware and software compatibility, and interoperability testing are accomplished. Reviews and evaluations with the customer are continued throughout the process. System delivery is accomplished, and actual system operation is completed to customer satisfaction.

Primary uses. Systems integration is essential to the design and development of systems of all types, and especially systems that are information technology- and software-intensive, that automate key operations for business and government. It is required for major procurements for the military services and for private businesses.

Systems integration involves management of the entire life-cycle process from requirements definition to system development and ultimate deployment. It also includes implementation and training. The key personnel required for systems integration are program managers, communications specialists, software specialists, functional specialists, and hardware specialists. The most critical person is the program manager. Program management furnishes planning, controlling, and operating functions and applies these to large-scale complex systems

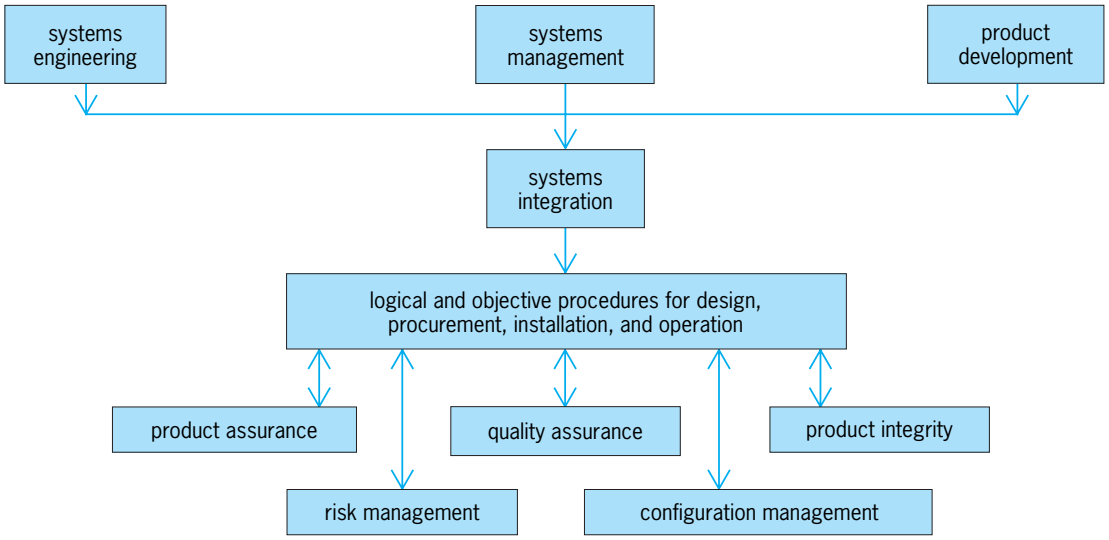


Fig. 2. Tactical and strategic perspective of systems integration, showing essential elements.

programs that include hardware, software, facilities, personnel, procedures, and training needed to meld existing systems and new technologies.

In the above example, systems integration required the coordination of preexisting or coexisting system components with newly developed or modified systems components. This resulted in significant growth of the existing system, merger of two other systems, and the combination of existing system fragments with commercial off-the-shelf components.

For the sample system, this included development and application of a strategic plan, assignment to subcontractors of all terms and conditions from the prime contract, assurance of full compliance with all requirements and constraints, development of an audit trail for complete traceability, and tracking of procurements and schedules. This process assured system growth, future modification, and expansion capabilities to meet initially unrecognized user needs. Systems integration processes and procedures include cost-control mechanisms and work breakdown schedules. They also include attempts to avoid bureaucratic procedures that increase costs to the client and delay deployment of the system (**Fig. 2**).

Application of processes and procedures. The application of systems integration to large-scale complex systems follows standard systems engineering lifecycle activities. The successful application of the systems integration approaches requires (1) completely and correctly defined client needs; (2) complete traceability; (3) consistent system performance criteria; (4) trade-off assessments for alternative solutions, rated as to efficacy and cost; and (5) operational test and evaluation of the integrated system to verify that it satisfies requirements.

The application of systems integration approaches requires that client needs are completely and correctly defined. Elicitation of client needs is one of the most difficult tasks in the development of large-scale complex systems. In most systems integration procurements, specifications are provided by the client and are to be interpreted and analyzed by the development team. This process requires significant interaction between client and developer to assure that the right system is being constructed and that it carries out the desired functions at the performance level needed. Generally, it is highly desirable to maintain close contact between the client and development team throughout the entire life cycle of the program to assure client satisfaction. Thus, requirements and specifications management and the maintenance of a complete audit trail of all actions taken in the program are essential for successful delivery and operation of the completed system.

Procurement criteria, based on function and performance need, must be established prior to consideration of trade-offs, alternatives, or options. This is a complex activity that involves understanding of the economic impact of the procurement, usually without full information on the current state of the art, as practiced by potential subcontractors. Selection criteria must be developed on a functional basis and not by reference to existing commercial configurations. An audit trial that leads from the functional performance requirements of the client to the final procurement choice must be maintained.

Operational test and evaluation criteria must be developed as early as feasible, preferably during requirements specifications, so as to assure that performance requirements will be met. These criteria are to be applied to each of the development phases and most especially to hardware and software procured from outside vendors.

Advantages. Systems integration approaches enable early capture of design and implementation needs. The interactions and interfaces across existing system fragments and new requirements are especially critical. It is necessary that interface and intermodule interactions and relationships across components and subsystems that bring together new and existing equipment and software be articulated. The systems integration approach supports this through application of both a top-down and a

bottom-up design philosophy; full compliance with audit-trail needs, system-level quality assurance, and risk assessment and evaluation; and definition and documentation of all aspects of the program. It also provides a framework that incorporates appropriate systems management application to all aspects of the program. One of the principal advantages of this approach is that it disaggregates large and complex issues and problems into well-defined sequences of simpler problems and issues that are easier to understand, manage, and build. Systems integration provides a suitable methodology that encompasses the entire integration program from requirements through design to construction, test, and finally deployment and maintenance; and supports problem understanding, risk management, and communication between all parties at all stages of development. It is generally necessary to develop plans for systems integration early in the systems engineering life cycle. Needs for systems integration should be established during the definition phase, and the systems architecture for integration is established early in the system development phase. *See* INFORMATION SYSTEMS ENGINEERING; RISK ASSESSMENT AND MANAGEMENT; SYSTEMS ANALYSIS; SYSTEMS ENGINEERING.

James D. Palmer

Bibliography. R. A. Beutel, *Contracting for Computer Systems Integration*, Michie Company, Charlottesville, VA, 1991; Computer Science and Telecommunications Board, *Systems Integration*: *Keeping the U.S. Industry Competitive*, National Academy Press, Washington, DC, 1991; J. O. Grady, *Systems Integration*, CRC Press, Boca Raton, FL, 1994; A. P. Sage, *Systems Engineering*, Wiley, 1992; A. P. Sage, *Systems Management for Information Technology and Software Engineering*, Wiley, New York, 1995; A. P. Sage and C. L. Lynch, Systems integration and architecting: An overview of principles, practices, and perspectives, *Sys. Eng.*, 1(3):176–227, November 1998; A. P. Sage and J. D. Palmer, *Software Systems Engineering*, Wiley, New York, 1990; A. P. Sage and W. B. Rouse (eds.), *Handbook of Systems Engineering and Management*, Wiley, New York, 1999; J. Wyzalek (ed.), *Enterprise Systems Integration*, CRC Press, Boca Raton, FL, 2000.

Syzygy

The alignment of three celestial objects within a solar system. Syzygy is most often used to refer to the alignment of the Sun, Earth, and Moon at the time of new or full moon. Although syzygy is strongly associated with these two lunar phases in many minds, it must be emphasized that the alignment of any three celestial objects within the solar system (or within any other system of objects in orbit about a star) constitutes syzygy. Alignments need not be perfect in order for syzygy to occur: because the orbital planes for any three bodies in the solar system rarely coincide, the geometric centers of three objects that are in syzygy almost never lie along the same line. *See* PHASE (ASTRONOMY); SOLAR SYSTEM.

In general, syzygy occurs whenever an observer on one of the three objects would see the other two objects either in opposition or in conjunction. Opposition occurs when two objects appear 180° apart in the sky as viewed from a third object. Conjunction occurs when two objects appear near one another in the sky as seen from a third object.

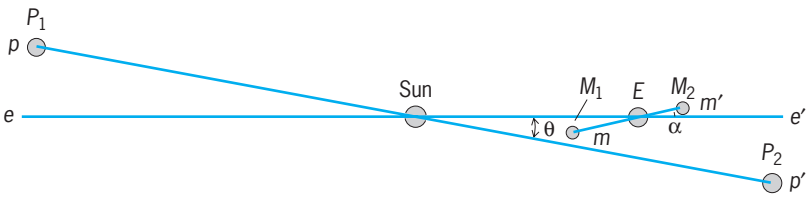
In the **illustration**, e-e' is the orbital plane of the Earth (E), p-p' is the orbital plane of an outer planet (P) that is tilted at an angle θ to the orbital plane of the Earth, and m-m' is the orbital plane of the Moon (M), which is tilted at an angle α to the Earth's orbital plane. Syzygy occurs for the Earth, Sun, and the planet when the planet is either at P_1 (conjunction) or P_2 (opposition). The Sun, Earth, and Moon are in syzygy when the Moon is either at M_1 (new moon) or M_2 (full moon). If the Moon is at M_2 when the planet is at P_2, then the Earth, Moon, and the planet are in syzygy. (As seen from the Earth, the Moon and the planet would be near one another in the sky; that is, the Moon and the planet would be in conjunction.) The Earth, Moon, and the planet will be in syzygy any time that the Moon and the planet are seen in the same direction from the Earth: the Moon need not be full and the planet need not be in opposition. The maximum angle for θ is about 17° (Pluto); the value for α is slightly more than 5°. There is no accepted upper limit for θ; if a new planet were discovered with an orbital plane tilted 25° from that of the Earth, syzygy would still occur at P_1 and P_2. *See* MOON; PLANET.

Solar and lunar eclipses are dramatic results of syzygy. During a solar eclipse, when the Moon is in its new phase, the alignment of the Sun, Earth, and Moon is so nearly perfect that the Moon's shadow falls on the Earth; during a lunar eclipse, which occurs at the time of the full moon, the Moon passes through the Earth's shadow. *See* ECLIPSE.

An occultation is another type of eclipse that can occur during syzygy. For an Earth-based observer, an occultation occurs when the Moon is seen to pass in front of a planet or other member of the solar system. The occultation of a star by the Moon does not qualify as syzygy, since the star is far beyond the limits of the solar system. *See* OCCULTATION.

Harold P. Coyle

Bibliography. V. Illingworth, *The Facts on File Dictionary of Astronomy*, 4th ed., 2000; J. Mitton, *Cambridge Dictionary of Astronomy*, 2001; J. M. Pasachoff and A. Filippenko, *The Cosmos: Astronomy in the New Millennium*, 3d ed., 2007.



Diagram, not to scale, of the Sun, the Earth, the Moon, and a planet, illustrating syzygy. Symbols are explained in text.

T Tauri star

A young low-mass star characterized by variability, the presence of hydrogen emission lines, and association with dark or bright nebulae. T Tauri stars are named after a variable star in the constellation Taurus that exhibits particularly strong hydrogen emission. They were originally believed to be ordinary field stars passing through and interacting with a star-forming nebula. Their association with these regions was soon discovered to be more than coincidental, however, and they are now identified with the earliest phase of young stellar evolution, in which a star emerges from its natal molecular cloud to be detectable at visible wavelengths. T Tauri stars with very strong emission lines are designated classical T Tauri stars; their counterparts with reduced hydrogen emission are known as weak-line T Tauri stars. In addition to the original defining properties, T Tauri stars are generally accompanied by excess x-ray, ultraviolet, infrared, and millimeter-wave emission that arises from an accreting circumstellar disk of dust and gas. *See* MOLECULAR CLOUD.

Young age. Independent lines of evidence confirm the extreme youth of T Tauri stars. They are typically located next to dark clouds of dust and gas in aggregates (T associations) with stellar densities that are too low to have withstood disruption by galactic tides for more than 10^7 years. They exhibit high variability, as expected for objects that have not yet achieved equilibrium. They have luminosities and surface temperatures that place them above the zero-age main sequence on the Hertzsprung-Russell diagram, and have values that are reproduced by theoretical simulations of young stars prior to the onset of nuclear fusion of hydrogen in their inner cores. Evolutionary models indicate stellar ages between 10^5 and 10^7 years. Finally, T Tauri stars exhibit strong lithium absorption lines. The presence of lithium in low-mass stars with turbulent envelopes is an indicator of youth, since the lithium is mixed down to interior regions where it is rapidly depleted by nuclear reactions in more mature stars. *See* HERTZSPRUNG-RUSSELL DIAGRAM.

Origin of emission lines. The presence of high-velocity gas is implied by Doppler broadening of strong emission lines in the spectra of classical T Tauri stars. These features were originally attributed entirely to gas ejected from the star, partly on the basis of observations of bipolar jets in spectrophotometric images at very high spatial resolution. Theoretical simulations indicated that the jets are fueled by coupling of accreting disk material to a rotating stellar magnetosphere. Material leaves the disk plane close to the star and travels along magnetic field lines in a "funnel flow." Although most material falls onto the stellar photosphere, some is flung out of the system along open field lines to appear as bipolar jets. Classical T Tauri stars thus resemble more exotic examples of disk-jet systems, such as accreting white dwarf stars, pulsars, and black holes at the center of active galactic nuclei. More recent comparisons of observed spectral features with simulations from this theoretical framework have led to the conclusion that the high-velocity wings of spectral lines result mostly from the gas that is falling directly onto the star, although some features may still result from outflow. *See* BLACK HOLE; DOPPLER EFFECT; PULSAR; WHITE DWARF STAR.

Properties of circumstellar disks. The existence of circumstellar dust disks around classical T Tauri stars was first established by observations of long-wavelength emission in excess of stellar photospheric values. The distribution of dust radiation from infrared to millimeter wavelengths coincides with that of a disk surrounding the star with temperatures that decrease with disk radius. Simulations of the emission indicate the disks have masses from 0.01 to 10% of the mass of the Sun, sizes of about 100 astronomical units (1.5×10^{10} km or 1×10^{10} mi), and constituent dust grains that are substantially larger than for the interstellar medium. These

conclusions are supported by imaging at high spatial resolution. Aperture-synthesis images of carbon monoxide (CO) emission at millimeter wavelengths display the Doppler signature of a disk that is rotating in accord with Kepler's laws. Scattered light images obtained with the *Hubble Space Telescope* and ground-based telescopes reveal the outline of disk surfaces. Silhouetted disks, or proplyds (protoplanetary disks), also appear around stars located in front of bright nebulae. *See* HUBBLE SPACE TELESCOPE; KEPLER'S LAWS.

The most sensitive infrared observations to date indicate that circumstellar disks are detected around virtually all classical T Tauri stars. Far fewer are observed around weak-line T Tauri stars. This is as expected by theoretical interpretations that link emission line strength and the accretion of gas from a disk onto the star. The low detection rate may indicate the presence of disks that have evolved toward the formation of a planetary system, or it may belie the existence of young stars with little or no circumstellar material to begin with. Small amounts of dust from the collisions of large boulders and planetesimals in "debris disks" with very little molecular gas are detected around some stars in the solar neighborhood. Disks with these properties are currently difficult to observe at the distance of nearby star-forming regions, but a handful of detections around weak-line T Tauri stars indicates that at least some are surrounded by transitional planetesimal disks. Taken as a whole, the properties, rate of occurrence, and inferred evolution for disks around classical and weak-line T Tauri stars provide strong evidence that planet formation is a common by-product of the star-forming process. *See* EXTRASOLAR PLANETS; INFRARED ASTRONOMY; PROTOSTAR; SOLAR SYSTEM; STELLAR EVOLUTION.

David Koerner

Bibliography. I. Appenzeller and R. Mundt, T Tauri stars, *Astron. Astrophys. Rev.*, 1:291–334, 1989; C. Bertout, T Tauri stars: Wild as dust, *Annu. Rev. Astron. Astrophys.*, 27:351–395, 1989; G. H. Herbig, The properties and problems of T Tauri stars and related objects, *Adv. Astron. Astrophys.*, 1:47–103, 1962; B. Reipurth, D. Jewitt, and K. Keil (eds.), *Protostars and Planets V*, University of Arizona Press, Tucson, 2006.

Tabulata

One of two principal orders of extinct Paleozoic corals. The Tabulata appeared in the Lower Ordovician and reached their acme in the Middle Devonian before being severely affected by the Late Devonian extinction event. Their subsequent rediversification was limited, and they became extinct at the end of the Permian. Tabulates were closely related to the other principal Paleozoic coral order, the Rugosa, but neither was ancestral to the post-Paleozoic Scleractinia, which evolved from a different group of anemones. Some Cambrian corals have been claimed to be Tabulata, but are more appropriately assigned to a new order.

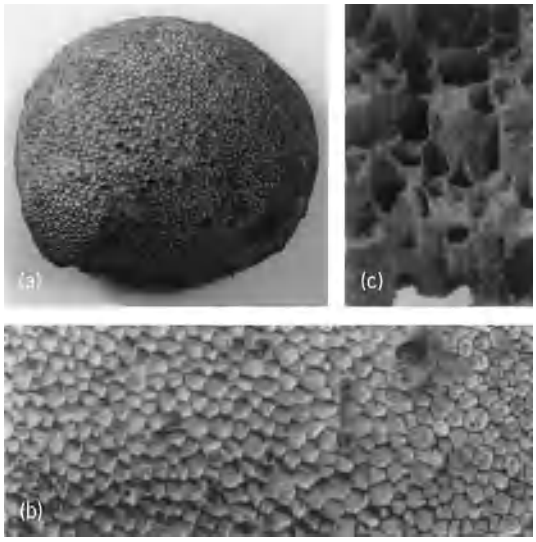


Fig. 1. Structure of the widespread tabulate coral *Favosites*. (*a*) Massive domal colony of *F. multipora* (Silurian, England), ×0.5. (*b*) Detail of surface showing polygonal corallites, ×1.5. (*c*) Oblique view of corallites in *Favosites* sp. (Silurian, Ontario). Note the short septal spines on walls, mural pores (bottom right), and tabulae (upper right, center and left), ×2.

Tabulate corals were exclusively colonial. Polyps secreted slender calcitic (calcium carbonate) tubes (corallites, ranging 0.5–20 mm in diameter, but predominantly 1–3 mm in diameter), polygonal in cross section when in contact, or cylindrical when surrounded by colonial skeletal material (coenenchyme) or not in contact (**Figs. 1** and **2**). The corallites are almost always partitioned by flat or curved, complete or incomplete plates (tabulae). Septal structures, as spines, or less commonly thin plates radially arranged in corallites, often number 12

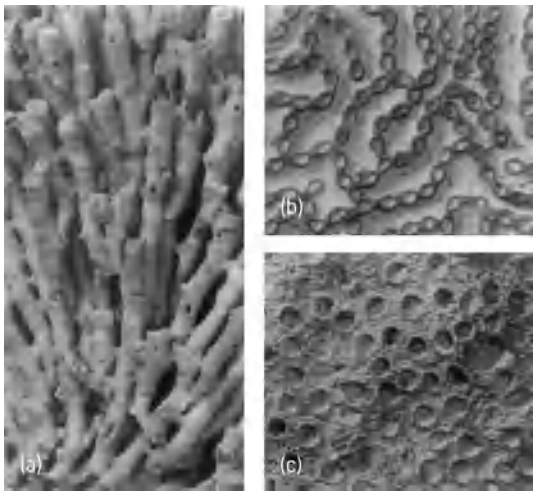


Fig. 2. Common tabulate corals. (*a*) *Syringopora geniculata* (Carboniferous, England), a fasciculate colony with interconnecting tubules (center-left). (*b*) *Halysites catenularius* (Silurian, England), a cateniform colony (chain coral). (*c*) *Heliolites megastoma* (Silurian, England), a massive colony with corallites set in colonial tissue (coenenchyme). All ×1.5.

when well developed. However, these structures are usually weakly developed or completely absent. The Tabulata is divided into six suborders—Lichenariina, Sarcinulina, Favositina, Halysitina, Heliolitina, and Auloporina—based mainly on the structural arrangement of corallites in the colony and the presence or absence of communication (mural pores or connecting tubules) between adjacent corallites. Exceptionally preserved polyps have been found in *Favosites* sp. from the Silurian of eastern Canada. Some groups formerly assigned to the tabulate corals, such as the Chaetetina, are now known to be calcareous sponges.

Tabulate coral colonies could take on a range of external forms. Shape was determined by the interaction of internal controls on colonial growth with prevailing environmental conditions. Solid masses of corallites, either close-packed or set in coenenchyme, are massive (Figs. 1*a* and 2*c*). Such colonies vary from thin, spreading sheets, through domal and bulbous forms, to columnar and branching forms (in which each branch is composed of a mass of corallites). Species vary in the range of forms they could adopt. Another group is characterized by corallites arranged like posts in a fence, giving a chainlike appearance (cateniform) in surface view (Fig. 2*b*). Others consist of short, discrete corallites ramifying across a host skeletal surface or the substrate, with only the aperture turned upward, or are bushy (fasciculate), with or without interconnecting tubules between adjacent corallites (Fig. 2*a*). Most have an external wall, the holotheca, which limits their ability to cement themselves to hard surfaces. Colonies are generally 30–600 mm (1.2–24 in.) in diameter, but massive colonies may reach 2 m (3.3 ft) and some fasciculate colonies may be even larger.

Tabulate corals were most abundant and diverse in temperate to warm shelf environments, particularly in biostromes and bioherms. Smaller, laminar to domal massive colonies inhabited deeper, cooler waters. They were an important component of true reefs, particularly in back-reef, reef-flat and fore-reef environments, but their limited ability of secure attachment to hard surfaces restricted their contribution to reef framework. *See* ANTHOZOA; HEXACORALLIA; RUGOSA; SCLERACTINIA.

Colin Scrutton

Bibliography. D. Hill, Rugosa and Tabulata, in C. Teichert (ed.), *Treatise on Invertebrate Paleontology*, pt. F: *Coelenterata*, suppl. 1, vols. 1, 2, 1981; W. A. Oliver, Jr., Origins and relationships of Paleozoic coral groups and the origin of the Scleractinia, in G. D. Stanley, Jr. (ed.), Paleobiology and biology of corals, *Paleontol. Soc. Pap.*, 1:107–134, 1996; C. T. Scrutton, Corals and other Cnidaria, in R. C. Selley, L. R. M. Cocks, and I. R. Plimer (eds.), *Encyclopedia of Geology*, vol. 2, pp. 321–334, Elsevier, Oxford, 2004; C. T. Scrutton, The Palaeozoic corals, *Proc. Yorkshire Geol. Soc.*, 51:177–208, 1997, and 52:1–57, 1998; J. E. Sorauf, Biocrystallization models and skeletal structure of Phanerozoic corals, in G. D. Stanley, Jr. (ed.), Paleobiology and biology of corals, *Paleontol. Soc. Pap.*, 1:159–185, 1996.

Tacan

A member of the rho-theta family of air navigation systems which define an aircraft's position by its distance and bearing to a single beacon. Such systems inherently answer the navigator's questions of "in what direction" and "how far," without additional computation, and are of particular value when the beacon has to be placed on a ship, oil-drilling rig, or small island. The main weakness of such systems is that bearing errors cause spatial error to increase with distance from the beacon. Major attention is therefore given to the reduction of bearing errors.

Principles of operation. Tacan, first proposed in 1947, is the major survivor of a number of military proposals following World War II. It allows the distance-measuring equipment (DME) to provide bearing service also, without the large antennas or site errors characteristic of the civil very high-frequency omnidirectional range (VOR). Range and accuracy are the same as DME (300 mi or 480 km, and 0.1 mi or 0.16 km, respectively), with a bearing accuracy of 1°. As in DME, operation is on 252 channels, spaced 1 MHz apart, 962–1213 MHz.

To provide the added bearing service, the DME transponder is first arranged to operate at constant duty cycle. This means that the number of output pulses is held constant, whether the beacon is being interrogated by one or a hundred aircraft. When few interrogations are present, the balance is made up of so-called squitter pulses (in early sets these were obtained by merely allowing receiver gain to increase until noise pulses reached a preassigned level).

The total output of the transponder is amplitude-modulated by the rotating directional antenna system (**Fig. 1**). At the center of this system is the central radiator connected to the DME transponder, just as in the conventional DME. However, rotating around this radiation at 15 revolutions per second

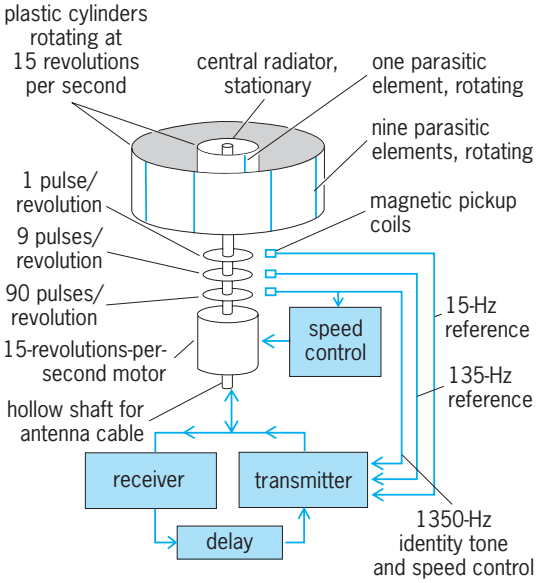


Fig. 1. Tacan transponder with rotating directional antenna system.

are two concentric dielectric cylinders. The inside one, about 6 in. (15 cm) in diameter, contains a single parasitic reflector which imparts a 15-Hz amplitude modulation to the DME replies, and the outside one, about 33 in. (84 cm) in diameter, contains nine parasitic elements which impart a 135-Hz amplitude modulation. On the same rotating shaft are mounted reference pulse generators which additionally modulate the transmitter with coded pulses, once per revolution for the 15-Hz signal (called the north reference burst) and nine times per revolution for the 135-Hz signal (called the auxiliary reference bursts). The composite radiated signal therefore looks like **Fig. 2**. To compensate for the reduced signal amplitude in the "valleys," Tacan beacons are usually of 5 kW power or more.

In the airborne receiver the 15- and 135-Hz sine waves are detected and filtered and compared with the decoded reference bursts to provide a two-speed or fine-coarse bearing display whose accuracy and site freedom are superior to those of conventional VOR, yet with a ground antenna system which is small enough to mount on a ship's mast. This is of particular interest to aircraft based on aircraft carriers, but is also of value in portable transponders.

Site freedom. Tacan's site freedom derives from three factors: (1) Its horizontal antenna aperture is about three wavelengths, as compared with a half wavelength of the standard VOR. (The Doppler VOR, with five wavelengths, is better in this respect, but requires a 150-ft- or 45-m-diameter antenna counterpoise.) (2) Because of Tacan's higher frequency, vertical antenna aperture of up to 12 elements is feasible, leading to reduced radiation into the ground and consequently less site error. Such vertical apertures have been found to be impractical at lower frequencies. (3) Whatever error remains is then divided by 9 by the airborne coarse-fine instrumentation.

Improvements. Major improvements in Tacan for both airborne and ground equipment include the use of all-solid-state components, even at the 5-kW radio-frequency level; replacement of analog signal-processing circuits by digital, primarily to reduce the need for adjustments during manufacture and maintenance; fixed ground equipment monitored from remote service centers to reduce maintenance calls; and reduction in sizes and weights along with increased reliability. Thus, a Vortac station has 5 cabinets, compared to 17 in 1980, and requires a service call once in 3 months instead of weekly.

Installation. About 30,000 military aircraft, chiefly in the NATO countries, are Tacan-equipped. These receivers, of course, also receive distance-only service from civil DME transponders.

In the United States and other countries having a common civil and military airways system, VOR, DME, and Tacan are combined into the Vortac system (**Fig. 3**). Here the VOR and Tacan stations are colocated, one above the other. Civil aircraft obtain their DME service from the Tacan transponder. There are about 750 such stations in the United States.

It might be asked why the VOR is necessary if Tacan provides the same service. The reasons are that

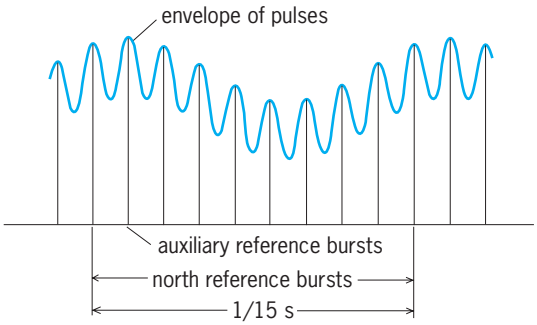


Fig. 2. Composite radiated signal of Tacan transponder. Spaces between reference burst are filled with 2700 random DME replies per second.

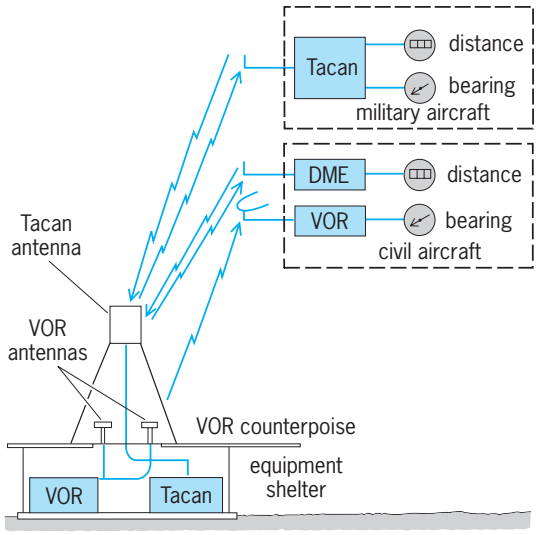


Fig. 3. Vortac system.

VOR preceded Tacan by about a decade and that the VOR receiver, by itself, not only is common to the instrument landing system (ILS) localizer, but also has inherently low cost. There are about 100,000 general-aviation aircraft in the world which carry VOR receivers, but no DME or Tacan. *See* DISTANCE-MEASURING EQUIPMENT; DOPPLER VOR; ELECTRONIC NAVIGATION SYSTEMS; INSTRUMENT LANDING SYSTEM (ILS); RHO-THETA SYSTEM; VOR (VHF OMNIDIRECTIONAL RANGE).

Sven H. Dodington

Bibliography. B. E. Bjerede, A unified signal processor for Tacan navigation sets, *Navigation*, 23:119–127, 1976; M. Kayton and W. Fried, *Avionics Navigation Systems*, 2d ed., Wiley, 1997; A. H. Lang, Second-generation Vortac equipment, *Elect. Commun.*, 58(3):256–262, 1984; P. C. Sandretto, *Electronic Avigation Engineering*, 1958; G. J. Sonnenberg, *Radar and Electronic Navigation*, 6th ed., 1988.

Tachometer

An instrument that measures angular speed, as that of a rotating shaft. The measurement may be in revolutions over an independently measured time interval,

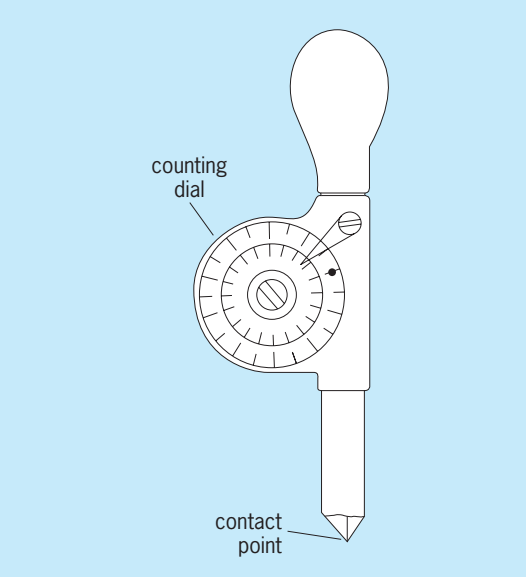


Fig. 1. Revolution counter, the simplest form of tachometer. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

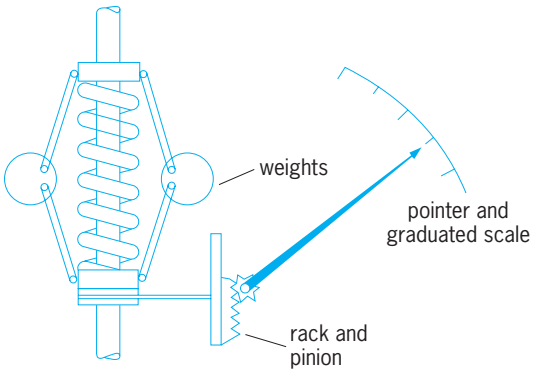


Fig. 2. Centrifugal-force tachometer indicates instantaneous speed. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

as in a revolution counter, or it may be directly in revolutions per minute. The instrument may also indicate the average speed over a time interval or the instantaneous speed. Tachometers are used for direct measurement of angular speed and as elements of control systems to furnish a signal as a function of angular speed.

Revolution counter. The simplest form of tachometer is the revolution counter shown in **Fig. 1**. Held in contact with the rotating shaft, it counts the number of shaft rotations. It is used in conjunction with a timer to obtain average speed over the elapsed period of time. Accuracy is best for uniform speeds measured over long time periods.

Chronometric tachometer. This tachometer counts the revolutions over a fixed interval of time and presents the measurement directly in terms of speed. The result is the average speed for the time interval. The shorter the time interval, the more nearly the instantaneous speed will be approached.

Centrifugal tachometer. The centrifugal force developed by a rotating mass is a function of speed. **Figure 2** shows the essential parts of such a meter. This form indicates instantaneous speed. It is capable of accuracy on the order of ±1% of full-scale value.

Vibrating-reed tachometer. The vibrating-reed tachometer consists of a group of reeds of different length. The lowest natural frequency of vibration of each reed is a function of its length, mass, and cross-sectional dimensions. Observation of the reed that is vibrating forms the means of measuring the frequency of vibration (**Fig. 3**). The instrument is brought into mechanical contact with the device whose mechanical frequency, oscillation, or rotation is to be measured by touching it to the bearing support, case, or frame. The reeds may be adjusted to an accuracy of ±0.3% and are usually guaranteed to ±0.5%.

Impulse tachometer. Tachometers falling into this group may be classified as capacitor charging-current type, inductor type, or interrupted dc type.

Capacitor charging-current tachometer. In the instrument shown in **Fig. 4**, the charging current of a capacitor is utilized. The pickup head usually contains a reversing switch, operated from a spindle, which reverses twice with each revolution. The indicator

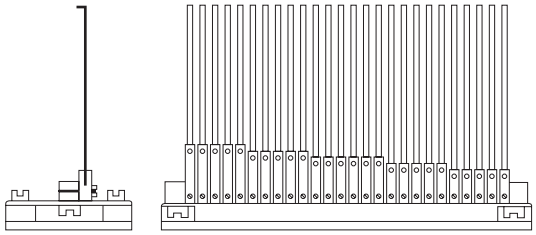


Fig. 3. Vibrating-reed tachometer consists of reeds of varying lengths. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, 2d ed., McGraw-Hill, 1974*)

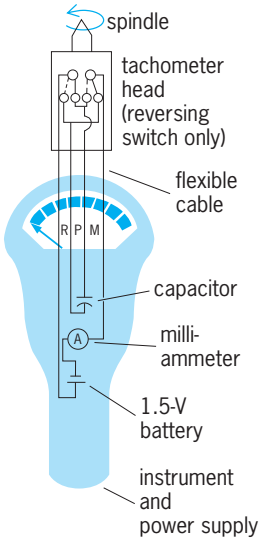


Fig. 4. Capacitor-type impulse tachometer. (*After G. K. McMillan and D. M. Considine, eds., Process/Industrial Instruments and Controls Handbook, 5th ed., McGraw-Hill, 1999*)

responds to the average value of these pulses. The indication is proportional to time rate of these pulses and therefore to the time rate of the spindle revolutions. The battery voltage must be steady and the circuit must be adjusted to the actual value of this voltage. With a steady voltage, this device is capable of good accuracy.

Inductor tachometer. With this instrument the rotating member, which could be a piece of soft iron or laminated iron, causes the magnetic flux of a circuit, containing a magnet and pickup coil, to rise and fall. The rise in flux produces a pulse of one polarity and the fall of flux produces a pulse of the opposite polarity. This is then rectified for a permanent-magnet movable-coil instrument. The pulses may be produced by a lump of magnetic material passing close to a coil having a permanent-magnet core. **Figure 5** shows a form of this tachometer in which the soft-iron piece, or magnet, rotates close to the pickup coil.

Interrupted dc tachometer. The interrupted dc of an ignition-circuit primary, whether battery-excited or magneto-excited, provides a frequency of pulses which can be used to measure speed. **Figure 6** shows a frequency-responsive circuit, in which current from the battery is interrupted by the contactor and excites the ignition coil. The voltage drops from these pulses excite the saturable-core transformer to produce current, which is rectified for the average-reading dc instrument.

Eddy-current tachometer. This form, also known as drag type, is widely used for automobile speedometers and for measuring aircraft-engine speed. It contains a permanent magnet, rotated by the shaft

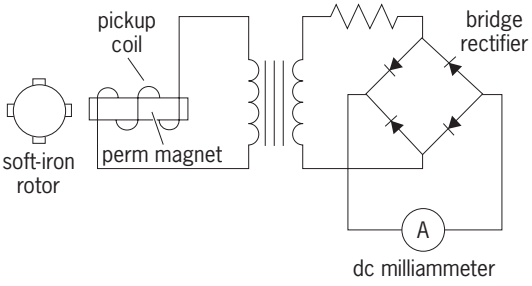


Fig. 5. Circuit for an inductor form of tachometer.

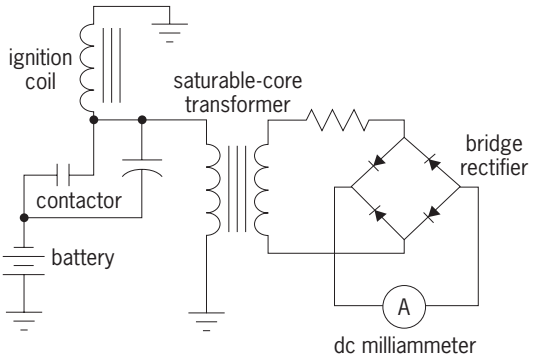


Fig. 6. Schematic circuit of an interrupted dc pulse type of tachometer, as used with the ignition circuit of internal combustion engine.

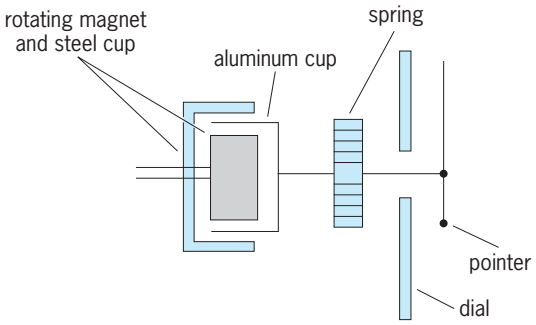


Fig. 7. Drag-type, or eddy-current, tachometer. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, 2d ed., McGraw-Hill, 1974*)

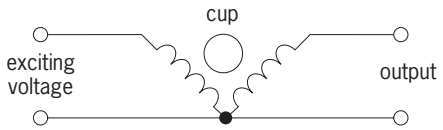


Fig. 8. Schematic circuit and components of a drag-cup tachometer generator.

whose speed is to be measured. Close to the revolving magnet is an aluminum disk or cup, mounted to a staff with a pointer, pivoted and free to turn against a spring (**Fig. 7**). The pointer is associated with a calibrated scale. As the permanent magnet revolves, eddy currents are produced in the disk or cup. The magnetic fields caused by these eddy currents produce a torque, which acts in a direction to resist the turning magnet field. The cup or disk will then turn against the spring, in the direction of the rotating-magnet field, and will turn (or be dragged) until the torque developed by the eddy currents equals that of the spring. An accuracy of 10 revolutions/min in a full scale of 3000 revolutions/min may be obtained. The movable member may be revolved as much as 1080°, affording high resolution in indication.

Drag-cup generators, which also use eddy currents, are used in control systems. They have two stationary windings, positioned so as to have zero coupling, and a nonmagnetic metal cup, which is revolved by the source whose speed is to be measured (**Fig. 8**). One of these windings is used for excitation, inducing eddy currents in the rotating cup.

These eddy currents in the cup produce a field which induces in the other winding an emf proportional to the speed of the rotating cup and at the same frequency as the exciting source. These generators usually have a high degree of linearity, low electrical noise, and low starting and running torque. The output emf and energy are also low.

Velocity-head tachometer. With this type of tachometer the device whose speed is to be measured drives a pump or blower, producing a fluid flow, which is converted to a pressure. **Figure 9** shows a form of this tachometer that not only indicates the speed but produces a tape recording. This form has been used on railroad locomotives. The tape is moved by the drive shaft whose speed is being measured.

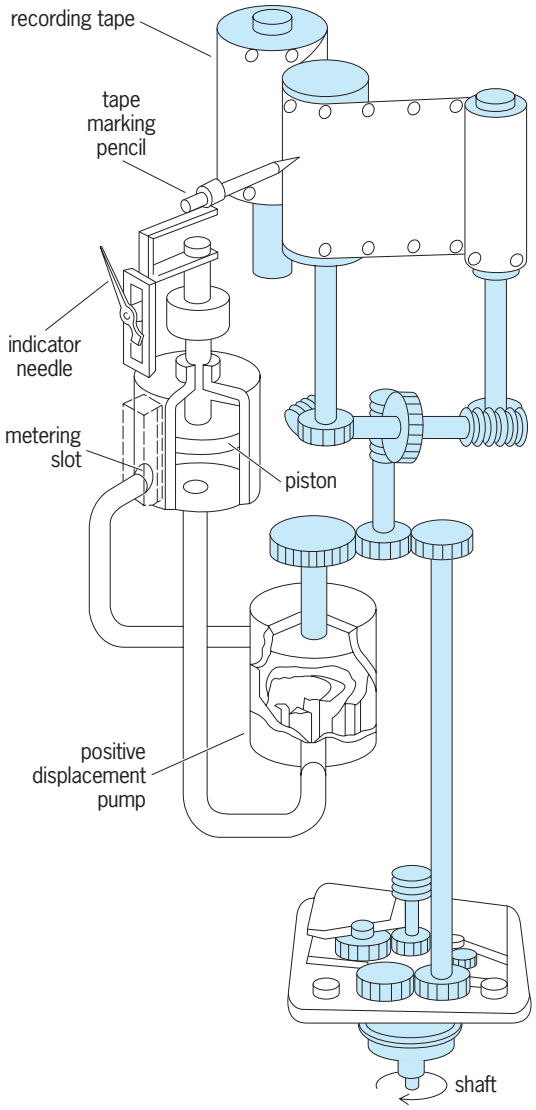


Fig. 9. Sectional view of velocity-head-type tachometer. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

Electronic tachometer. Circuits including vacuum tubes or transistors are sometimes needed to amplify weak pulses or to shape pulse waves. These pulses may be produced by photoelectric, magnetic, or inductor transducers. These shaped waves are applied to frequency-responsive circuits, are rectified, and then are applied to a dc milliammeter.

Electronic counters, also known as EPUT (events per unit of time) meters, are designed to measure frequency, whether sine-wave or pulse. The input pulse waves, clipped and shaped by a discriminating circuit, are allowed to pass through a gating circuit to the displaying decade counters. The open time of the gate is controlled by a crystal oscillator or other suitable time base. The displaying counters are responsive to the number of pulses passed by the gate in a definite time. The accuracy, established by the time-base generator, can be better than 0.01%. Since the display counters or readout may be read to ± 1 count, the number of pulses applied to the input should be sufficiently high to realize its inherent accuracy. *See* FREQUENCY COUNTER.

Electric-generator tachometer. A widely used and flexible form of electric tachometer comprises a combination of electric generator and indicator. There are two principal forms: a dc generator with a dc voltmeter, and an ac generator with an ac voltmeter (or dc voltmeter with rectifier). In either case, the emf developed is proportional to the shaft speed.

The ac generator form may include a circuit that is responsive to frequency and affected only slightly by voltage, giving greater accuracy.

Direct-current generator. The dc tachometer is a small permanent-magnet generator with an output of 2–10 V per 1000 revolutions/min. A high-resistance voltmeter, calibrated in revolutions per minute, indicates the speed.

The dc generator assumes a polarity dependent upon direction of rotation. When used with a zero-center indicator, the instrument will indicate the direction of rotation. With standardizing, an accuracy in the order of 0.25–0.1% may be realized. *See* DIRECT-CURRENT GENERATOR.

The low starting and running torque makes it useful for measuring wind velocity. Besides measuring speed, the dc tachometer is used as a stabilizing component in velocity servomechanisms. *See* CONTROL SYSTEMS.

Alternating-current generator. The ac tachometer generator can be constructed with a stationary winding and a revolving permanent-magnet field. Both generated voltage and frequency are proportional to speed of rotation. The voltage may be rectified and applied to a permanent-magnet moving-coil instrument calibrated in revolutions per minute. *See* ALTERNATING-CURRENT GENERATOR.

By the addition of a saturable-core transformer to the instrument circuit, as shown in **Fig. 10**, the instrument indication becomes frequency-responsive and only slightly affected by voltage, affording greater accuracy.

The difference of two speeds may be measured or indicated by connecting the outputs of two speed-measuring circuits to a differential bridge as shown in **Fig. 11**. The individual speeds may also be indicated.

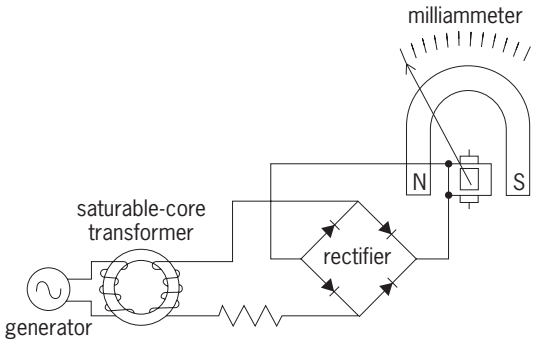


Fig. 10. Schematic circuit of an ac tachometer to which a saturable-core transformer has been added to increase accuracy. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

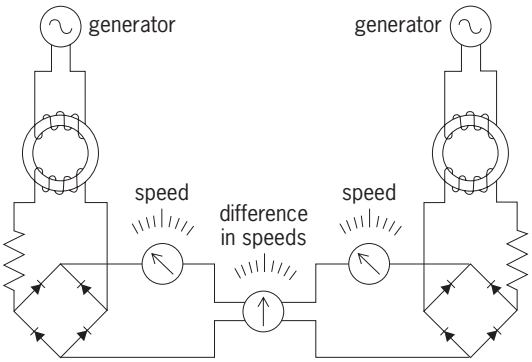


Fig. 11. System employing two ac electric-generator tachometers for indicating individual and differential speeds. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

This difference of speed indication is independent of the actual speeds.

If one of the generators is replaced by a fixed frequency, the other speed may be measured in reference to this frequency, or the scale may be calibrated in terms of the speed. This circuit also affords the means for suppressing as much as 90% of a speed range and allowing the top 10% to occupy the entire scale. Alfred H. Wolferz

Bibliography. D. M. Considine and S. D. Ross (eds.), *Handbook of Applied Instrumentation*, 1964; reprint 1982; G. K. McMillan and D. M. Considine (eds.), *Process/Industrial Instruments and Controls Handbook*, 5th ed., 1999; P. H. Sydenham (ed.), *Handbook of Measurement Science*, vol. 2: *Practical Fundamentals*, 1983.

Tachyon

A hypothetical faster-than-light particle consistent with the special theory of relativity. According to this theory, a free particle has an energy E and a momentum $\mathbf{p}$ which form a Lorentz four-vector. The length of this vector is a scalar, having the same value in all inertial reference frames. One writes Eq. (1), where

$$E^2 - c^2p^2 = m^2c^4 \tag{1}$$

c is the speed of light and the parameter m^2 is a property of the particle, independent of its momentum and energy. Three cases may be considered: m^2 may be positive, zero, or negative. The case $m^2 > 0$ applies for atoms, nuclei, and the macroscopic objects of everyday experience. The positive root m is called the restmass. If $m^2 = 0$, the particle is called massless. A few of these are known: the electron neutrino, the muon neutrino, the tau neutrino, the photon, and the graviton. The third case, $m^2 < 0$, was studied originally by S. Tanaka and by O. M. P. Bilaniuk, V. K. Deshpande, and E. C. G. Sudarshan. Further contributions were made by G. Feinberg, who gave the name tachyons (after a Greek word for swift) to the particles with $m^2 < 0$. Whether such particles exist is an interesting speculation, but there has been no experimental evidence for them.

In general, the particle speed is given by Eq. (2).

$$v = \frac{cp}{E}c \tag{2}$$

If $m^2 > 0$, Eq. (1) implies $E > cp$ and Eq. (2) gives $v < c$. If $m^2 = 0$, then $E = cp$ and $v = c$. In case $m^2 < 0$, one finds $E < cp$ and $v > c$. Tachyons exist only at faster-than-light speeds.

The quantities p or E can be eliminated from Eqs. (1) and (2) to get expressions for E and p in terms of the speed. In case $m^2 > 0$, the familiar results are given by Eqs. (3), where the radical signs

$$E = \frac{mc^2}{\sqrt{1 - \dfrac{v^2}{c^2}}} \qquad p = \frac{mv}{\sqrt{1 - \dfrac{v^2}{c^2}}} \tag{3}$$

imply the positive roots. The way to make the analogous formulas for tachyons is to introduce the positive number μ such that $m^2 = -\mu^2$ and Eq. (4) holds. Then Eqs. (2) and (4) give Eqs. (5). It is seen that for

$$E^2 - c^2p^2 = -\mu^2c^4 \tag{4}$$

$$E = \frac{\mu c^2}{\sqrt{\dfrac{v^2}{c^2} - 1}} \qquad p = \frac{\mu v}{\sqrt{\dfrac{v^2}{c^2} - 1}} \tag{5}$$

ordinary particles as v increases, E increases, but to speed them up to $v = c$ would involve an infinite amount of energy. In contrast for tachyons, as v decreases, E increases, but to slow them down to $v = c$ would involve an infinite amount of energy.

For an ordinary free particle with $m^2 > 0$, there is always a special reference frame, called the rest frame, in which $\mathbf{p} = 0$. From Eq. (1) it is seen that the energy there has the minimum value mc^2. No such special frame exists for massless particles. For tachyons also the situation is quite different. For a free tachyon a special frame can be found in which $E = 0$ and, according to Eq. (4), p has the minimum value μc. Since the tachyons may exist at zero energy, there is no energy obstacle in creating them in elementary particle reactions.

According to electromagnetic theory, a charged particle moving at a speed greater than the speed of light in a medium emits light, the Cerenkov radiation. If charged tachyons existed, they would spontaneously radiate light even in a vacuum. *See* CERENKOV RADIATION.

Attempts to detect tachyons have been made by looking for the Cerenkov radiation (T. Alväger and M. N. Kreisler) and by analyzing for negative m^2 values in elementary particle reactions (C. Baltay, R. Linsker, N. K. Yeh, and G. Feinberg). The conclusions were negative, and present indications are that this type of particle does not exist. *See* ELEMENTARY PARTICLE; RELATIVITY. Roland H. Good, Jr.

Bibliography. O. M. P. Bilaniuk, V. K. Deshpande, and E. C. G. Sudarshan, Meta relativity, *Amer. J. Phys.*, 30:718–723, 1962; G. Feinberg, Particles that go faster than light, *Sci. Amer.*, 222(2):68–73, February 1970.

Taconite

The name given to the siliceous iron formation from which the high-grade iron ores of the Lake Superior district have been derived. It consists chiefly of fine-grained silica mixed with magnetite and hematite. As the richer iron ores approach exhaustion in the United States, taconite becomes more important as a source of iron. To recover the ore mineral in a usable form for the production of iron, taconite must be finely ground, and the magnetite or hematite concentrated by a magnetic or other process. Finally, the concentrate must be agglomerated into chunks of size and strength suitable for the blast furnace. *See* IRON METALLURGY; ORE AND MINERAL DEPOSITS; ORE DRESSING.

Cornelius S. Hurlbut, Jr.

Taeniodonta

An extinct order of quadrupedal eutherian land mammals known from the early Cenozoic deposits of the Rocky Mountain intermontane basins of western North America and, based on a single tooth of *Ectoganus gliriformis*, from early Cenozoic rocks in South Carolina. The nine known genera of taeniodonts are classified into two families: (1) the medium-size (5–15 kg or 11–33 lb), relatively primitive, omnivorous conoryctids (*Schochia, Onychodectes*, and *Conoryctella*, early Paleocene; *Huerfanodon* and *Conoryctes*, middle Paleocene); and (2) the larger (15–110 kg or 33–243 lb) and more advanced stylinodontids (*Wortmania*, early Paleocene; *Psittacotherium*, middle Paleocene; *Ectoganus*, late Paleocene to early Eocene; and *Stylinodon*, middle Eocene).

Conoryctids developed enlarged canines, but the lower jaws were unspecialized and the cheek teeth were low-crowned, enamel-enclosed, and cuspidate. Stylinodontid taeniodonts developed deep massive jaws, peglike teeth that were ever-growing, large curved canines bearing enamel bands, and large laterally compressed and recurved claws on the front paws. The most derived taeniodonts, such as *Stylinodon mirus*, were apparently active diggers, rooters, and grubbers living in warm temperate to subtropical environments. In terms of modern analogs, an advanced stylinodontid may be thought of as an aardvark with the head of a pig.

Taeniodonts were never common members of the faunas in which they occurred; they reached their maximum species diversity during the early and middle Paleocene and then persisted at low diversity and absolute numbers into the middle Eocene. *See* EUTHERIA.

Robert M. Schoch

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, W. H. Freeman, 1988; C. M. Janis, K. M. Scott, and L. L. Jacobs (eds.), *Evolution of Tertiary Mammals of North America*, vol. 1: *Terrestrial Carnivores, Ungulates, and Ungulatelike Mammals*, Cambridge University Press, 1998; R. M. Schoch, *Systematics, Functional Morphology and Macroevolution of the Extinct Mammalian Order Taeniodonta*, Yale Univ. Peabody Mus. Nat. Hist. Bull. 42, 1986.

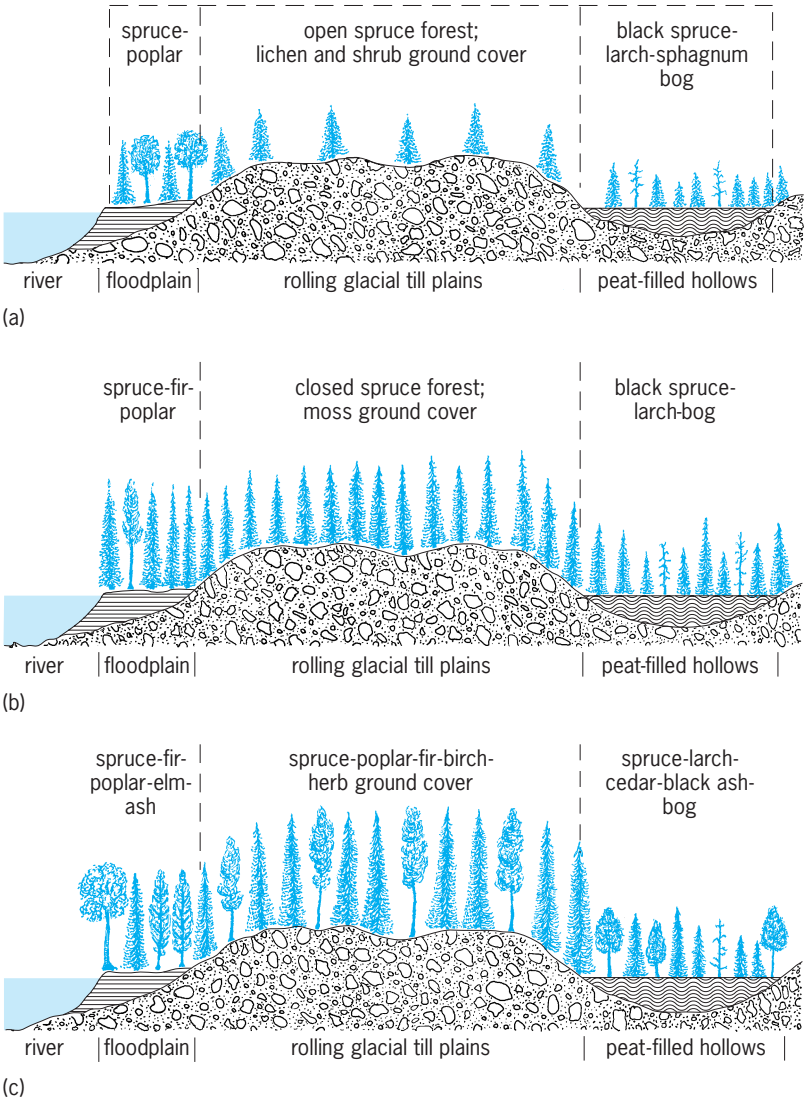
Taiga

A zone of forest vegetation encircling the Northern Hemisphere between the arctic-subarctic tundras in the north and the steppes, hardwood forests, and prairies in the south. The chief characteristic of the taiga is the prevalence of forests dominated by conifers. The taiga varies considerably in tree species from one major geographical region to another, and within regions there are distinct latitudinal subzones. The dominant trees are particular species of spruce, pine, fir, and larch. Other conifers, such as hemlock, white cedar, and juniper, occur locally, and the broad-leaved deciduous trees, birch and poplar, are common associates in the southern taiga regions. Taiga is a Siberian word, equivalent to "boreal forest." *See* FOREST AND FORESTRY; TUNDRA.

Climate. The northern and southern boundaries of the taiga are determined by climatic factors, of which temperature is most important. However, aridity controls the forest-steppe boundary in central Canada and western Siberia. In North America there is a broad coincidence between the northern and southern limits of the taiga and the mean summer and winter positions of the arctic air mass. In the taiga the average temperature in the warmest month, July, is greater than 50°F (10°C), distinguishing it from the forest-tundra and tundra to the north; however, less than four of the summer months have averages above 50°F (10°C), in contrast to the summers of the deciduous forest farther south, which are longer and warmer. Taiga winters are long, snowy, and cold—the coldest month has an average temperature below 32°F (0°C). Permafrost occurs in the northern taiga. It is important to note that climate is as significant as vegetation in defining taiga. Thus, many of the world's conifer forests, such as those of the American Pacific Northwest, are excluded from the taiga by their high precipitation and mild winters.

Subzones. The taiga can be divided into three subzones (see **illus.**) in most of the regions which it occupies; these divisions are recognized mainly by the particular structure of the forests rather than by changes in tree composition. These subdivisions are the northern taiga, the middle taiga, and the southern taiga.

Northern taiga. This subzone is characterized on moderately drained uplands by open-canopy forests, dominated in Alaska, Canada, and Europe by spruce and in Siberia by spruce, larch, and pine. The well-spaced trees and low ground vegetation, usually rich lichen carpets and low heathy shrubs, yield a beautiful parkland landscape; this is exemplified in North America by the taiga of Labrador and Northern Quebec. This subzone is seldom reached by roads and railways, in part because the trees seldom exceed 30 ft (9 m) in height and have limited commercial value. These forests are important as winter range of Barren Ground caribou, but in many parts of the



Schematic profile of the three main subzones of the North American taiga, showing the main forest assemblages on three of the more important landform types. (*a*) Northern taiga. (*b*) Middle taiga. (*c*) Southern taiga.

drier interior of North America their area has been decreased by fire, started both by lightning and by humans.

Middle taiga. This subzone is a broad belt of closed-canopy evergreen forests on uplands. The dark, somber continuity is broken only where fires, common in the drier interiors of the continents, have given temporary advantage to the rapid colonizers, pine, paper birch, and aspen poplar. The deeply shaded interior of mature white and black spruce forests in the middle taiga of Alaska and Canada permits the growth of few herbs and shrubs; the ground is mantled by a dense carpet of mosses. Here, as elsewhere in the taiga, depressions are filled by peat bogs, dominated in North America by black spruce and in Eurasia by pine. Everywhere there is a thick carpet of sphagnum moss associated with such heath shrubs as bog cranberry, Labrador tea, and leatherleaf. Alluvial sites bear a well-grown forest yielding merchantable timber, with fir, white spruce, and black poplar as the chief trees.

Southern taiga. This subzone is characterized on moderately drained soils throughout the Northern Hemisphere by well-grown trees (mature specimens up to 95 ft or 29 m) of spruce, fir, pine, birch, and poplar. These trees are represented by different species in North America, Europe, and Siberia. Of the three taiga zones, this has been exploited and disturbed to the greatest extent by humans, and relatively few extensive, mature, and virgin stands remain. In northwestern Europe this subzone has been subject to intensive silviculture for several decades and yields forests rich in timber and pulpwood. In Alaska and Canada the forests have a much shorter history of forest management, but they yield rich resources for the forest industries. *See* FOREST MANAGEMENT.

Fauna. In addition to caribou, the taiga forms the core area for the natural ranges of black bear, moose, wolverine, marten, timber wolf, fox, mink, otter, muskrat, and beaver. The southern fringes of the taiga are used for recreation. J. C. Ritchie

Bibliography. B. V. Barnes and S. H. Spurr, *Forest Ecology*, 4th ed., 1998; A. Bryson, *Geographical Bulletin*, vol. 8, 1966; *The New Oxford Atlas*, 1975.

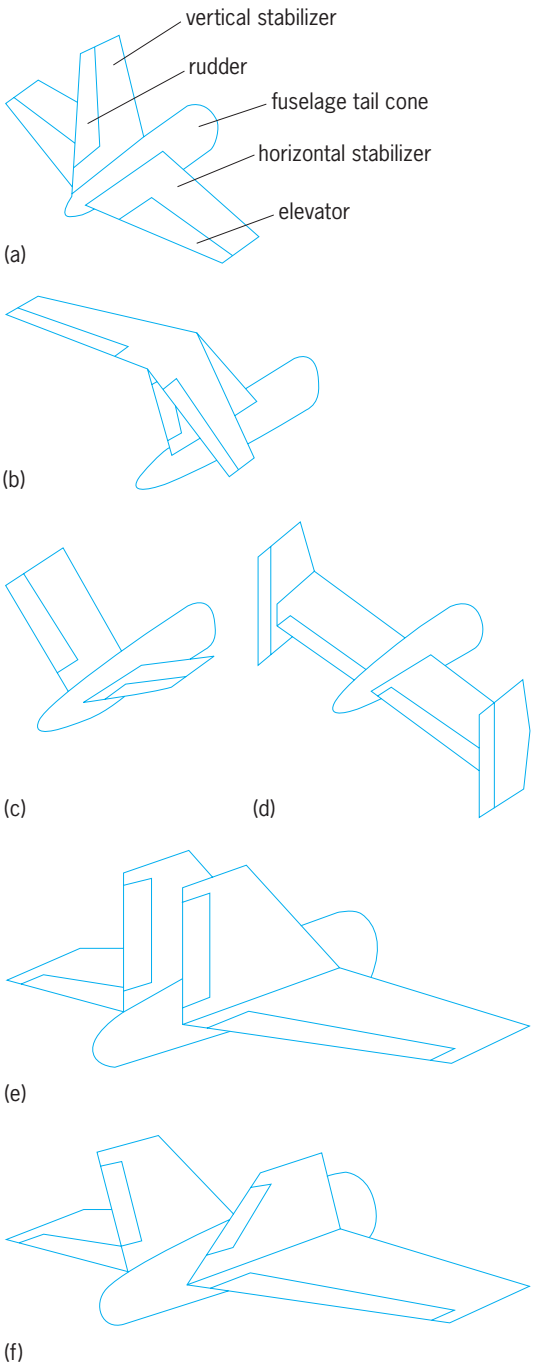
Tail assembly

An assembly at the rear of an airplane, consisting of the tail cone, the horizontal tail, and one or more vertical tails.

Normal configuration. The tail assembly, or empennage, of an airplane is normally composed of a vertical tail and a horizontal tail attached to the rear, or tail cone, of the airplane's fuselage. The vertical tail is composed of the vertical stabilizer and the rudder (**illus.** *a*). The vertical stabilizer is attached rigidly to the fuselage and is intended to provide stability about a vertical axis through the airplane's center of gravity. The rudder is attached by hinges to the rear of the vertical stabilizer and can rotate from side to side in response to pilot control input. It also contributes to stability, but its main function is to provide a yawing moment about the airplane's vertical (yaw) axis, thereby causing the airplane to yaw (turn) to the left or right. *See* AIRCRAFT RUDDER; FLIGHT CONTROLS; FUSELAGE; STABILIZER (AIRCRAFT).

The horizontal tail, similar to the vertical tail, is composed of the horizontal stabilizer and the elevator. The horizontal stabilizer is fixed rigidly to the fuselage and provides stability about a horizontal axis directed along the wing and through the center of gravity, and known as the pitch axis. The elevator is hinged to the rear of the horizontal stabilizer and rotates up and down as the pilot moves the control column fore and aft. The elevator also contributes to stability about the pitch axis, but its main purpose is to provide a pitching moment about the pitch axis, which causes the airplane to nose up or down. *See* ELEVATOR (AIRCRAFT).

Airplanes that operate at supersonic speeds usually have the horizontal tail swept back and in one



Tail assembly configurations. (*a*) Normal configuration. (*b*) T-tail. (*c*) V-tail. (*d*) Twin-tail. (*e*) Uncanted twin vertical tails. (*f*) Canted twin tails.

piece that is movable and is controlled by the motion of the pilot's control stick. Such a surface is frequently called a stabilator. *See* SUPERSONIC FLIGHT.

Other configurations. Many airplanes employ empennages that depart from the normal configuration.

T-tail. In this configuration (illus. *b*), the horizontal tail is attached to the top of the vertical tail in order to raise it above the wake from the wing. Most jet transports employ this type of tail. The T-tail is an extreme example of the more general case where the horizontal tail may be placed vertically at any location along the vertical tail. Also, when the horizontal tail is attached to the fuselage, its location may be ahead of, in line with, or behind the vertical tail. The horizontal placement of the horizontal tail is frequently done to improve the spin-recovery characteristics of the airplane.

V-tail. This tail (illus. *c*) has been used sparingly in the past and its best-recognized application is in the Beechcraft Bonanza, a single-engine, piston-powered, four-place airplane. The single tail, shaped like a V, provides aerodynamic forces about both the yaw and pitch axes, thus providing the functions of both the horizontal and vertical tail surfaces. Pitch control is accomplished by control surfaces which, when moved up or down together, provide a pitching moment in a manner similar to the conventional elevator. If these control surfaces are moved differentially, that is, one up and the other down, then a yawing moment results similar to that produced by a rudder.

Multiple vertical tails. It is not uncommon for airplanes to have more than one vertical tail. An empennage with two vertical tails is referred to as a twin tail (illus. *d*). The purpose of having two or more vertical tails may be aerodynamic, but it can also be to avoid the use of one large vertical tail. The Lockheed Constellation, one of the first four-engine, propeller-driven transports of the 1950s, is reputed to have resorted to three vertical tails because the height of a single tail would have exceeded the hangar doors then in use by the airline that first ordered it.

Many modern fighters use two vertical tails mounted near the center just outboard of the fuselage. These tails may be oriented vertically (Fig. 1*e*), as in the case of the F-15 fighter, or they may be canted outward at the top (Fig. 1*f*), as used on the F-14 and F-18 fighters. An appreciable amount of cant is employed in the F-18 tails for the purpose of enhancing the stability and control of the airplane at angles of attack in excess of 50°. *See* AIRFRAME; AIRPLANE; MILITARY AIRCRAFT. Barnes W. McCormick, Jr.

Bibliography. J. D. Anderson, Jr., *Introduction to Flight*, 4th ed., 1999; B. W. McCormick, *Aerodynamics, Aeronautics and Flight Mechanics*, 2d ed., 1994; L. M. Nicolai, *Fundamentals of Aircraft Design*, 1975; D. Stinton, *The Design of the Airplane*, 1983; D. B. Thurston, *Design for Flying*, 2d ed., 1994; E. Torenbeek, *Synthesis of Subsonic Airplane Design*, 1982, reprint 1986.

Takakiales

An order of liverworts in the subclass Jungermanniidae, consisting of a single genus and two species: *Takakia lepidozioides* with the smallest chromosome number, 4, known in bryophytes, and *T. ceratophylla* (5 chromosomes).

Some authors put the Takakiales in the Calobryales owing to branching from a prostrate branched stem, lack of rhizoids, copious mucilage secretion, and massive frequently scattered archegonia that lack protective envelopes. However, sporophytes have

never been seen, and so it is difficult to demonstrate any meaningful relationship.

The members of the Takakiales consist of a very small gametophyte made up of a branched system of prostrate, leafless stolons and erect, radially organized branches with terete appendages. The "leafy" branches are simple or forked and have a weak central strand of narrow cells enclosed by larger cells. The leafy appendages are small and scalelike below, larger and crowded above, and variable in arrangement. They are dissected to the base into two to four segments consisting of one to several central cells and an epidermis of somewhat smaller cells. The leaf cells lack oil bodies. Mucilage is secreted by simple filaments on leafy branches and by branched filaments clustered on both leafy and stoloniform branches. The archegonia are scattered and not enclosed by protective structures, but both antheridia and sporophytes are unknown. *See* BRYOPHYTA; CALOBRYALES; HEPATICOPSIDA. Howard Crum

Talc

A hydrated magnesium layer silicate (phyllosilicate) with composition close to $Mg_3Si_4O_{10}(OH)_2$. Limited substitution of aluminum (Al) or titanium (Ti) for silicon (Si) and of iron (Fe), manganese (Mn), or aluminum for magnesium (Mg) is possible. Steatite is massive talc aggregate that contains less than 1.5% calcium oxide (CaO), 1.5% iron(II) oxide + iron(III) oxide ($FeO + Fe_2O_3$), and 4% aluminum oxide (Al_2O_3). Talc commonly is white, but it may appear pale green or grayish depending on the amount of minor impurities. It usually occurs in coarse to fine platy or fibrous aggregates that often have a more or less parallel arrangement. Shreds and plates are often bent; some tabular crystals exhibit perfect basal cleavage, yielding flexible, slightly elastic lamellae. Talc has a greasy feel and pearly luster, and has been used as one of the hardness standards for rock-forming minerals with the value 1 on the Mohs scale. Because talc is soft, it can be scratched by fingernails.

Talc frequently occurs in magnesium-rich metamorphosed serpentinites and siliceous dolomites. Talc-rich rocks include massive soapstones, massive steatite, and foliated talc schists. In regional and thermal metamorphic terranes, talc is associated with magnesite, dolomite, tremolite, or chlorite in medium-grade rocks, and with enstatite, olivine, or anthophyllite in high-grade rocks. Progressive changes of talc-bearing assemblages with increasing grade of metamorphism toward granitic plutons have been well documented in the Alps and in California. In the upper amphibolite facies, talc dehydrates to enstatite and quartz at temperatures above 1380°F (750°C). At lower temperatures in the greenschist facies, both tremolite and chlorite may be converted to talc by carbon dioxide (CO_2) metasomatism, but at still lower temperatures talc is unstable in the presence of calcium oxide and carbon dioxide, and is replaced by dolomite or magnesite.

Talc is also often a hydrothermal mineral formed at the expense of serpentine and tremolite in shear zones of serpentinites, as lenticular veins, or as actinolite–chlorite–talc rinds around some blocks enclosed within serpentinite. Steatization is commonly associated with serpentinization; the conversion of serpentine to talc may occur by the addition of silica and removal of magnesium, or by the addition of carbon dioxide to form magnesite + talc. Although not abundant in sediments, authigenic talc has been identified in evaporites and in unmetamorphosed carbonates.

Hydrothermal experiments and geologic observations suggest that talc is stable over a wide range of temperatures. In the presence of quartz, talc may form from serpentine at $T < 660$°F (350°C). It is stable up to 1380°F (750°C), where it dehydrates to enstatite or anthophyllite. The stability field of talc becomes very restricted in environments of low activity of silicon dioxide (SiO_2). A decrease in the activity of silicon dioxide in hydrothermal fluid causes talc to be replaced by serpentine (at low temperatures) or olivine (at high temperatures). Occurrence of talc in serpentinitized ultramafic rocks suggests introduction of silicon dioxide or carbon dioxide during hydrothermal alteration. Talc-bearing assemblages can be used to estimate temperature and fluid composition for thermal or regional metamorphism of siliceous dolomite and serpentinite.

Talc is a good insulating material. It has been commonly used in industry as a raw material for ceramics, paints, plastics, cosmetics, papers, rubber, and many other applications. *See* SILICATE MINERALS. J. G. Liou

Bibliography. S. W. Bailey (ed.), *Hydrous Phyllosilicates (Exclusive of Micas), Reviews in Mineralogy*, vol. 19, 1998; W. A. Deer, R. A. Howie, and J. Zussman, *The Rock-Forming Minerals*, vol. 3: *Sheet Silicates*, 1966; J. J. Hemley et al., Mineral equilibria in the MgO-SiO_2-H_2O system: I. Talc-chrysotile-forsterite-brucite stability relations, *Amer. J. Sci.*, 277:322–351, 1977; K. L. Kimball, F. S. Spear, and H. J. B. Dick, High temperature alteration of abyssal ultramafics from the Islas Orcadas Fracture zone, South Atlantic, *Contrib. Mineral. Petrol.*, 91:307–320, 1985.

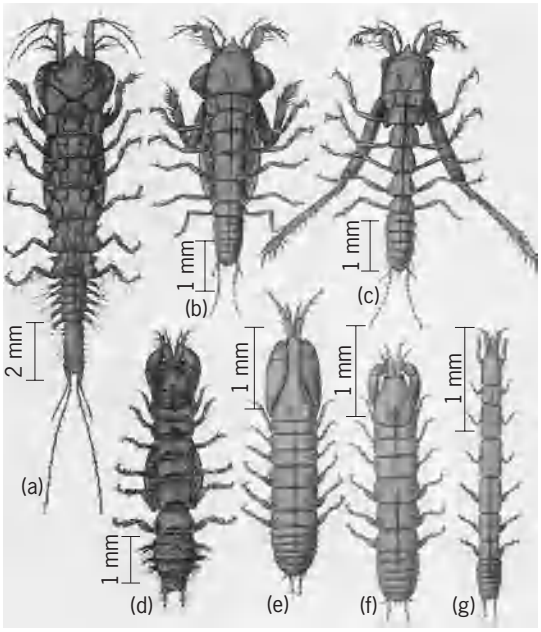
Tall oil

A by-product from the pulping of pine wood by the kraft (sulfate) process. In the kraft process the wood is digested under pressure with sodium hydroxide and sodium sulfide. (The alternate process name, sulfate, is derived from the sodium sulfate used to replace sulfur lost in the process.) The volatilized gases are condensed to yield sulfate turpentine. During the pulping the alkaline liquor saponifies fats and converts the fatty and resin acids to sodium salts. Concentration of the pulping solution (black liquor) prior to recovery of the inorganic pulping chemicals allows the insoluble soaps to be skimmed from the surface. Acidification of the skimmed soap yields crude tall oil.

Crude tall oil from southern pines contains 40–60% resin acids (rosin), 40–55% fatty acids, and 5–10% neutral constituents. Abietic and dehydroabietic acids comprise over 60% of the resin acids, while oleic and linoleic acids predominate in the fatty acid fraction. The neutral components of tall oil are principally sterols (sitosterol) and diterpene alcohols and aldehydes. During heating of tall oil, esterification takes place, reducing the acid number.

Tall oil is refined by vacuum distillation. Fractionation of 1.1 tons (1 metric ton) of tall oil yields approximately 660 lb (300 kg) of fatty acids, 770 lb (350 kg) of rosin, and 770 lb (350 kg) of intermediate (distilled tall oil), head, and pitch fractions. United States fractionation capacity is over 1×10^6 tons (9×10^5 metric tons) per year.

Fatty acids from tall oil distillation may contain as much as 10–40% resin acids (distilled tall oil) or as little as 0.5%. About 35% of United States fatty acid production and more than half of the unsaturated fatty acid part come from tall oil. Major uses of tall oil fatty acids as chemical raw materials are in coatings, resins, inks, adhesives, and soaps and detergents, and as flotation agents. Tall oil is an important source of rosin in the United States. *See* PINE TERPENE; ROSIN; WOOD CHEMICALS. Irving S. Goldstein



Tanaidacea. (*a*) *Apseudes spinosus*, female; (*b*) *Sphyrapus anomalus*, female; (*c*) *Sphyrapus anomalus*, male; (*d*) *Tanais cavolini*, female; (*e*) *Heterotanais oerstedi*, male; (*f*) *Heterotanais oerstedi*, female; (*g*) *Leptagnathia filiformis*, female.

Tanaidacea

An order of the eumalacostracans of the superorder Percarida, derived from the genus *Tanais*. These animals have a worldwide distribution and with few exceptions are marine. They occur from the shore down to abyssal depths. They are free-living and benthonic. The order is divided into 2 suborders with 5 families, 44 genera, and about 350 species. The body is linear, more or less cylindrical or dorsoventrally depressed (see **illus.**). Thoracid segments 1 and 2 are fused with the head, forming a carapace enclosing a respiratory chamber on each side, and the last abdominal segment is fused with the telson. The left mandible has a lacinia mobilis, while this structure may be present or absent on the right mandible. Eight pairs of thoracic legs are present, of which the first pair are maxillipeds, the second pair chelipeds (the first pair of pereiopods), and the following six pairs pereiopods. The third pair in *Monokonophora*, as a rule, is fossorial, and the chelipeds and first pair of pereiopods usually have vestigial exopodites. Pleopods may be present or absent, and the uropods are filiform.

The nervous system consists of a brain, a subesophageal mass, and a ventral chain. Eye lobes are present and sessile, with or without visual elements. Antennulae, especially in the males, have aesthetes. Sense hairs or sense spines are found on the segments and legs.

In the reproductive organs, the gonads are double. The oviducts open laterally at the base of the fifth pair of pereiopods. The vasa deferentia have a common vesicula seminalis, which is ventromedian on the last thoracic segment. Hermaphroditism, protandry, and protogyny can occur, and sex dimorphism is common. The antennulae are always different in the two sexes. Differences can occur also in the shape of the head, the mouthparts, the chelipeds, the second pair of pereiopods, and the pleopods.

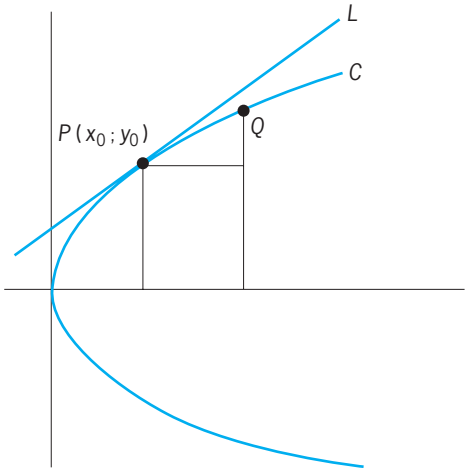
The alimentary canal consists of a ventral mouth, a stomach with a complicated filter and masticatory apparatus, a syncytial midgut, and a terminal anus. As a rule, there are two pairs of hepatopancreas. The excretory organs are a single pair of maxillary glands.

The females produce several broods, each preceded by a molting, by which the exterior morphology may undergo considerable alterations. The eggs develop in an incubatory pouch formed by one or four pairs of oostegites in which the embryos are dorsally flexed. The newly hatched larva lacks the last pair of pereiopods and the pleopods. The young probably undergo four larval (the manca) stages. *See* PERACARIDA; SEXUAL DIMORPHISM. Karl Lang

Bibliography. K. Lang, Neotanaidae nov. fam., with some remarks on the phylogeny of the Tanaidacea, *Arkiv Zool.*, 9:469–475, 1956; K. Lang, The postmarsupial development of the Tanaidacea, *Arkiv Zool.*, 4:409–422, 1953; K. Lang, Protogynie bei zwei Tanaidaceen-Arten, *Arkiv Zool.*, 11:535–540, 1958; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; R. Siewing, Besteht eine engere Verwandschaft zwischen Isopoden und Amphipoden?, *Zool. Anz.*, 147:166–180, 1951; R. Siewing, Morphologische Untersuchungen an Tanaidaceen und Lophogastriden, *Z. Wiss. Zool.*, 157:333–426, 1954; T. Wolff, Crustacea Tanaidacea from depths exceeding 6000 meters, *Galathea Rep.*, 2:187–241, 1956.

Tangent

A term describing a relationship of two figures (usually of the same dimension) in the neighborhood of a common point. The figures are tangent at a point P if they touch at P but do not intersect in a sufficiently small neighborhood of P. To be more precise, if P denotes a point of a curve C (see **illus.**), a line L is a



Line *L* tangent to curve *C* at *P*.

tangent to C at P provided L is the limit of lines joining P to a variable point Q of C, as Q approaches P along C (that is, for Q sufficiently close to P, the line PQ is arbitrarily close to L). If curve C has equation $y = f(x)$ and point P on C has coordinates (x_0,y_0), it is shown in the calculus that the slope of the line tangent to C at P is the value of the derivative $f'(x)$ for $x = x_0$. *See* CALCULUS.

Leonard M. Blumenthal

Tangerine

A name applied to certain varieties of a variable group of loose-skinned citrus fruits belonging to the species *Citrus reticulata*. Although mandarin and tangerine are often used interchangeably to designate the whole group, tangerine is applied more strictly to those varieties (cultivars) having deep-orange or scarlet rinds, whereas the term mandarin is more properly used to include all members of this quite variable group of citrus fruits. *See* MANDARIN.

The two principal tangerine varieties are the Dancy, grown mainly in the United States, and the Algerian (Clementine), grown extensively in the Mediterranean countries but also in California, Arizona, and Texas. Robinson and Murcott tangerines are also important in Florida. Trees are medium-sized with rounded tops bearing leaves pointed at the tip, rounded at the base, and having narrowly winged petioles. The flowers are white and small. The fruits are deep orange, loose-skinned, and small to medium-sized, and possess small seeds with green cotyledons. Tangerines are easily peeled and eaten out of hand as fresh fruit. About one-third of the crop is utilized in juice, sherbets, and canned sections.

A number of tangerine hybrids, the result of controlled breeding programs by the U.S. Department of Agriculture and California experiment stations, are delicious fruits. Important among these are the tangelos, which are hybrids of Dancy tangerine by grapefruit. *See* BREEDING (PLANT).

Frank E. Gardner; C. Jack Hearn

Tantalum

A chemical element, symbol Ta, atomic number 73, and atomic weight 180.948. It is a member of the vanadium group of the periodic table and is in the 5*d* transitional series. Oxidation states of IV, III, and II are also known. *See* PERIODIC TABLE; TRANSITION ELEMENTS.

1												13	14	15	16	17	18
1 H	2																2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Tantalum metal is used in the manufacture of capacitors for electronic equipment, including citizen band radios, smoke detectors, heart pacemakers, and automobiles. It is also used for heat-transfer surfaces in chemical production equipment, especially where extraordinarily corrosive conditions exist. Its chemical inertness has led to dental and surgical applications. Tantalum forms alloys with a large number of metals. Of special importance is ferrotantalum, which is added to austenitic steels to reduce intergranular corrosion.

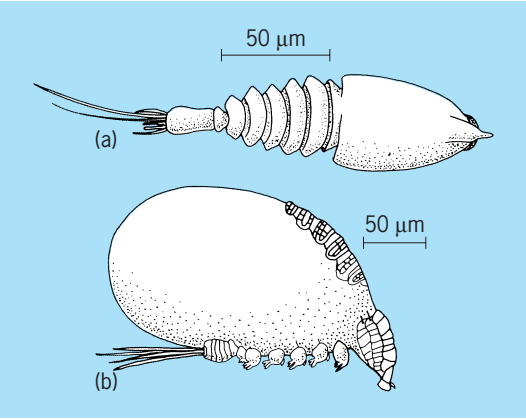
The metal is quite inert to acid attack except by hydrofluoric acid. It is very slowly oxidized in alkaline solutions. The halogens and oxygen react with it on heating to form the oxidation-state-V halides and oxide. At high temperature it absorbs hydrogen and combines with nitrogen, phosphorus, arsenic, antimony, silicon, carbon, and boron. Tantalum also forms compounds by direct reaction with sulfur, selenium, and tellurium at elevated temperatures.

Edwin M. Larsen

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; R. N. Crockett, *Niobium (Columbium) and Tantalum: International Strategic Minerals Inventory Summary Report*, 1994; P. Moeller, P. Cerny, and S. Saupe (eds.), *Lanthanides, Tantalum, and Niobium*, 1989.

Tantulocarida

A subclass of the Crustacea. The Tantulocarida was initially proposed as a class of the Crustacea to accommodate two species formerly assigned to the Copepoda or Cirripedia, but now, with four additional genera that have been described, it is ranked as a subclass. Tantulocarids are minute ectoparasites, less than 0.01 in. (0.3 mm) in length, that infest deep-sea copepods, isopods, tanaids, and ostracods (see **illus.**). As a result of their parasitic mode of



Tantulocarida. (*a*) *Deoterthron aselloticola*. (*b*) *Basipodella harpacticola*. (*After G. A. Boxshall and R. J. Lincoln, Tantulocarida, a new class of Crustacea ectoparastic on other crustaceans, J. Crust. Biol., 3:1–16, 1983*)

life, adult females have lost all resemblance to crustaceans; males are free-living but nonfeeding. Infection of the host occurs at the tantulus larval stage, which is believed to occur immediately after the larva is released. The head, or cephalon, of the tantulus larva is covered by a dorsal shield that may protrude anteriorly to form a rostrum. No cephalic appendages are present, although some of the specialized structures found in tantulocarids may possibly be derived from them. The mechanism of attachment, the mouth tube, is located on the ventral surface with a distal mouth opening, and is surrounded by an oral disk. The walls of the mouth tube consist, at least in part, of a pair of chitinous bars, each with a basal swelling. A second tubular structure, the longitudinal organ, projects into the mouth tube, reaching almost to the oral opening. A structure distinguished by its striated appearance, and referred to as the striated organ, lies above, in proximity to the dorsal surface of the head. A prominent, elongate cephalic stylet is located behind the oral region.

Posterior to the cephalon, the trunk is made up of six free somites, each with well-developed tergal plates and a pair of appendages (thoracopods). Each of the first five pairs consists of a basal segment with median, spined endite and a pair of rami. The sixth pair of thoracopods is much narrower; rami and endites are lacking. The abdomen varies in segmentation from two in *Deoterthron* and *Microdajus* to seven in *Stygotantulus*, and it terminates in a pair of small caudal rami.

Development of the maturing larva can give rise to either the adult male or female. In the former, the tantulus swells between the fifth and sixth thoracic somites or posterior to the sixth. Although the head and remainder of the thorax remain unchanged, a major reorganization of the body occurs within the trunk sac. From an undifferentiated cellular mass, the body of the male develops. When mature, the male possesses a cephalothorax; four free thoracic somites, each with a pair of thoracopods; and an unsegmented abdomen terminated by a pair of long caudal setae. In the development of the female, the tantulus swells behind the cephalic shield, the larval trunk is sloughed off, and the female body, still attached to the host by the head, becomes a simple saclike structure.

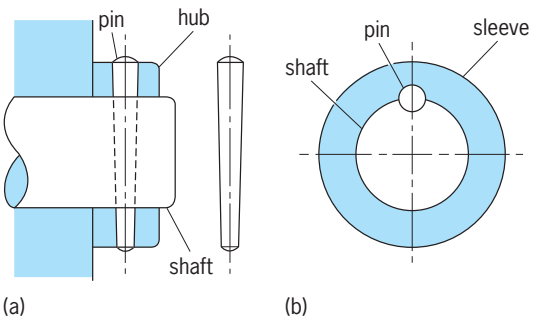
Phylogenetic relations of the Tantulocarida are not clearly understood. Although tantulocarids lack the antennules and antennae characteristic of Crustacea, two clusters of aesthetascs, or sensory hairs, in adult males are probably antennulary in origin. The basic tagmosis exhibited and male gonopore position suggest an affinity with some cirriped taxa. Some researchers include the Tantulocarida as a subclass of the Maxillopoda; others accept it as a distinct class. *See* CIRRIPEDIA; COPEPODA; CRUSTACEA; MAXILLOPODA.

Patsy A. McLaughlin

Bibliography. G. A. Boxshall and R. Huys, New tantulocarid, *Stygotantulus stocki*, parasitic on harpacticoid copepods, with an analysis of the phylogenetic relationships within the Maxillopoda, *J. Crust. Biol.*, 9:126–140, 1989; G. A. Boxshall and R. J. Lincoln, The life cycle of the Tantulocarida (Crustacea), *Phil. Trans. Roy. Soc. Lond.*, ser. B, 315:267–303, 1987; G. A. Boxshall and R. J. Lincoln, Tantulocarida, a new class of Crustacea ectoparastic on other crustaceans, *J. Crust. Biol.*, 3:1–16, 1983.

Taper pin

A tapered self-holding pin used to connect parts. Standard taper pins have a diametral taper $^1/_4$ in. in 12 in. (0.6 cm in 30 cm) and are driven in holes drilled and reamed to fit. The pins are made of soft steel or are cyanide-hardened. They are sometimes used to connect a hub or collar to a shaft (see **illus.**). Taper pins are frequently used to maintain the location of



Taper pins. (*a*) Connecting hub or collar to shaft. (*b*) Connecting sleeve to shaft.

one surface with respect to another. A disadvantage of the taper pin is that the holes must be drilled and reamed after assembly of the connected parts; hence they are not interchangeable. Paul H. Black

Taphonomy

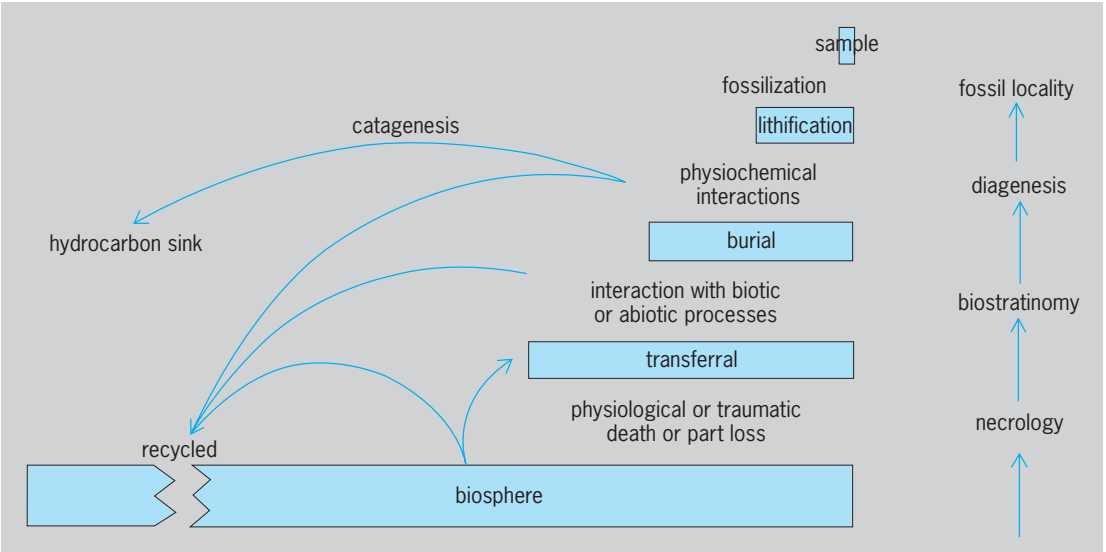
A subdiscipline of paleobiology that investigates the processes of preservation and their influence on information in the fossil record. Coined by J. A. Efremov in 1940, the term "taphonomy" involves all processes that affected the organism during its life, its transferral from the living world (biosphere) to the geological realm (lithosphere), and all physical and chemical interactions from the time of burial until collection. Besides the conspicuous characteristics of the preserved organism that can be seen easily—either morphological (external) and/or anatomical (internal) features—there are often less prominent details that record the fossil's history. Taphonomists are forensic scientists. By analyzing preserved details, paleontologists can understand an organism's mode of death or disarticulation; the biological processes that may have modified the remains before burial, including their use by hominids; the response of the organism or one of its parts to transport by animals, sediment, water, or wind; its residency time in a depositional setting before final entombment; and the alterations of tissues or skeletal parts within a wide range of chemical settings. The processes of fossilization appear to be environmentally site-specific, resulting in a mosaic of preservational traits in terrestrial and marine environments. Few fossil assemblages are exactly identical with regard to formative processes, but general patterns exist. An understanding of taphonomic assemblage features within its environmental context allows for a more accurate interpretation of the fossil record.

Early investigations focused on vertebrates, with a surge of studies encompassing all groups of organisms after 1970, and subsequent applications to forensics and archeology. Comparisons of the consistency of processes and assemblage attributes through geologic time have been made both in the rock record and in modern analog settings. These results have influenced the way in which paleontologists collect, analyze, and interpret their data. Before it was realized that all fossilized material in an assemblage offered information about preburial interactions, paleontologists expended their efforts on collecting the best-preserved (museum-quality) and most diagnostic fossils. Interpretations based upon these data sets often were heavily biased because only a small part of the fossil record was used. However, paleontologists have recognized that the inherent taphonomic biases in a fossil assemblage may not provide a complete and wholly accurate picture but, when taken into consideration, provide the most accurate means of assessing the record of life. *See* TRACE FOSSILS.

A very small percentage of organisms that have lived on Earth have had the potential to become fossils, because most biomass produced on a yearly basis is either ingested by organisms higher on the food chain or microbially degraded and recycled back into the environment. Ultimately, nearly all biomass is recycled (see **illustration**). The possibility of fossilization exists only when biomass accumulates under geologically rapid and site-specific physiochemical conditions. *See* BIOMASS.

Stages. Once fossilized, organic remains successfully have passed through several taphonomic stages, including necrology, biostratinomy, and diagenesis. However, remains may be affected at any stage, effectively removing them from the fossilization process.

Necrology is the death or loss of a part of an organism. Death is a prerequisite for most animals, but not



Subdisciplines of taphonomy and their relationships. (*After A. K. Behrensmeyer and S. M. Kidwell, Taphonomy's contribution to paleobiology, Paleobiology, 11:105–119, 1985; and A. Traverse, ed., Sedimentation of Organic Particles, Cambridge University Press, 1994*)

for plants that shed many different parts during their life cycle. Necrology may be induced by physiological means (old age, disease, or temporal or climatic shedding) or traumatic means (sudden catastrophic death or part loss in response to natural disruptions such as violent storms, volcaniclastic ash fall, mudflows, or flood events). The potential for any organic remains to successfully pass into the fossil record is dependent on the original biochemical composition. Organic remains of more resistant biochemicals or mineralized hard parts have a higher probability of fossilization than those of easily degraded compounds.

Biostratinomy focuses on the processes and interactions that follow necrology, until final burial. Biologic interactions include the effects of scavenging on carcasses, the use of discarded parts as domiciles, or borings in resistant structural parts. Abiotic processes include mechanical and physical alteration or breakdown under different transport conditions (fragmentation and rounding in river channels), orientation or concentration of resistant parts under varying hydrological regimes, or reexposure and reworking of previously buried remains in response to changing geological circumstances. This stage ends when organic remains are entombed within sediment, effectively isolating the material from the effects of biological degradation. The integrity of the buried part and resistance to diagenetic degradation will ultimately determine its residency time and potential for long-term burial and fossilization.

Diagenesis affects organic remains following burial and involves the physical (compaction) and chemical (cementation, recrystallization) changes in the sediment before, during, and after lithification (conversion of soft, unconsolidated sediments into hard rock). Mineralized skeletal elements of animals originally are composed of calcium carbonate (shell), calcium phosphate (bone and Precambrian shell), or silica (plankton test); resistant skeletal elements of plants are composed of biomacromolecules such as algaenans (microalgal cell walls), lignins (wood), cutins and cutans (cuticles), and sporopollenin (pollen and spores). Inorganic skeletal elements may be altered texturally and replaced mineralogically during diagenesis, whereas organic molecules undergo devolatization (loss of gases and water) and carbon concentration. Reactivity of pore waters with the organics soon after burial may result in pyritization (sulfur-reducing bacterial activity) or authigenic cementation (iron carbonate precipitation). Cellular voids may be infilled with crystallized hydrous minerals, resulting in the permineralization and preservation of internal features or cast replicas. If organics are subjected to high temperatures and pressures generated under deep burial conditions, catagenesis may result in the production of hydrocarbon fuels. *See* ORGANIC GEOCHEMISTRY; PALYNOLOGY.

Fidelity. The fidelity of any fossil assemblage is important because it provides information about the quantitative faithfulness of the record of body plans, age classes, species richness and abundances, trophic structures, and other aspects of the original biological community. Laboratory and field studies of modern communities have been used deductively to assess fossil assemblages in both terrestrial and marine realms. Information from live:dead studies, where the composition of a recent death assemblage is compared to the living community, shows a range of quantitative relationships. There is now a clear sense of how the construction and life habits of organisms and their postmortem environment in several systems combine to determine how well the death assemblage resembles the community from which it was sourced. There is tremendous variance in fidelity even among organisms that have a high probability of preservation, and this is related to the durability of their hard parts.

Fossil record and time resolution. The completeness of the stratigraphic record is restricted in time and space because of the assortment of Earth processes operating globally throughout geologic history. Taphonomic processes operating within this framework bias fossil-assemblage formation to specific depositional settings. By their nature, such assemblages generally reflect contributions from one primary systematic group: plants, invertebrates, or vertebrates. Fossils within these assemblages provide paleontological data concerning ontogenetic completeness (spectra of developmental stages within species), taxonomic completeness (the proportion of a higher taxon's component species), and ecosystem completeness (the share of an ecosystem's component species) in the fossil record at the time of accumulation.

The time represented within a fossil accumulation varies between depositional sites. Taphonomic assemblage characteristics allow paleontologists to estimate the relative or absolute time interval that has elapsed from biosphere to lithosphere transfer, which in turn provides a means of evaluating assemblage fidelity. Assemblages may represent the spectrum from the instantaneous picture (such as volcaniclastic ash burial) to accumulations that span thousands of years [for example, chenier plains (continuous ridges of beach material built upon swampy deposits) of shell debris]. Accumulations that represent short time intervals provide the highest-fidelity record, whereas time-averaged deposits represent contributions into a single bed or assemblage from successive populations. These latter assemblages preserve a mixture of noncontemporaneous remains that introduce another set of biases that must be recognized before credible interpretations can be formulated. Various statistical techniques have been devised to assess quantitative data with the prospect of deciphering patterns within time-averaged accumulations.

Relationship to paleoecology. Historical ecosystem responses to changing intrinsic short-term (local) or extrinsic long-term (global) factors are recorded in the fossil record at various resolutions. From these records it is possible to evaluate paleocommunity structure and composition and how these change through time in response to environmental

disruption, especially those associated with climate change. In turn, the dynamics associated with biodiversification of Earth, from individual assemblages to biomes, can be determined and rates of evolutionary events understood. These retrodictive (infering a past state from present data) patterns can be used as models to predict ecological reorganization under changing global conditions into the future. *See* FOSSIL; PALEOBOTANY; PALEOECOLOGY; PALEONTOLOGY.

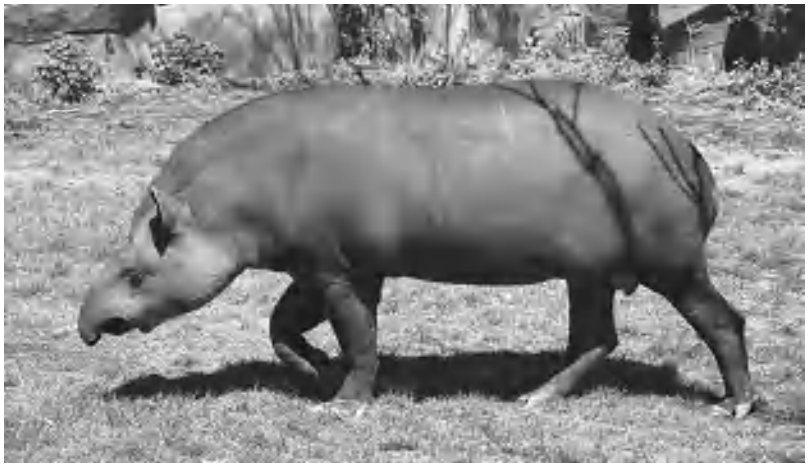
Robert A. Gastaldo

Bibliography. R. A. Allison and D. E. G. Briggs (eds.), *Taphonomy: Releasing the Data Locked in the Fossil Record*, 1991; A. K. Behrensmeyer et al., Taphonomy and paleobiology, *Paleobiology*, 26(4)[suppl.]:103–144, 2000; A. K. Behrensmeyer et al., *Terrestrial Ecosystems Through Time*, 1992; D. J. Bottjer et al. (eds.), *Exceptional Fossil Preservation*, 2002; C. E. Brett and S. E. Speyer, *Comparative Taphonomy: Pattern and Processes in Fossil Preservation*, 2005; S. K. Donovan (ed.), *The Processes of Fossilization*, 1991; C. T. Gee and R. A. Gastaldo (eds.), Plant Taphonomy Special Issue, *PALAIOS*, vol. 20, no. 5, 2005; J. P. Huntley and S. Stallibrass (eds.), *Taphonomy and Interpretation*, 2000; S. M. Kidwell and A. K. Behrensmeyer, *Taphonomic Approaches to Time Resolution in Fossil Assemblages: Short Courses in Paleontology*, vol. 6, 1993; V. A. Krasilov, *Paleoecology of Terrestrial Plants: Basic Principles and Techniques*, 1975; R. L. Lyman et al. (eds.), *Vertebrate Taphonomy*, 2001; R. E. Martin, *Taphonomy: A Process Approach*, 1999; A. Traverse (ed.), *Sedimentation of Organic Particles*, 1994.

Tapir

An odd-toed ungulate of the family Tapiridae. These animals, with four species, have a discontinuous distribution in South America and Asia and consequently have a theoretical importance biologically.

Tapirs are nocturnal, timid animals; they spend the daytime in dense thickets. They are generally silent and solitary except during the rutting season. Their diet consists of fruits and leaves, and they can cause considerable damage to cultivated crops. The tapir has a dental formula of I 3/3 C 1/1 Pm 3/3 M 3/3 for a total of 40 teeth. The forefoot has four toes, the hindfoot three. These animals have a prehensile trunklike muzzle, a short tail, and small eyes. The female gives birth to a single young after a gestation period of 13 months. The male leaves the female to care for the young alone.



Lowland tapir (*Tapirus terrestris*). (*Courtesy of Brent Huffman/Ultimate Ungulate Images*)

South American species differ from the Malayan species in the coloring and markings of the adults. The most common species is the Brazilian tapir (*Tapirus terrestris*) [see **illustration**]. This animal is found in the lowlands of northern South America and prefers marshy areas, where it can browse on water plants and other vegetation; it is a good swimmer. It requires a large intake of water. This species stands about 3 ft (0.9 m) high and is about 6 ft (1.8 m) long, the adult attaining a maximum weight of 650 lb (295 kg). The life-span is about 30 years.

The Malayan tapir (*T. indicus*) is common in the dense forests of Sumatra and Malaya, where it browses during the night on vegetation. The adults may weigh up to 400 lb (180 kg) and reach a length of 6 ft (1.8 m). Vision is relatively poor, but the olfactory and hearing senses are good. This species spends considerable time in the water and is a good swimmer. The coat of the New World species is dark, whereas that of the adult Malayan tapir is white from the shoulder to rump region and black elsewhere. *See* MAMMALIA; PERISSODACTYLA.

Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Tardigrada

A phylum of microscopic, bilaterally symmetrical invertebrates which are generally less than 1 mm in length. About 400 species are known. Commonly called water bears, bear animalcules, or urslets, they are worldwide in distribution and are found in all habitats.

Anatomy. The tardigrade body (**Fig. 1***a*) consists of an anterior prostomium and five segments. The mouth is located in the prostomium in a centroterminal position. A soft, nonchitinous cuticle surrounds the body and lines the fore- and hindgut. The cuticle may be smooth or sculptured and forms innervated cephalic appendages and spines on the trunk and legs. Four pairs of ventrolateral legs arise from the trunk and terminate in claws or other modified structures. A pair of oral glands and stylets are present which are partly cuticular and partly chitinous. Both the pharynx and esophagus are muscular structures. The digestive tract is tubular and more or less lobed due to the presence of diverticular dilations. In the heterotardigrades, separate anal and genital openings occur, while in the eutardigrades there is a single anogenital opening, the cloaca. The sexes are separate and the gonads are unpaired dorsal sacs with paired gonoducts in the male and in theory also in the female.

Histologically, the musculature of the body, legs, and mouth is of the smooth type.

The brain is a voluminous supraesophageal ganglion connected to a subesophageal ganglion by a pair of connectives. The brain has two or three inner and two outer lobules. Each of the latter, in most species except Arthrotardigrada, has one eyespot. This structure is an ocellus which is cup-shaped, and consists of one retinal and one pigmented cell which may be black, red, or pigment free. The ventral nerve cord, with four ventral ganglia, is a continuation of the subesophageal ganglion. Food storage cells float in the spacious body cavity, the coelom, which lacks a parietal or visceral peritoneum in the adult. Circulatory and respiratory structures are lacking.

These animals exhibit the phenomenon known as cell constancy. The number of epidermal cells is the same in all species of a genus. The pharynx has 27 epithelial and 24 myoepithelial cells. This number is constant in all genera except *Milnesium*, which has 24 and 39, respectively.

The lumen of the pharynx is triradial. Body muscles are metameric and comprise dorsal, ventral (Fig. 1*b*), and lateral groups. Each muscle is either a single fibrillar, uninucleated cell or a chain of such structures. The number of myocytes is constant in all individuals of the same species. The muscles retract and bend the body, while during relaxation the pressure of the fluid in the body cavity extends and stretches the animal. Storage cells of newly hatched *Echiniscoides* are constant in number and appressed to the epidermis; older cells multiply when dissociated from the epidermis.

Embryology. These animals lay eggs and development is direct. Embryonic development lasts 3–40 days, varying according to the species and surrounding temperature. Fertilization, as known in the Eutardigrada, may be external while eggs are laid in the old cuticle during molting, or internal within the ovary. Internal fertilization occurs after ejaculation of sperm into the female cloaca. Parthenogenesis has been observed in two species.

The oocytes mature with abortive oocytes serving as nurse cells. The ovarian endothelium produces the shell (chorion), which is smooth or has processes (ornamentations) that are of taxonomic importance (**Fig. 2***a*). The diploid number of chromosomes is 10–14 for species of Eutardigrada. Liberated eggs vary in diameter from 60 to 200 micrometers. The number of eggs in a single oviposit varies from 3 to 30. Cleavage is total and nearly equal, and is irregular because of asynchronous division of the blastomeres.

Gastrulation, by coeloblastic delamination, occurs at the 80–100-cell stage. Primordial germ cells are recognizable in the primary entoderm. The enterocoelous origin of the entomesoderm results in the formation of five pairs of coelomic sacs (Fig. 2*b*). The four anterior pairs give rise to muscles and storage cells, while the posterior pair unites to form the gonad from which gonoducts develop. The nervous system originates from the epidermis by delamination. The secondary entoderm becomes the midgut.

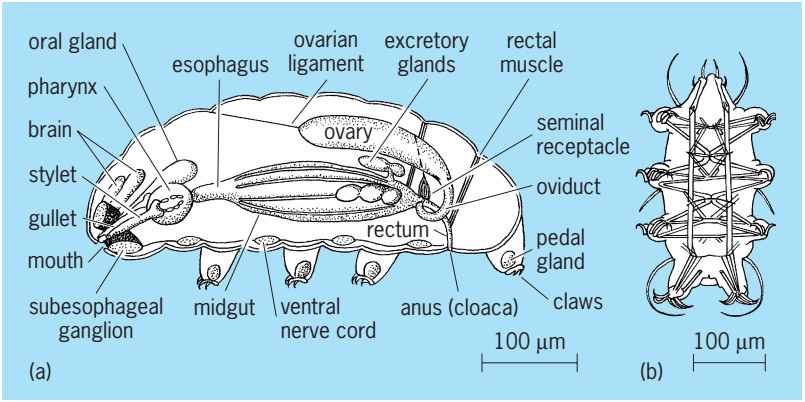


Fig. 1. Tardigrade morphology. (*a*) *Macrobiotus*, principal organs. (*b*) Muscles, ventral view.

Newly hatched tardigrades measure about 51 μm, while the largest adult is about 1200 μm. The very young animals swell as a result of water intake. Fewer claws and appendages are characteristic of the juvenile. These animals ordinarily grow to three to five times their initial size, and the immature tardigrades may lay eggs.

Molting. During their active life of 18 months, tardigrades molt about 12 times. They are unable to feed in the 5–10 days of molting; the buccal cavity requires at least 5 days for renewal (**Fig. 3**). Molting begins when the animal passes to the simplex stage, which is characterized by expulsion of the buccal

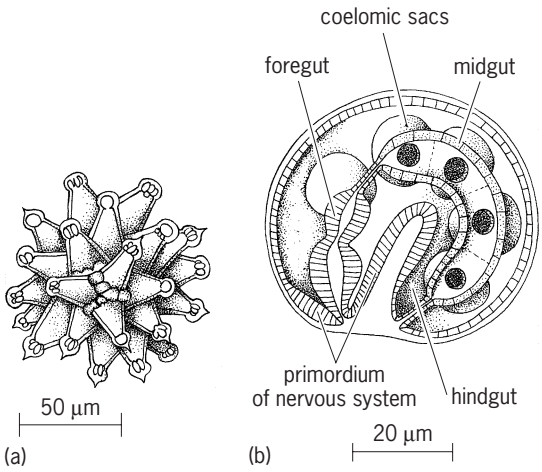


Fig. 2. Embryology of tardigrades. (*a*) Egg with processes. (*b*) Development of coelomic sacs.

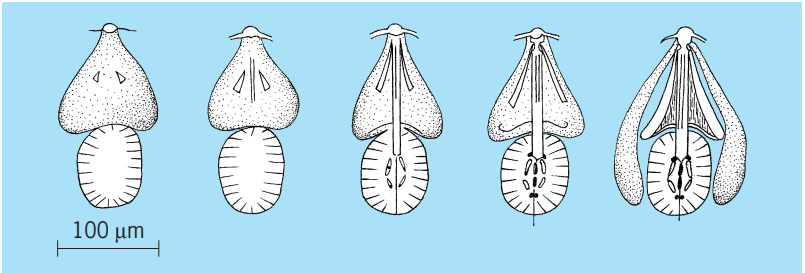


Fig. 3. Tardigrade buccal apparatus in 5 consecutive days after molting.

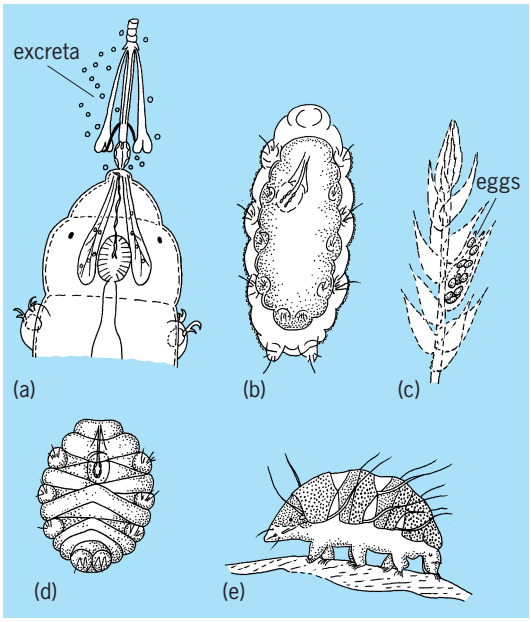


Fig. 4. Active and inactive forms of Tardigrada. (*a*) Expulsion of the buccal apparatus and excreta. (*b*) Cyst. (*c*) Eggs fastened to a moss leaf. (*d*) Barrel. (*e*) Scutechiniscid in the walking position.

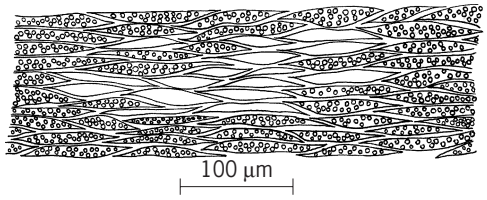


Fig. 5. Sixteen moss cells emptied by feeding action of tardigrade.

cuticular parts (**Fig. 4***a*). During molting, the epidermis secretes a new cuticle, and epidermal foot glands renew the toes or claws. The oral glands shorten and then regenerate stylets and bearers. In *Diphascon*, epithelia of the gullet, pharynx, and rectum replace corresponding cuticles. Under certain conditions, molting tardigrades, at least in the Eutardigrada, may remain in the old cuticle while the new one thickens to form a cyst (Fig. 4*b*). Hunger and perhaps unusual warmth seem to elicit encystment. Histolysis within a cyst appears improbable. It is possible that consumption of stored food causes the tardigrade to leave its cyst upon formation of the third cuticle. Molting is completed when the newly formed stylets pierce the old cuticle and allow the animal to escape. Duration of active life, without encystment, for as long as 30 months has been recorded for these animals.

Physiology. Tardigrada live as active forms, without encystment, only when surrounded by a pellicle of water. They are mainly herbivorous, and feed by piercing the wall of plant cells with their stylets. They ingest the contents of these cells (**Fig. 5**) by means of pumping pharynx. As a result, many ingested chloroplasts are fragmented. The foregut is acid in reaction (pH 4.4–5.2), while the midgut is alkaline (8.4–8.7). Digestion occurs only in the cells of the midgut (**Fig. 6**) which absorb the green-colored droplets containing plant matter. These cells contain digestive ferments. Fats, starch, and glycogen have been demonstrated in the storage cells.

Sometimes tardigrades, especially *Milnesium*, pierce the cuticle of rotifers, nematodes, other tardigrades, and nauplii to ingest their soft parts. Active animals can fast as long as 5 weeks. In active life in the Heterotardigrada, defecation occurs during molting with deposition of waste in the old cuticle (Fig. 4). In the Eutardigrada, it is independent of molting. Excretion occurs from the intestinal cells into the lumen of the gut, also from the oral glands at the beginning of a molt, and from the epidermis (part of pigments) during the formation of the new cuticle.

When the surrounding medium dries up, most tardigrades continue to live as inactive or anabiotic barrel-shaped structures called cysts without any protective cover (Fig. 4*d*). Desiccation begins when there is a loss of oxygen from the water. The animal responds by contraction and loss of body water. Dried eggs also survive. Moistened animals usually revive, but anabiosis and revival cannot be repeated indefinitely. Experimentally, the limit of a survival was 14 cycles of desiccation-hydration. Anabiosis can last for as long as $6^1/_2$ years, but minimal manifestations of muscular activity were observed when barrels which had remained dry for 120 years were moistened. By computing maximum periods of desiccation plus a total active life of 30 months, it is found that a tardigrade might live for 67 years, hence the name *Macrobiotus*. The limit of life during anabiosis does not depend on the amount of stored food, since those

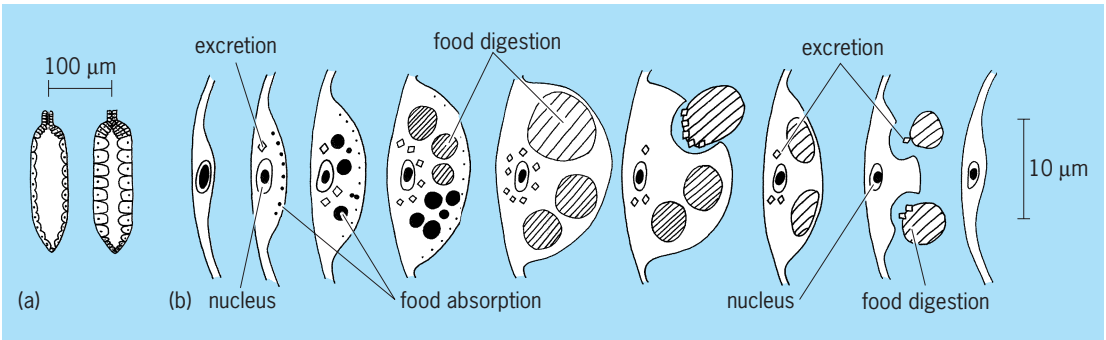


Fig. 6. Digestive activities in tardigrade. (*a*) Gut before and after absorption. (*b*) Digestive processes in gut cells.

with a reserve food supply do not differ in anabiotic capacity from those without one. These facts imply that the cause of death must be endogenous.

A 100-year-old discussion on anabiotic animals was characterized by two opposing views. One argument maintained that anaerobiosis in these animals was actually a very slow metabolism or "vita minima." The other claimed that life stopped without death supervening, and the organism was compared to a wound clock which had stopped ticking. A. Pigón and B. Weglarska settled the matter in 1953 by establishing a vita minima. They found that cysts were resistant to 198°F (92°C) heat in the air for 1 h. The cysts withstood cold at temperatures to −314°F (−192°C) for 20 months, and to −458°F (−272°C) for $8^1/_2$ h. In liquid air or helium used to study low-temperature effects, cysts are not completely dry, but contain water within and between their cells.

Tardigrades contract to the characteristic barrel shape when dissolved oxygen decreases in the water (Fig. 4*d*). They cannot maintain this state in water, however, and within 48 h they become maximally distended and immobile. Such asphyctic tardigrades may continue to live for 5 days and can revive if oxygen and food are made available. Animals emerging from anabiosis pass into a state of asphyxy, since the cells have lost their osmotic capacity. The arthrotardigrade, *Batillipes*, does not endure desiccation. For *Echiniscoides sigismundi*, about 42% of a population withstood dryness for 10 days, but none survived after 4 weeks. All were found to be alive in rainwater after 3 days, and in an asphyctic state.

Movement in the tardigrades is varied. Scutechiniscidae walk with legs which are nearly perpendicular, under the trunk (Fig. 4*e*), while *Tetrakentron* crawls. No species swims. *Echiniscoides* shows positive phototaxis, while the negative response of *Macrobiotus dispar* is a directed reaction. Eggs are frequently laid with positive thigmotaxis either between the leaflet and stalk of moss or in the shells of water fleas (Fig. 4*c*).

Ecology. Most species are widely distributed. Dissemination may be by wind, birds, and certain terrestrial animals which transport tardigrade eggs and barrels. Other moss dwellers, such as rhizopods, rotifers, and nematodes, are more easily transported than tardigrades. The tropics have few species, and Scutechiniscidae are rare on the Antarctic continent.

Echiniscoides sigismundi is the most euryokous species since it is worldwide in distribution and found both in littoral algae and at altitudes up to 3000 ft (1000 m) in the eastern Congo. In this environment tardigrade populations probably survive the dry periods. Population density varies considerably. In one sample of water containing *Enteromorpha*, 1 ml yielded 60 specimens.

A few scutechiniscids and several eutardigrades are limnetic organisms. The limnetic fauna is not sharply delimited from the terrestrial fauna, so that more limnetic than marine species are found, usually among algae, aquatic mosses, sand, and mud. Sandy beaches of soft-water lakes contain more specimens that do those of hard-water lakes. Some populations of permanently aquatic habitats cannot survive desiccation. Most species are eurythermal, having the ability to live through a wide range of temperature conditions. They are active between near-freezing temperature and 77–86°F (25–30°C). One known species is continuously exposed to temperatures of 104°F (40°C).

Most tardigrades are terrestrial. They are found among lichens, liverworts, densely growing soft-leaved mosses, and also in rather hard-leaved Pottiacea and Grimmiacea. They are rarer in some ferns, lycopods, and phanerogams such as *Sedum*, *Saxifraga*, and *Haastia*. The soil among fallen needles and leaves will yield tardigrades. As many as 22,000 animals have been recovered from 0.04 oz (1 g) of air-dried moss. An average population produces 780 lb (354 kg) of humus per hectare yearly, chiefly by means of fecal decomposition.

Fewer species are found in permanently wet or damp mosses than in land mosses. Oxygen supply is better in the land mosses, and desiccation excludes many rival organisms.

Tardigrades and their eggs are preyed upon chiefly by Amoebozoa, especially *Difflugia*. Nematodes attack the eutardigrades. Common pathogens of tardigrades are Phycomycetes (*Macrobiotophthora vimariensis*) and Microsporidia (*Pleistophora*). *See* CELL CONSTANCY; CLEAVAGE (DEVELOPMENTAL BIOLOGY); EUTARDIGRADA; HETEROTARDIGRADA; POLYCHAETA.

Eveline Marcus

Bibliography. K. W. Cooper, The first fossil tardigrade, *Psyche*, 71(2):41–48, 1964; E. Marcus, Tardigrada, in H. G. Bronn (ed.), *Klassen und Ordnungen des Tierreichs*, 1929; E. Marcus, Tardigrada, in F. Schulze and W. Kükenthal (eds.), *Das Tierreich*, pt. 66, 1936.

Tarragon

A herb of the genus *Artemesia* in the aster family (Asteraceae) that is used as a spice. Tarragon is often divided into French and Russian types. French tarragon can be distinguished by its highly aromatic leaves, low seed set or fertility, and compact growth habit. Due to its long history of vegetative propagation to preserve its fine scent characteristics, French tarragon has all but lost its ability to form viable seeds. Russian tarragon does produce seeds, but the plant is much lower in overall oil content and is not as fine-scented.

Tarragon is a perennial that grows to a height of 2–3 ft (60–120 cm) in one season, and then dies back to the ground with frost. In spring the plant sprouts from underground rhizomes and resumes its growth. The leaves contain 1.5–3% volatile oil with an odor similar to anise and chervil.

Tarragon requires special cultivation techniques because of its sterility. Root division, or more properly rhizome division, is the predominant method of propagation. This entails digging up the tarragon plants, then separating the rhizomes and replanting a larger area. Rooting stem tips and tissue culture

techniques have also been developed with some success.

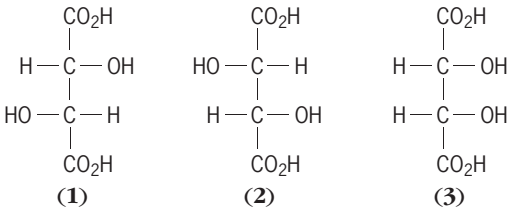
Tarragon is susceptible to a number of diseases, including wilts and dodder (a parasitic plant). Nematodes are a particular problem because they inhabit the rhizomes and are often passed on to new plantings. Weed control is usually accomplished by hand, but some chemical control is practiced outside of the United States. *See* ASTERALES; SPICE AND FLAVORING.

Seth Kirby

Bibliography. L. H. Bailey, *Manual of Cultivated Plants*, rev. ed., 1975; F. Rosengarten, *The Book of Spices*, 1969.

Tartaric acid

Any of the stereoisomeric forms of 2,3-dihydroxybutanedioic acid: L(+), D(−), and meso [(**1**), (**2**), and (**3**), respectively]. L(+)-Tartaric acid is present



in the juice of various fruits and is produced from grape juice as a by-product of the wine industry. The monopotassium salt precipitates in wine vats, and L(+)-tartaric acid is recovered from this residue. On heating in alkaline solution, the L(+) acid is converted to the racemic mixture of (**1**) and (**2**), plus a small amount of the meso acid (**3**).

Several systems have been used to designate the stereochemical configuration of the tartaric acid isomers, as shown in the **table**. The early names were based on the sign of optical rotation; the carbohydrate convention or the nomenclature of the International Union of Pure and Applied Chemistry (IUPAC system) are now preferred.

Tartaric acid has played a central role in the discovery of several landmark stereochemical phenomena. In 1848, L. Pasteur isolated enantiomers (**1**) and (**2**) by mechanical separation of hemihedral crystals of the racemic mixture. He also used tartaric acid and its salts to demonstrate a distinction between the meso isomer (**3**) and the racemic mixture (**1**) + (**2**), and between enantiomers and diastereoisomers in general. The difference in properties between (**1**) [or (**2**)] and the meso form (**3**) was later a key in establishing the relative configuration of the pentose and hexose sugars.

In 1951 the absolute configuration of L(+)-tartaric acid was determined to be structure (**1**) by anomalous x-ray scattering of the potassium rubidium salt. This experiment established the correctness of the arbitrary assignment of the configuration of the hexose sugars and of all other compounds whose configuration had been correlated with that of tartaric acid. *See* STEREOCHEMISTRY.

Both L(+)- and D(−)-tartaric acid and the esters are inexpensive compounds and are used as chiral auxiliary reagents in the oxidation of alkenes to enantiomerically pure epoxides. This method employs a hydroperoxide oxidant, titanium alkoxide catalyst, and L(+)- or D(−)-tartrate, and involves chirality transfer from the tartrate to the product. *See* ASYMMETRIC SYNTHESIS; EPOXIDE.

Tartaric acid has some use as an acidulant in foods and also as a chelating agent. Potassium hydrogen tartrate (cream of tartar) is an ingredient of baking powder. The potassium sodium salt, commonly called Rochelle salt, was the first compound used as a piezoelectric crystal. *See* CHELATION; PIEZOELECTRICITY.

James A. Moore

Stereoisomers of tartaric acid†

	(**1**)	(**2**)	(**3**)	(**1**) + (**2**)
Nomenclature				
Early usage	*d*-	*l*-	meso-	racemic acid
Carbohydrate convention	L(+)-	D(−)-	*meso*-	*rac*-
IUPAC	*2R, 3R*	*2S, 3S*	*2R, 3S*-	—
Chemical Abstracts	*R*-(R*R*)	*S*-(R*R*)	R*, S*	*R**, R*(±)
Properties				
Melting point °C (°F)	168–170 (334–338)	168–170 (334–338)	159 (318)	206 (403)
Optical rotation, [α]	+12°	−12°	0°	0°

†Structures (**1**)–(**3**) correspond to those displayed in the text.

Taste

Taste, or gustation, is one of the senses used to detect the chemical makeup of ingested food—that is, to establish its palatability and nutritional composition. Flavor is a complex amalgam of taste, olfaction, and other sensations, including those generated by mechanoreceptor and thermoreceptor sensory cells in the oral cavity. Olfactory (smell) sensory cells of the nose are particularly important in the perception of flavor. Taste sensory cells respond principally to the water-soluble chemical stimuli present in food, whereas olfactory sensory cells respond to volatile (airborne) compounds. *See* SENSATION.

Anatomy. The sensory organs of gustation are termed taste buds. In humans and most other mammals, taste buds are located on the tongue in the fungiform, foliate, and circumvallate papillae and in adjacent structures of the throat. There are

approximately 5000 taste buds in humans, although this number varies tremendously from person to person. Taste buds are goblet-shaped clusters of 50 to 100 long slender cells. Microvilli protrude from the apical (upper) end of sensory cells into shallow taste pores. Taste pores open onto the tongue surface and provide access to the sensory cells. Taste cells are modified epithelial cells that develop at the base of the taste bud, differentiate into functional sensory cells, and ultimately die. This cycle of differentiation, decline, and replacement, which lasts from 8 to 10 days, continues as long as the nerve supply to the taste buds remains intact. When the taste nerves are cut, taste buds disappear. Taste buds reappear when the peripheral taste nerve fibers regenerate and reinnervate the tongue epithelium. Individual sensory nerve fibers branch profusely within taste buds and make contacts (synapses) with taste bud sensory cells. Taste buds also contain supporting and developing taste cells (**Fig. 1**). *See* TONGUE.

The anterior two-thirds of the tongue is innervated by a sensory branch of the facial nerve (cranial nerve VII), the posterior tongue by the glossopharyngeal nerve (cranial nerve IX), and the throat and larynx by branches of the vagus nerve (cranial nerve X). These nerves carry information from touch, temperature, and pain sensory cells, as well as from taste buds. The taste fibers from the anterior part of the tongue branch from the lingual nerve in a smaller nerve called the chorda tympani (sensory branch of the facial nerve). The chorda tympani nerve traverses the eardrum en route to the brainstem. When the chorda tympani is damaged, as in the surgical removal of the tympanum, taste sensitivity is lost on the anterior two-thirds of the tongue on the same side as the surgery because the nerve is interrupted.

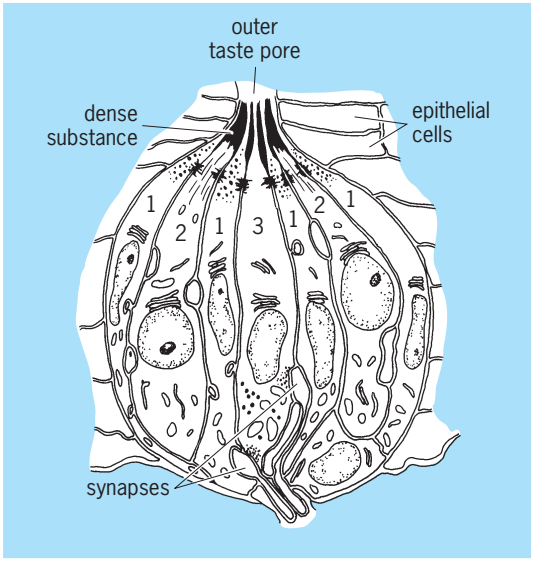


Fig. 1. Drawing of an electron microscopic section through a rabbit taste bud. A taste receptor cell (no. 3) is shown making two synaptic contacts with sensory nerve fibers. Several cells (no. 2) that make numerous physical contacts with nerve fibers also may be involved in impulse transmission. Darker cells (no. 1) are presumed to be supporting cells. (*After R. G. Murray, The ultrastructure of taste buds, in I. Friedmann, ed., The Ultrastructure of Sensory Organs, pp. 3–81, 1973*)

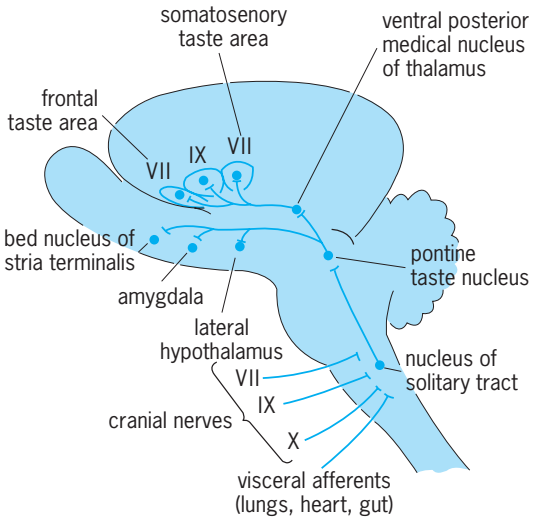


Fig. 2. Schematic drawing of the two neural pathways for taste in the rodent brain. (*After R. Norgren, Taste pathways to hypothalamus and amygdala, J. Comp. Neurol., 166:17–30, 1976*)

The cell bodies (somata) of taste nerve fibers (primary sensory afferent fibers) are clustered in several small ganglia nestled within the cranium but outside the brain itself. In the brain, taste fibers from the lingual, glossopharyngeal, and vagus nerves converge in the solitary tract and its nucleus in the medulla oblongata. These fibers terminate on neurons in a region contiguous to the area where touch and temperature sensory nerve fibers from the tongue also end. Second-order fibers from these neurons ascend to a cluster of small neurons located in the ventrobasal thalamus. (In rodents, another relay site—the pontine taste nucleus, or parabrachial nucleus—is interposed between the nucleus of the solitary tract and the thalamus; **Fig. 2**.) From the thalamus, neurons project to the lateral sensory cortex to two areas, one a mixed tactile-taste zone and the other largely a taste area, the primary gustatory cortex. This thalamocortical system is concerned largely with discrimination among different tastes. Outputs from the primary gustatory cortex terminate in the orbitofrontal cortex, where axons from the olfactory sensory cortex and other sensory cortical areas converge. This area of the brain appears to be dedicated to processing more complex perceptions such as flavor, where information from several sensory modalities is integrated. A second projection from the pontine taste area travels more ventrally, sending collaterals to the hypothalamus and other limbic structures to terminate ultimately in the central nucleus of the amygdala. This ventral pathway is presumed to mediate taste preferences, aversions, and hedonic responses (pleasantness or unpleasantness) of taste (Fig. 2). *See* BRAIN; NERVOUS SYSTEM (VERTEBRATE).

Qualities. The basic taste qualities experienced by humans include sweet, salty, sour, and bitter. (In some species, pure water also strongly stimulates taste bud cells.) A fifth taste, umami, is now

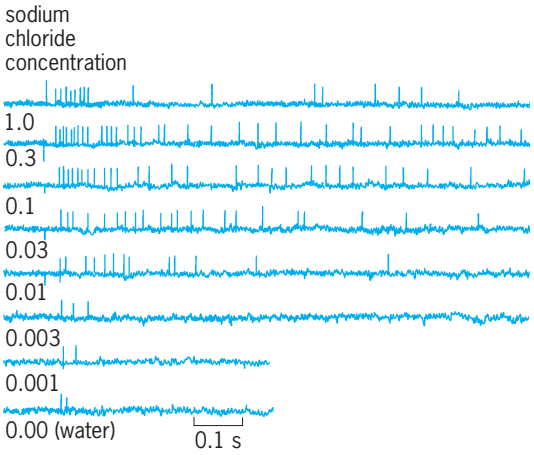


Fig. 3. Typical oscillographic record of impulses in a single taste sensory nerve fiber of a rat. (*After C. Pfaffmann, Gustatory nerve impulses in rat, cat and rabbit, J. Neurophysiol., 18:432–433, 1955*)

recognized by many as distinct from the other qualities. Umami is a Japanese term roughly translated as "good taste" and is approximated by the English term "savory." It refers to the taste of certain amino acids such as glutamate (as in monosodium glutamate) and certain monophosphate nucleotides. These compounds occur naturally in protein-rich foods, including meat, fish, cheese, and certain vegetables.

The middorsum (middle top portion) of the tongue surface is insensitive to all tastes. Only small differences, if any, exist for the taste qualities between different parts of the tongue (contrary to the "tongue maps" of taste sensitivities that are commonly published). Electrophysiological records from individual taste bud sensory cells show increased depolarization with an increased concentration of an effective stimulus. This leads to the release of neurotransmitters at synapses on the afferent taste nerves (Fig. 1) and an increase in frequency of nerve impulses therein (**Fig. 3**). No simple direct relationship exists between chemical stimuli and a particular taste quality except, perhaps, for sourness (acidity). Sourness is due to H^+ ions. The taste qualities of inorganic salts are complex, and sweet and bitter tastes are elicited by a wide variety of diverse chemicals.

Sensitivity. Specific sensitivity to particular chemicals is attributed to ion channels and to molecular receptors on the surface of sensory cells in taste buds, particularly on the apical membrane. Taste bud sensory cells are believed to have a spectrum of chemical sensitivities that differs from cell to cell. Some taste bud sensory cells probably respond to only one of the basic taste qualities, whereas others may be less discriminating and respond to multiple chemicals. This suggests that taste bud sensory cells possess more than one type of molecular receptor on their membranes. Alternatively, the surface membrane receptors themselves may be somewhat broadly sensitive to chemical stimuli. The sensitivity of individual taste bud sensory cells and their taste receptors is currently a topic of intense investigation.

More is known about the chemical sensitivity of sensory nerve fibers that innervate taste buds (chorda tympani, glossopharyngeal, and vagal). These fibers represent the sensory information output from taste buds. A fraction of the fibers are activated when the tongue is stimulated with a specific chemical or closely related compounds such as sucrose and fructose, and thus are narrowly receptive to chemical stimulation. However, even fibers with a broad sensitivity to diverse chemicals usually respond best (that is, give the highest frequency of action potentials) to a single type of chemical stimulus. The perception of taste quality (sweet, sour, salty, bitter, umami) is probably coded by the concurrent activation of some sensory fibers that are highly selective for certain chemical compounds as well as by other fibers that are less selective. That is, the pattern of activity within a heterogeneous population of sensory nerve fibers elicited by a particular taste stimulus determines taste perception (**Fig. 4**). No single homogeneous population of nerve fibers carries the information for a given taste quality. Various statistical procedures (such as factor analysis and multidimensional scaling) have been used to analyze these patterns of activity in taste fibers and have provided quantitative evidence for the existence of sweet, sour, salty, bitter, and umami taste qualities. *See* CHEMORECEPTION.

Sour. Studies by different investigators of human sour taste (acidity) reveal a wide range of thresholds for detection. For instance, hydrochloric acid (HCl)

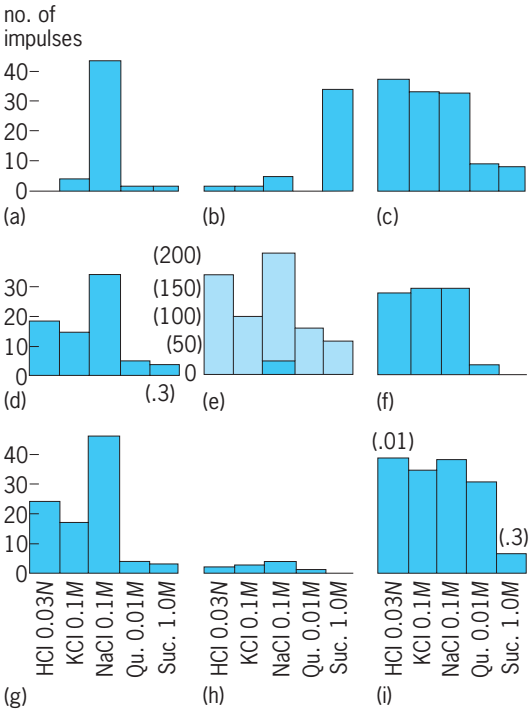


Fig. 4. Bar diagram of the different sensitivity patterns in (*a–i*) each of nine single sensory fibers from a rat. Dark color indicates number of impulses during first second of discharge to each of five test solutions shown along base: HCl (hydrochloric acid); KCl (potassium chloride); NaCl (sodium chloride); Qu (quinine); Suc (sucrose). Light color on fiber *e* shows relative amount of neural activity in total nerve (sum of all fibers) to same test solutions. (*After C. Pfaffmann, Gustatory nerve impulses in rat, cat and rabbit, J. Neurophysiol., 18:432–433, 1955*)

can be tasted at a concentration of about 0.0005 molar (*M*). Sourness increases with increases in hydrogen ion concentration, but weak organic acids are more sour than would be predicted from their hydrogen ion concentration. Increasing carbon chain length in the aliphatic acid series, for example, appears to enhance the potency of taste stimulation.

A number of explanations for how acid-sensitive taste sensory cells respond to hydrogen ions have been put forth. Possible mechanisms include the following:

1. Penetration of positively charged hydrogen ions (protons, H^+) through specific ion channels such as epithelial sodium channels (ENaC) in the apical membrane. The influx of H^+ would depolarize and excite the taste sensory cell. Taste bud sensory cells express ENaC channels, and pharmacological agents that block these channels reduce sour sensitivity in certain species, including humans.

2. Activation of cation-selective ion channel receptors by protons. Mammalian degenerin-1 (MDEG1) is one such receptor that has been identified in taste bud sensory cells in mice. Activation of MDEG1 by H^+ would allow the influx of cations such as sodium (Na^+), again depolarizing and exciting the taste sensory cell.

3. Block of potassium-selective ion channels in the apical membrane by H^+. Potassium ions (K^+) tend to leave the cell through these channels, keeping the sensory cell at a negative resting potential. Any agents, such as protons, that block these channels will result in depolarization and excitement of the taste sensory cell, as has been demonstrated in amphibian taste bud sensory cells.

4. Possible stimulation of specialized ion channels called hyperpolarization-activated, cyclic-nucleotide-gated (HCN) or pacemaker channels by H^+. Such channels have been identified in taste bud cells in some species. Their stimulation would lead to depolarization of the cell.

Perhaps some combination of all the above events underlies sourness, and sour taste mechanisms may differ from species to species. The final answer is not yet achieved. *See* BIOPOTENTIALS AND IONIC CURRENTS.

Salt. Most salts, including sodium chloride (NaCl), elicit other qualities, such as bitter or sour, in addition to salty. The dominant taste elicited by table salt is salty. Low-molecular-weight salts are predominantly salty, many higher-molecular-weight salts are bitter, and the salts of lead and beryllium are sweet. The median human threshold for detecting sodium chloride is approximately 0.01 *M*, but a wide range of values has been reported. Both the anion (for example, Cl^-) and cation (for example, Na^+) contribute to saltiness and to stimulus potency. According to one theory, salt-sensitive taste bud cells are stimulated by the influx of Na^+ through ENaC channels in their cell membrane. ENaC channels are found in other tissues in the body that are involved in Na^+ transport, such as renal tubules. Blocking ENaC channels with pharmacologic agents has been reported to reduce salt taste responses in some species. (However, the effects of blocking ENaC channels on human salt taste are controversial at present.) Further, H^+ also passes through ENaC channels, and agents that block ENaC channels alter sour taste. These observations raise the conundrum that the same mechanism (ENaC channels) appears to underlie both salt and sour tastes. However, there is no complete explanation yet for these observations. *See* SALT (FOOD).

Sweet. Sweet taste is associated largely with organic compounds such as sugars, alcohols, glycols, and sugar derivatives, with the exception of certain lead and beryllium salts. Sucrose thresholds for humans have a median value of 0.1 *M*, whereas the synthetic sweetener saccharin is 700 times more potent. The discovery that L-aspartyl-L-phenylalanine methyl ester (aspartame) is extremely sweet has led to further study of other dipeptides, but not all are sweet; some are sour, bitter, umami, or even tasteless. Some proteins have been discovered that elicit sweet taste with thresholds for humans in the 10^{-7} *M* concentration range. Slight changes in spatial arrangement render a sweet molecule tasteless or even bitter (**Fig. 5**). *See* ASPARTAME.

Studies of the specific blockage of sweet sensitivity by gymnemic acid, which does not affect salt, sour, or bitter, point to the existence of specific sweet receptors on the surface of taste bud sensory cells. One theory attributes the sweet taste to substances with a molecular system AH-B, where A and B are electronegative atoms separated by a distance of 0.3 nanometer. The AH moiety is a proton donor and the B moiety a proton acceptor. Sweet taste receptors are assumed to possess a complementary

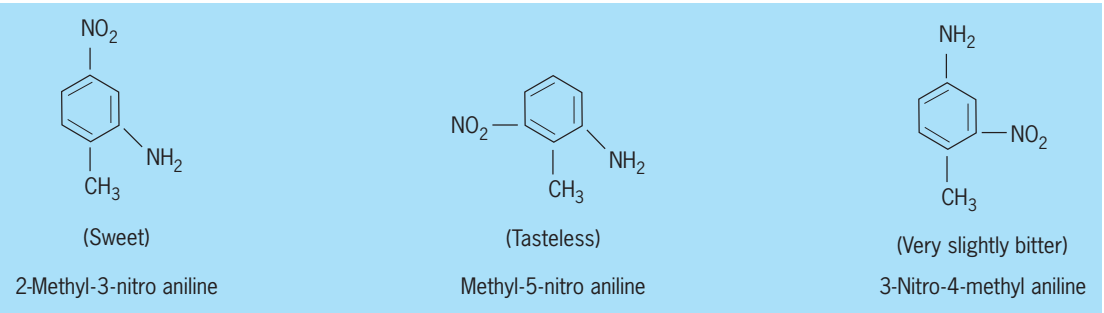


Fig. 5. Taste quality changes when spatial arrangements within the molecule change. (*After S. S. Stevens, ed., Handbook of Experimental Psychology, Wiley, 1951*)

AH-B system with which the stimulus system can interact, forming two intersystem hydrogen bonds. More recent formulations add a third molecular feature. Two candidate receptors for sweet taste have recently been cloned from human and rodent taste buds. These receptors (T3Rs) are G protein–coupled receptors, meaning that stimulation of the sweet receptor activates a cascade of intracellular enzymatic reactions involving G proteins. One such G protein that is associated with sweet taste is gustducin, which is found in taste bud cells and in some related cells. The net result of activating the sweet receptors on sweet-sensitive taste bud cells appears to be the block of K^+ channels, which depolarizes and excites taste bud sensory cells.

Bitter. Bitter is elicited by many chemical compounds and may be found in association with sweet and other taste qualities. An increase in the molecular weight of inorganic salts or in the length of a carbon chain of organic molecules may be associated with increased bitterness. Typical of substances with the bitter taste are the alkaloids such as quinine, caffeine, and strychnine, which are often toxic. A median threshold value for quinine has been cited at 0.000008 *M*. Taste blindness is an inherited inability to taste the bitterness of specific compounds such as phenyl thiocarbamide (PTC) and substances with the thiocarbamide group (N—C═S). About one-third of Caucasians are nontasters of PTC.

Two mechanisms have been proposed to explain how bitter-sensitive taste bud cells respond to stimulation. For example, certain bitter compounds directly block K^+ channels, a mechanism shared by sour and sweet sensing cells, and bitter taste receptors have been cloned and identified in human and rodent taste bud cells. Bitter taste receptors (T2Rs), like the candidate sweet taste receptors (T3Rs), are also G protein–coupled receptors. Unlike the candidate sweet receptors that have been identified to date, a large family of genes (about 30) encode bitter taste receptors. The G protein gustducin is also believed to interact with bitter receptors and activate an intracellular signaling cascade. Mice lacking gustducin show reduced taste responses to sweet and bitter chemicals. Thus, the determining factor for whether a given taste bud sensory cell responds to bitter or sweet is which receptor or receptors are expressed on the apical surface.

Umami. Umami, or savory taste, is stimulated by L-glutamate and certain nucleotide monophosphates such as 5′-inosine monophosphate and 5′-guanosine monophosphate. The threshold for human perception of L-glutamate is about 0.0008 *M*. Humans and other mammals find L-glutamate a preferred taste. Consequently, the natural glutamate content of prepared foods is often increased to enhance palatability by adding glutamate in the form of monosodium glutamate (MSG) or as protein hydrolysates. An important characteristic of umami is the synergistic interaction between glutamate and nucleotide monophosphates. Low concentrations of glutamate and nucleotide monophosphates interact to generate a robust savory taste.

A candidate G protein–coupled receptor for umami has been identified. It has been termed taste-mGluR4 and is a variant of the synaptic glutamate receptor mGluR4. The G proteins and enzymes stimulated by taste-mGluR4 have not yet been identified.

Adaptation. Prolonged exposure of the tongue to a taste stimulus decreases the response to that stimulus. This phenomenon, seen as a decrease in nerve activity or in the perceived intensity of the stimulus, is called adaptation. Adaptation by flowing taste solutions continuously over the tongue may lead to a rise in threshold or even complete disappearance of the particular taste sensation. Not only does that taste sensation disappear, but water and all concentrations of the stimulus at or below that used for adaptation show a contrasting taste that is often quite intense. Sodium chloride and sucrose adaptation tends to produce a bitter subadapting taste; adaptation to HCl and quinine produces a sweet subadapting taste. A salty taste is less common but may follow adaptation to compound sour-bittersweet substances such as urea. That bitter or sour taste can be masked by sweetening agents is well attested by the use of sweetners in coffee or in sour lemonade.

Nutrition and taste. The ability of an organism to select nutritious or necessary ingredients of the diet by taste can be demonstrated with the self-selection technique, in which an animal is given free choice among individual containers with necessary nutrients in pure form. After certain physiological stresses, like glandular imbalances, the selection of the various nutritive agents often shows compensatory changes. A severely salt-deficient rat (adrenalectomized) may show a significant increase in the intake of sodium chloride sufficient to counteract the usually fatal outcome in the absence of salt replacement therapy. Similar effects have been noted in children; but in adults, food habits, cultural conditioning, learning, and other complex psychological factors play significant enough roles in food acceptance to override the physiological factors controlling behavior. Nonetheless, food palatability, as determined by taste, olfaction, and other sensory effects, has such a profound effect on food acceptance that a substantial applied science of flavor technology has developed. *See* CHEMICAL SENSES.

Carl Pfaffmann; Nirupa Chaudhari; Stephen Roper

Bibliography. T. E. Finger, W. L. Silver, and D. Restrepo, *Neurobiology of Taste and Smell*, 2d ed., Wiley-Liss, 2000; T. A. Gilbertson, S. Damak, and R. F. Margolskee, The molecular physiology of taste transduction, *Curr. Opin. Neurobiol.*, 10(4):519–527, 2000; B. Lindemann, Taste reception, *Physiol. Rev.*, 76(3):718–766, 1996; S. D. Roper, Gustatory and Olfactory Sensory Transduction, Ch. 49 in N. Sperkelakis (ed.), *Cell Physiology Sourcebook: A Molecular Approach*, 3d ed., Academic Press, San Diego, 2001; S. A. Simon and S. D. Roper, *Mechanisms of Taste Transduction*,

CRC Press, Boca Raton, FL, 1993; D. V. Smith and R. F. Margolskee, Making sense of taste, *Sci. Amer.*, 284(3):32–39, 2001.

Taurus

The Bull, a large zodiacal northern constellation prominent in the evening winter sky (see **illustration**). Taurus is usually said to be the form assumed by Zeus in Greek mythology in order to carry off the princess Europa. *See* ZODIAC.

The head of the bull is marked by the Hyades, a large V-shaped open cluster, against which is projected the red giant star Aldebaran, though it is only half as far away. From the V near Aldebaran, two bright stars farther out mark the points of Taurus's long horns. Near one of those stars, Zeta, is the Crab Nebula (M1, the first object in Messier's eighteenth-century catalogue, though it had been previously discovered by John Bevis), the remnant of a supernova whose light reached Earth in AD 1054, and which is detectable all across the spectrum of gamma rays, x-rays, ultraviolet, light, infrared, and radio waves. *See* ALDEBARAN; CRAB NEBULA; HYADES; MESSIER CATALOG.

The Pleiades, the seven sisters of Greek mythology, are a star cluster riding on Taurus's back. Only six are visible to the unaided eye, though binoculars or telescopes show dozens or hundreds more stars in the cluster. The Pleiades (M45, the 45th object in Messier's catalogue) have the shape of a small dipper, and long exposures show dust surrounding many of the stars, which are hot and therefore bluish. The dust reflects the starlight toward us. *See* PLEIADES.

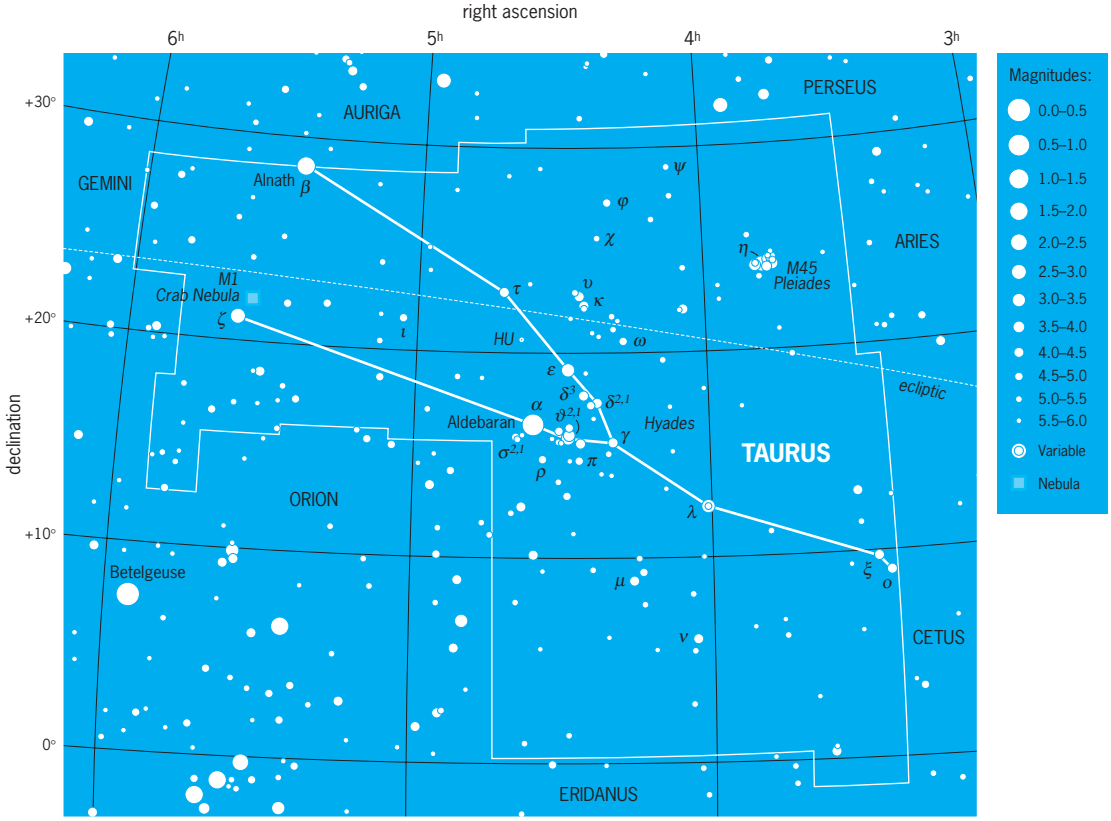
The modern boundaries of the 88 constellations, including this one, were defined by the International Astronomical Union in 1928. *See* CONSTELLATION.

Jay M. Pasachoff

Tautomerism

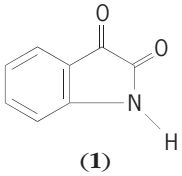
The reversible interconversion of structural isomers of organic chemical compounds. Such interconversions usually involve transfer of a proton (prototropy), but anionotropic (allylic, Wagner-Meerwein) rearrangements may be reversible and so be classed as tautomeric interconversions.

Lactam-lactim tautomerism. A cyclic system containing the grouping —CONH—is called a lactam, and the isomeric form, —COH═N—, a lactim. These terms have been extended to include the same structures in open-chain compounds when considering the shift of the hydrogen from nitrogen to oxygen.

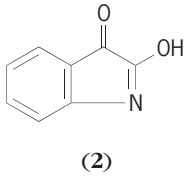


Modern boundaries of the constellation Taurus, the Bull. The celestial equator is 0° of declination, which corresponds to celestial latitude. Right ascension corresponds to celestial longitude, with each hour of right ascension representing 15° of arc. Apparent brightness of stars is shown with dot sizes to illustrate the magnitude scale, where the brightest stars in the sky are 0th magnitude or brighter and the faintest stars that can be seen with the unaided eye at a dark site are 6th magnitude. (*Wil Tirion*)

Isatin (**1**) appears to react in either the lactam (**1**) or the lactim (**2**) structure. Thus, there is a



(**1**)

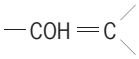


(**2**)

precedent for proving the existence of lactim-lactam tautomerism—that chemical behavior may be inferred from the structure of a reaction product. Spectroscopic techniques are more reliable, and it is often possible to determine from the absorption spectrum whether a given substance has either one structure or both. *See* LACTAM.

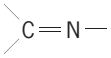
Keto-enol tautomerism. The molecular grouping

may in certain substances exist partly or wholly as



The former constitutes the keto form and the latter the enol form. Kurt Meyer first studied the keto ($CH_3COCH_2CO_2C_2H_5$) and enol ($CH_3OHC{=}CHCO_2C_2H_5$) forms of ethyl acetoacetate, and recognized them respectively by reactions specific for the carbonyl group and the carbon-carbon double bond. Both forms may be obtained in relatively pure condition, the former by freezing it out of the mixture and the latter by slowly distilling the mixture in quartz apparatus. However, each is slowly converted into the equilibrium mixture of the two. Extensive chemical and spectroscopic studies showed that the enol content of such an equilibrium mixture is a function of the physical state of any given substance. The gas phase or solution in a nonpolar solvent (hexane) favors the enol form, whereas more polar solvents (chloroform, alcohols) repress its formation.

The existence of an enol in an acyclic system requires that a second carbonyl group, or its equivalent, for example,

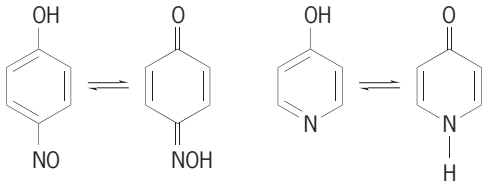


be attached to the same

as an aldehyde or ketone carbonyl. Thus, ethyl acetoacetate tautomerized demonstrably, but ethyl malonate ($C_2H_5O_2CCH_2CO_2C_2H_5$) does not. Occasionally an enol form exists, these requirements notwithstanding. For example, ethyl pyruvate is partially enolized ($CH_3COCO_2C_2H_5 \rightleftharpoons CH_2{=}COHCO_2C_2H_5$), and α-hydroxy ketones (or aldehydes) exhibit the characteristics of the tautomeric enediols, for example, benzoin [reaction (1)].

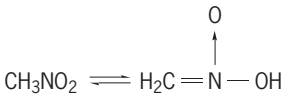
$$C_6H_5COCHOHC_6H_5 \rightleftharpoons C_6H_5OHC{=}COHC_6H_5 \qquad (1)$$

Where the enol form includes an aromatic ring such as phenol, the existence of the keto form is often not demonstrable, although in some substances such as 4-nitrosophenol (**3**) and 4-hydroxypyridine (**4**) there may be either chemical



(**3**) (**4**)

or spectroscopic evidence for both forms. Closely related to keto-enol tautomerism is the prototropic interconversion of nitro and aci forms of aliphatic nitro compounds such as nitromethane (**5**).



(**5**)

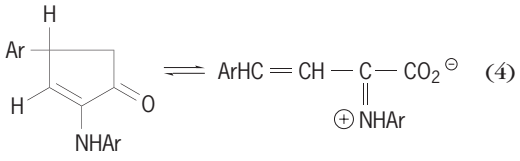
Ring-chain tautomerism. An acyclic hydroxyaldehyde may exist in equilibrium with its cyclic hemiacetal. The failure of glucose to form a normal acetal with an alcohol and the production of two isomeric glucosides instead led to the postulate that carbohydrates exist principally as inner or cyclic hemiacetals in equilibrium with only enough free aldehyde to permit typical aldehyde reactions with reagents that either oxidize the carbonyl group or form derivatives that effectively remove it from the equilibrium [reaction (2)]. The glycosides are formed by the elimination of water between the hydroxyl (OH) derived by hemiacetal formation and an alcohol, with two structures being possible, and the hydroxyl lying above or below the hetero ring. *See* GLUCOSE; GLYCOSIDE.

$$CH_2OHCHOHCHOHCHOHCHOHCHO \rightleftharpoons CH_2OH\,CHCHOHCHOHCHOHCHOH \quad (2)$$

In general, tautomeric forms exist in substances possessing functional groups that can interact additively and that are so placed that intramolecular reaction leads to a stable cyclic system. The cyclic form usually predominates (especially if it contains five or six members).

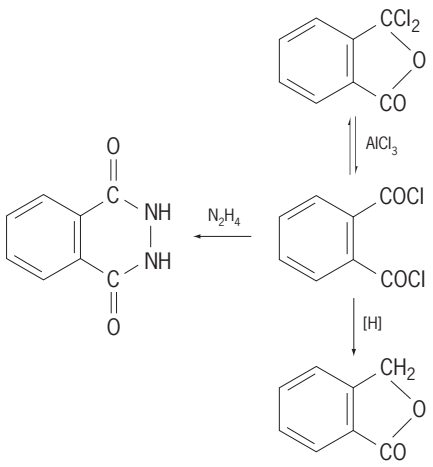
Certain alkenic acids are tautomeric with their lactones [reaction (3)], and this is called lacto-enoic tautomerism. However, lacto-enoic tautomerism not involving a prototropic shift has also been observed [reaction (4) Ar = aromatic group].

$$RCH{=}CHCH_2CO_2H \rightleftharpoons RCHCH_2CO \quad (3)$$



(4)

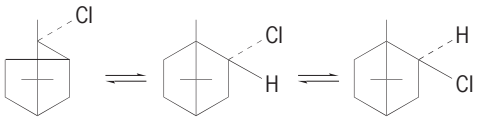
Still another type of ring-chain isomerism not involving a prototropic shift is demonstrable by the reactions of phthaloyl chloride [reaction (5)].



(5)

The latter type of ring-chain tautomerism is closely related to anionotropic rearrangements such as the allylic [reaction (6)] and Wagner-Meerwein [reaction (7)], which may thus be considered examples

$$RCHClHC{=}CH_2 \rightleftharpoons RHC{=}CHCH_2Cl \qquad (6)$$



(7)

of anionotropic tautomerism. *See* MOLECULAR ISOMERISM.

Wyman R. Vaughan

Valence tautomerism. The term valence tautomerism applies generally to relatively low-energy processes in which concerted, synchronous σ- or π-electron shifts occur over several atoms. Neither diradical nor dipolar species intervene in such processes, nor do atom or functional group migrations occur in conjunction with these electron shifts. The phenomenon of valence tautomerism is distinct from and should not be confused with resonance. In the former case, individual tautomers represent energy minima that interconvert via a transition structure that is an energy maximum on the potential surface that leads from one tautomer to the other. When this concerted electron-shift process occurs in systems for which the energy barrier to interconversion of valence tautomers is relatively high, this process frequently is referred to as valence isomerization. *See* CHEMICAL BONDING.

Alan P. Marchano

Bibliography. Z. Rappoport (ed.), *The Chemistry of Enols*, 1990; R. E. Valters, and W. Flitsch, *Ring-Chain Tautomerism*, 1985.

Taxis

A mechanism of orientation by means of which an animal moves in a direction related to a source of stimulation. There exists a widely accepted terminology in which the nature of the stimulus is indicated by a prefix such as phototaxis, chemotaxis, geotaxis (gravity), thigmotaxis (contact), rheotaxis (water current), and anemotaxis (air current). The directions toward or away from the stimulus are expressed as positive or negative, respectively. Finally, the sensory and locomotory mechanisms by means of which the orientation is achieved are denoted by a second type of prefix forming a compound noun with taxis. Positive phototropotaxis thus describes a mechanism by means of which an animal carries out a directed movement toward a source of light along a path which permits the animal's paired eyes to receive equal intensities of light throughout the movement. The following are examples of various types of taxes.

Klinotaxis. A well-analyzed case of this type of taxis is the way in which a fly maggot moves away from the light immediately before pupating in a dark and sheltered place. When such a maggot is exposed to a horizontal beam of light above the substrate, it moves along a fairly straight path away from the light in the direction of the beam. In doing so it waves its front end from side to side, exposing alternately the left and right lateral aspects of its front end to the light shining from behind. As long as the light intensity falling on the light-sensitive surfaces on either side remains equal in subsequent exposures, the animal follows a straight path away from the light source. This may be explained by the hypothesis that the extent of the swing toward one side is a function of the light intensity falling on the other. If the animal starts its course from a position at an angle with the light beam, the differential light intensities falling on its anterior flanks automatically steer it into a path curving into line with the beam. The prefix "klino" in this case denotes that "exploratory" side-to-side bending of the body brings about orientation by directed, though waving movement, related in its direction to the stimulus.

Tropotaxis. Tropotaxis is a term closely related to Jacques Loeb's original notion of tropism, which has now become restricted to the description of orientation phenomena in sessile organisms. The essential point is the unwavering turn of the organism into the stimulus direction by means of innate reflex mechanisms linking bilaterally symmetrical receptors with the organs of locomotion. Paths toward or away from a single source are straight. In the case of two or more sources, they run along the resultant of incident intensities. After unilateral receptor loss, tropotactic steering leads, in a uniform field of illumination, to continued "circus movement" away from or toward the injured side in the case of positive or negative taxis, respectively.

Telotaxis. Whereas bilateral intensity balance on receptors is essential in klinotaxis and tropotaxis, orientation in telotaxis occurs, as it were, by the orientation of one or the other of two bilaterally symmetrical receptors toward the stimulus source. Of a number of simultaneously offered stimuli, all but one may be "ignored" at any given instant by means of built-in switch mechanisms based on inhibition or

block. Unilateral sense organ loss does not lead to circus movement.

A special case of telotaxis is the light-compass reaction. In this the animal moves at a temporarily fixed angle with respect to the stimulus direction. The angle can be changed at will. The term pharotaxis has been suggested to describe an orientation toward a progressively changing direction of stimulus, such as the change of distribution of polarization of the light of the sky during the course of the day (navigation by honeybees and other arthropods).

The description of taxes given so far is based on concepts formulated by Alfred Kühn and after him by G. Fraenkel and D. Gunn. They apply largely to mechanisms of orientation found in small-brain animals which rely in their basic orientation on a relatively small number of innate responses. These are characterized by their stereotyped and rather inflexible nature. O. Koehler and with him a number of students of animal behavior have attempted to make the taxis concept part of the more complex and plastic behavior of vertebrates, including humans.

Thus Koehler's definition of a taxis is as follows: Every purely reflexive and every voluntary act begins with certain postures and movements which orient the animal's head, limbs, and body with reference to the direction of the eliciting stimulus or of the final goal of the movement. These intentional postures or movements of postural adjustment are to be considered the taxis components of the animal's act. Even in humans, the turning of eyes or head into the direction of an object of interest or desire is held to fall within this definition. *See* POSTURAL EQUILIBRIUM; PLANT MOVEMENTS.

Otto E. Lowenstein

Taxodonta

An order of bivalves in subclass Lamellibranchia whose hinge dentition is characterized by a series of numerous similar alternating teeth and sockets. Typical genera of so-called ark shells are *Arca*, *Barbatia*, and *Anadara*, all with heavy coarse or matted periostracum, and mostly with strong radial ribs on globose or purse-shaped shells. The filibranch gills are enlarged to form the four double lamellae of the filter-feeding organs as in all true lamellibranchs, and the labial palps form complex sorting surfaces.

Ark-shell bivalves are probably distantly related to the true marine mussels (Mytilacea), with which they share certain features of gill ciliation and byssal organization. Confusion arises because many primitive protobranch bivalves (totally unrelated and now placed in the subclass Protobranchia), using palp proboscides in feeding and lacking lamellate gills, also have taxodont hinge dentition. Formerly the name Taxodonta was used to designate a subclass of bivalves consisting of a mixed group of lamellibranch ark shells and primitive protobranchs. *See* BIVALVIA; LAMELLIBRANCHIA; MOLLUSCA; PROTOBRANCHIA.

W. D. Russell-Hunter

Taxonomic categories

Any one of a number of formal ranks used for organisms in a Linnaean classification. Biological classifications are orderly arrangements of organisms in which the order specifies some relationship. Taxonomic classifications are usually hierarchical and comprise nested groups of organisms. The actual groups are termed taxa. In the hierarchy, a higher taxon may include one or more lower taxa, and as a result the relationships among taxa are expressed as a divergent hierarchy that is formally represented by tree diagrams. In Linnaean classifications, taxonomic categories are devices that provide structure to the hierarchy of taxa without the use of tree diagrams. By agreement, there is a hierarchy of categorical ranks for each major group of organisms, beginning with the categories of highest rank and ending with categories of lowest rank, and while it is not necessary to use all the available categories, they must be used in the correct order.

Categories. Categories commonly used in botanical and zoological classifications are listed below, from highest to lowest rank:

Botanical categories	*Zoological categories*
Divisio	Phylum
Classis	Class
Ordo	Order
Familia	Family
Genus	Genus
Species	Species

Conceptually, the hierarchy of categories is different than the hierarchy of taxa. For example, the taxon Cnidaria, which is ranked as a phylum, includes the classes Anthozoa (anemones), Scyphozoa (jellyfishes), and Hydrozoa (hydras). Cnidaria is a particular and concrete group that is composed of parts. Anthozoa is part of, and included in, Cnidaria. However, categorical ranks are quite different. The category "class'' is not part of, nor included in, the category "phylum.'' Rather, the category "class'' is a shelf in the hierarchy, a roadmark of relative position. There are many animal taxa ranked as classes, but there is only one "class'' in the Linnaean hierarchy. This is an important strength of the system because it provides a way to navigate through a classification while keeping track of relative hierarchical levels with only a few ranks for a great number of organisms.

When Linnaeus invented his categories, there were only class, order, family, genus, and species. These were sufficient to serve the needs of biological diversity in the late eighteenth century, but were quite insufficient to classify the increasing number of species discovered since 1758. As a result, additional categorical levels have been, and continue to be, created. These categories may use prefixes, such as super- and sub-, as well as new basic levels such as tribe. An example of a modern expanded botanical hierarchy of ranks between family and species is:

Familia
Subfamilia
Tribus
Subtribus
Genus
Subgenus
Sectio
Subsectio
Series
Subseries
Species

Conceptual issues. Linnaean categories are the traditional devices used to navigate the hierarchy of taxa. But categories are only conventions, and alternative logical systems are frequently advocated. Taxonomic categories could be dispensed with, and relative position could be indicated by using only indentation. Taxa indented at the same level would have equivalent places in the hierarchy. But this method is cumbersome, and especially so when the classification occupies many pages and there are many subordinate taxa separating two closely related taxa of the same rank. Actual physical measurements would be needed using some arbitrary standard. Alternatively, a numbering scheme for ranks might be substituted. Taxa with equivalent codes would have the same position in the classification. While logical, this system is cumbersome. The Linnaean system of ranks seems more familiar, even to those who are not taxonomists.

Exactly what a categorical rank is supposed to indicate is also controversial. Evolutionary taxonomists adjust the ranks of taxa to reflect subjective judgments as to the significance of one taxon compared to another. For example, birds are accorded the rank of class while their closest living relatives, crocodiles, are considered an order in another class (Reptilia). Thus, the classifications of evolutionary taxonomists do not always reflect common ancestry relationships. Phylogenetic systematists (cladists) rank taxa strictly by their common ancestry relationships. As a result, sister taxa, that is, taxa of the same age or origin that uniquely share a common ancestor, always have the same rank.

The last controversy concerns the comparability of taxa of similar rank. Sister taxa are biologically comparable, but taxa in different groups that are ranked at the same level have no necessary biological equivalency. In preevolutionary concepts, rank could be correlated with increasing perfection. Thus, a family of butterflies might be comparable to a family of buttercups. Modern taxonomists acknowledge that categorical rank is a poor to misleading criterion on which to base comparative biology and that strict comparisons must either be restricted to sister groups or to groups where the phylogeny can be consulted to determine relative comparability. *See* SYSTEMATICS; CLASSIFICATION, BIOLOGICAL; PLANT TAXONOMY; SPECIES CONCEPT; TAXONOMIC CATEGORIES; ZOOLOGICAL NOMENCLATURE. Edward O. Wiley

Bibliography. R. E. Blackwelder, *Taxonomy: A Text and Reference Book*, 1967; P. H. Davis and V. H. Heywood, *Principles of Angiosperm Taxonomy*, 1963; E. Mayr and P. D. Ashlock, *Principles of Systematic Zoology*, 2d ed., 1991; E. O. Wiley, *Phylogenetics, The Theory and Practice of Phyolgenetic Systematics*, 1981.

Taxonomy

The human activity of naming organisms and organizing these names according to a given criterion. In biology, taxonomy has evolved into formal rules and systems to classify organisms. It is part of the broader discipline of systematics, a discipline that covers all of comparative biology, such as comparative morphology, genomics, and biogeography. Systematists who develop formal biological classifications and publish taxonomic revisions and keys are practicing taxonomy.

Taxonomy is one of the oldest activities of humans, originating with language. Such "folk" taxonomies survive today, even in developed cultures, in the form of common names for plants and animals. Many of these folk taxonomies are organized around the function of the organism or the importance of the organism to the society. For example, we might recognize groups of predators and prey, those who hunt us and those we hunt. Among plants, we might recognize edible and poisonous plants. Peoples in hunter-gatherer societies are quite good at recognizing the same species of plants and animals recognized by taxonomists. The dividing line between such folk taxonomies and scientific taxonomies can be thin, but in general, scientific taxonomies are attempts to organize diversity in a manner that humans think exists in nature in order to reflect some general scientific principle. Ultimately, this takes the form of a classification, an attempt to organize the world around us using names. Taxonomies are not restricted to biology; astronomers place stars into classes that reflect their mass and brightness, two characteristics that predict how any individual star formed and how it will behave before its eventual death. The classification has predictive power that reflects what we think we know about nature. In a similar manner, modern biologists have developed taxonomies that reflect evolutionary descent, which opens the door to all sorts of comparative predictions ranging from the ecologies of species to the potential for some species to be pathogens. *See* ANIMAL SYSTEMATICS; CLASSIFICATION, BIOLOGICAL; PLANT TAXONOMY.

History. The ancient Greeks applied science as an exercise in explaining the world around them in natural rather than supernatural terms, and so the first scientific taxonomies were Greek. Aristotle (384–322 BCE) classified animals with similar characteristics. For example, he recognized animals with blood, what we now call vertebrates, and grouped them into five genera: fishes, whales, mammals (quadrupeds who gave birth to live young), reptiles and amphibians (quadrupeds that lay eggs), and birds. He also recognized animals without blood (invertebrates) and distinguished between shelled

animals (such as clams), cephalopods such as octopi and squids, insects, crustaceans, and plantlike animals such as jellyfishes. The larger groups were called genera and the smaller were called species.

Aristotle's work was amazingly detailed, even if he did not recognize whales as mammals. A student of Aristotle, Theophrastus (ca. 370-285 BCE), produced one of the first plant classifications, distinguishing plants on the basis of their form (trees, scrubs, herbs, etc.) and did fundamental work on plant morphology. Because plants were important in medicine, this work was continued, and by the first century CE the Roman physician Dioscorides had cataloged some 600 medicinal plants. His work was copied for physicians for more than a millennium. Except for the work of Albertus Magnus (1193–1280), who recognized vascular and nonvascular plants as well as monocots and dicots, little progress was made in plant taxonomy until the dawn of the Renaissance. Aristotle's work preserved by Arab culture through the translation of Averroes of Cordoba (Ibn Rushd; 1126–1198) was the accepted work on animals until the sixteenth century.

World exploration by Europeans coupled with the rise of their science brought renewed interest in animal and plant diversity as newly discovered specimens flooded European centers of learning. By 1555 the Swedish naturalist Konrad Gessner published the first of his *Historiae Animalium* (1555–1558), considered by many as the foundation for modern zoology. Andrea Cesalpino, an Italian physician, published *De plantis libri XVI* in 1583. His classification of plants by their seed types, fruits, and form rather than providing a simple list of medicinal plants laid the foundation for modern plant classification. Between these early works and the standardization of taxonomic nomenclature were important works on plants and animals by such workers as John Ray (1627–1705). Modern taxonomy began to emerge. The simple "genus" and "species" dichotomy was replaced with something resembling modern categories by Augustus Rivinus (1652–1723) and Joseph Pitton de Tournefort (1656–1709), who used more categories (class, order, section) and a consistent generic name along with a modifier (a phrase) that later evolved into Linné's binomial nomenclature. Before these works, plants in different "higher genera" frequently had the same name; to Ray apples were *Malus* and peaches were *Malus Persica*. This evolution of taxonomy reached its first peak with the work of Carl Linné (Linnaeus), the Swedish botanist who introduced the idea of a consistent binomial for each species of plant and animal and adopted or invented higher categories to classify larger groups, producing what become known as the Linnaean hierarchy.

The primary purposes of early classifications, such as those of Linné, were first, the classifications organized diversity and second, they provided a means of identification. In this second role, many early taxonomies worked more like what we recognize as keys, dichotomous choices of characters designed to identify a specimen. Thus, early classifications were built around a limited number of characters thought to be important in identifying organisms or showing where unidentified organisms might fit given the characters used. Also, the form of the classification made a place for any newly discovered organisms. Many kinds of taxonomies were proposed during the eighteenth and nineteenth centuries, each built around a concept thought to produce a natural classification and thus an enduring taxonomy. For example, the Quinterians thought that groups came in fives and attempted to classify around this principle. Some, such as Richard Owen (England), built taxonomy around the concept of idealized types (archetypes) that defined major groups. Darwin provided a different concept; natural classifications could be built around the principles of evolutionary descent and common ancestry. This concept was taken to its logical end by the German entomologist Willi Hennig in the mid-twentieth century and emerged as the dominant paradigm of biological classification used to form modern taxonomies.

Taxa and categories. Most modern taxonomies are hierarchical classifications of groups within groups, resulting in a biological classification. In addition, modern taxonomies include detailed descriptions of the organisms and a history of the various names by which these organisms have been called. Such works are usually referred to as taxonomic revisions.

In classifications and taxonomic revisions, each group of plants or animals is given a name. The group itself is a taxon (examples: *Homo sapiens*, Vertebrata). The place of the group within the taxonomy is denoted by ranking. In Linnaean taxonomies, rank is denoted by placing the taxon name with a category rank (example: class Vertebrata). Each category has a place on the hierarchy relative to other categories. Each taxon is unique, but the categories are not unique. For example, there are over 400 different families of bony fishes. This relative ranking is set by convention for each system of taxonomy. There are several systems of taxonomy, including systems for bacteria (prokaryotes) and viruses. The two most familiar are for plants and animals. Some of the common categorical ranks and examples of taxa are provided below for humans and pine trees:

Humans
- Domain Eukarya
  - Kingdom Animalia
    - Phylum Chordata
      - Class Synapsida
        - Order Primates
          - Family Hominidae
            - Genus *Homo*
              - Species *Homo sapiens*

Pine trees
- Domain Eukarya
  - Kingdom Plantae
    - Division Pinophyta
      - Class Pinopsida
        - Order Pinales (=Coniferales)

Family Pinaceae
Genus *Pinus*
Species *Pinus ponderosa*

See TAXON; TAXONOMIC CATEGORIES.

Modern taxonomy in practice. There may be 12–20 million different species of organisms alive today. Each one requires a name so that scientists can communicate about the species. Recognition of species and their classification is fundamental to fields as diverse as conservation biology and medicine, and the status of a particular population of organisms as a species can have international consequences as nations attempt to conserve and use their biological resources. One major goal of taxonomy is naming species. Each species has a two-part name, a genus name and a specific epitaph. *Homo sapiens* (humans) and *Pinus ponderosa* (ponderosa pine trees) are examples. Species names are always "set apart" from the rest of the text, usually by italics (*Homo sapiens*). The second major goal of taxonomy is to organize these species into larger groups and give these groups names. Each group of species is given a single name (example: Mammalia). Groups of species are frequently referred to as supraspecific taxa or higher taxa, with "higher" meaning that their rank category is higher than the rank of species. *See* SPECIES CONCEPT.

From the time of Linnaeus until the end of the twentieth century, there was a dialog among systematists as to what might constitute the most general kind of classification. Intense debates were held in the literature and in scientific meetings from the 1950s to 1990s among proponents of three distinct points of view. The Pheneticists asserted that general classifications should be based on some measure of overall similarity (counting differences and similarities) in order to be repeatable. The Phylogeneticists (Cladists) asserted that general classifications should be based on common ancestry. The Evolutionary Taxonomists grabbed the middle ground and asserted that some groups could be recognized based on common ancestry while others should be based on overall similarity (especially overall genetic similarity). By the end of the twentieth century, it became apparent that the Phylogeneticists had won the battle, primarily because they had demonstrated that it was possible to reconstruct evolutionary histories of organisms (phylogenies) objectively, thus blunting the objections of the Pheneticists who claimed that reconstructing phylogenies was too problematic to be objective. They also demonstrated that the hybrid systems of the Evolutionary Taxonomists (the middle ground) resulted in loss of information or even misleading information about common ancestry relationships and similarity. Thus, in the twenty-first century, taxonomy is returning to Darwin's assertion that taxonomies should be, whenever possible, based on common ancestry relationships. This paradigm unites the diverse fields of systematics, bringing taxonomies in line with phylogenies when phylogenies are available. Taxon names become meaningful in an evolutionary sense because the organisms classified in the taxon share a common ancestral species only among themselves. Such a taxonomy is said to be logically consistent with the evolutionary history of the group.

The effects of this paradigm shift can be profound for taxonomy. Consider birds. Old-style taxonomies (and many current school texts) recognize birds as a taxon removed from reptiles and rank both groups as classes (class Reptilia, class Aves). Modern classifications recognize birds as an order (or even suborder) within a larger group, Archosauria, that includes crocodiles (superorder Archosauria, order Crocodilia, order Aves). The old-style classification is one that emphasizes the similarities between crocodiles and lizards and the obviously distinctive nature of birds. The modern classification recognizes that birds share a common ancestor with crocodiles (as well as with dinosaurs and other archosaurs).

The key to understanding why modern biologists prefer the modern classification is the predictive power of the taxonomy: the ability of the classification to predict and explain aspects of both similarity and difference in a manner that is superior to the old-style classification. Crocodiles and birds have certain bones that are hollow; lizards do not. The old-style classification predicts that these hollow bones evolved independently in crocodiles and birds. The modern classification predicts that hollow bones evolved in the common ancestor of crocodiles and birds—an observation consistent with the accepted phylogeny where such hollow bones are interpreted as evolutionary innovations of a common ancestor shared by all archosaurs, crocodiles, dinosaurs, and birds, but not in a common ancestor these archosaurs share with lizards. The old-style classification predicts that a scale-covered body is an evolutionary innovation of the common ancestor of lizards and crocodiles. The modern classification predicts that while a scale-covered body was, indeed, a homology of the common ancestor of lizards and crocodiles, it evolved at a much earlier date and the ancestor is shared with birds as well. The old-style classification predicts that parental care evolved independently in crocodiles and birds, but it is no surprise to modern taxonomists that dinosaurs practiced parental care, given that crocodiles do so. Such care is best explained as a behavioral evolutionary innovation that originated in the common ancestor of all archosaurs.

Taxonomic formalities. Taxonomy as an activity that gives names to groups of organisms has evolved over the last two centuries. Although binomial nomenclature of Linné was adopted by 1800, his set of rules for naming plants fell into disuse. Each taxonomist created his own set of naming rules, and this produced chaos as different names were applied to the same species or group by different authors. This prompted both the zoological and botanical communities to formulate rules for naming in different countries. But this hardly solved the problem: Americans might have a set of rules that differed from Germans, hardly the way to foster international understanding. By the beginning of the twentieth

century, international rules governing the formation and use of names were adopted that were meant to be universal in application, irrespective of borders or language. These rules were formalized by the adoption of Codes of Nomenclature, first for animals and plants, and much later for prokaryotes, cultivated plants, and viruses.

All codes are independent of each other, so some have slightly different rules. In general, each Code of Nomenclature performs two basic functions. First, within the scope of the code, it specifies what names are to be used and how this is determined. In general, disputes (two names for the same taxon) are arbitrated by declaring that the oldest published name is to be used. This is the Law of Priority. Younger names, called junior synonyms, are common because of a lack of communication or because of taxonomic disputes or failure to appreciate sexual or ontogenetic differences. In addition, each Code specifies how names are to be formed for taxa placed at specific categorical ranks. This naming principle requires, for example, that a taxon ranked at the family level have an ending of "-idea" (zoology) or "-aceae" (botany, Prokaryota). The intention of the person who names a species is signaled by that person pointing to a type specimen or series of type specimens. Such type specimens are meant to leave an objective record of the intention of the original describer; if no type specimen is designated, a later worker can designate a type specimen but cannot arbitrarily change the name. This is termed the Principle of Typification and can extend at various levels above the rank of species by designating type species, type genera, etc., as required by the Codes. These rules of nomenclature can be quite complex, leading to a group of systematists who specialize in understanding and interpreting the various Codes. Disputes can be brought to an international commission that has the power to make decisions to accept or change names as needed.

The various Codes of Nomenclature have survived and prospered for several reasons. (1) The names of species (*Homo sapiens*) and higher (more inclusive) taxa such as *Homo*, Mammalia, and Canidae are formed in Latin, a holdover of a time that all scholarship in Western society was written in Latin. As a "dead language," it is a convenient neutral language for taxonomists of different cultures and native languages. (2) Designation of hierarchical relationships through the use of ranks (genus, family, class) provides a reusable system of ranks. But perhaps the greatest strength is the fact that (3) the Codes are neutral as to biological meanings of the taxa. The Codes specifically state that their purposes are directed toward the use and formation of names, not the biological meaning or significance of the names. They leave it up to the biological community to make that determination. For example, to a phylogeneticist the name *Homo* is used to name a group of bipedal mammals that share a common ancestor unique to them (*Homo* is a unique evolutionary unit or monophyletic group), while to a pheneticist *Homo* might be used to name a group of bipedal mammals that are more similar to each other than to another group of bipedal mammals, like *Australopithecus*. Each generation of biologists can use or reject names based on their ideas of what constitutes a good taxonomy, but must work within the Codes; the Codes only govern which names are available for use. This has protected the Codes from changing ideas as to the nature of biological diversity while ensuring that names are used in a consistent manner that can be understood by everyone. *See* BACTERIAL TAXONOMY; PLANT NOMENCLATURE; TYPE METHOD; ZOOLOGICAL NOMENCLATURE.

Phylogenies: basic to modern taxonomy. The triumph of phylogenetic systematics and its attendant methods of discovering the evolutionary relationships among species have revolutionized taxonomy and biological classification. The principle insight provided by Hennig and his predecessors (such as the German systematist Walter Zimmerman) was that different homologies have different value in reconstructing phylogeny. Homologies are similarities in structure, behavior, and other characteristics that have a single evolutionary origin and thus have potential to act as "markers" of common ancestry. In contrast, homoplasies are similarities that have two to several evolutionary origins. For example, having legs is a homologous similarity marking the origin of tetrapod vertebrates such as ourselves. In contrast, homothermy, the ability to maintain a constant body temperature, is a homoplasous similarity observed in birds and mammals and does not have a single origin, given our current understanding of vertebrate relationships. One might think that all homologies are useful in reconstructing common ancestry relationships, and this is true, but not all are useful at the same time. For example, humans and frogs have multiple digits on their legs while living horses have a single digit. Horses and humans have hair, but frogs lack hair. Just given this, we might say that humans are as similar to frogs as they are to horses. But having more than one digit, while each is homologous with the other, is a character that appeared very early in the evolution of land vertebrates and is a characteristic of the common ancestor not only of frogs and humans but also of horses. In contrast, having hair evolved much later and is characteristic of the common ancestor of horses and humans but not of the ancestors of all three. Thus, we can conclude that horses and humans shared a more recent common ancestor and thus are more closely related. The presence of hair is a marker for all mammals sharing a common ancestor not shared with frogs. In short, all homologies can be used to reconstruct phylogenies but not at the same level of the hierarchy.

A large number of systematists now devote their research to reconstructing phylogenies. There are detailed protocols for determining which homologies are relevant to particular levels of the Tree of Life and which similarities are actually homoplasies and not homologies. Critical to such research are the 300 years of taxonomic research that have resulted in our present classifications. In some cases, the "phylogenetic signal" was so strong that clades (common

ancestry groups) recognized today were recognized by those who came before the rise of modern methods. After all, Aristotle himself recognized animals with backbones. In other cases, the clades were more subtle, or preconceived ideas about the distinctiveness of a group caused the classification to be different from what we now think are the common ancestry relationships (birds and crocodiles; great apes and humans). In these cases, the taxonomies must be changed to reflect the phylogenies. In other cases, there are no phylogenies, and we must rely on the present taxonomies. *See* PHYLOGENY.

Challenges to Linnaean taxonomy. Can a system of nomenclature invested by an eighteenth-century Swedish botanist survive the discovery of 20 million species of organisms and the attendant phylogeny of their relationships? This question is now being debated. Some have suggested that it cannot survive and have formulated an alternative system that dispenses with the current Codes and substitutes their own, the PhyloCode. Others suggest that the Linnaean system can be modified to suit the future needs of taxonomy. Given that major organizations that use taxonomy have adopted the Linnaean system, it is doubtful that any alternative system will soon replace it. For example, GenBank, the major repository of all genetic sequences, is organized around the system of Linnaean names and ranks, adhering to the present Codes. However, just as the debate over phenetics, evolutionary taxonomy, and phylogenetics invigorated systematic and taxonomic research, the debate over retaining the Linnaean system or replacing it is also healthy in stimulating debate over the form and function of modern taxonomy. E. O. Wiley

Bibliography. J. Cracraft and M. J. Donoghue (eds.), *Assembling the Tree of Life*, Oxford University Press, New York, 2004; W. Hennig, *Phylogenetic Systematics*, University of Illinois Press, Urbana, 1966; J.-W. Wägele, *Foundations of Phylogenetic Systematics*, Verlag Dr. Friedrich Pfeil, Munich, 2005; E. O. Wiley, *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*, Wiley, New York, 1981.

Tea

A small tree (in cultivation, constant pruning makes it a shrub 3–4 ft or 0.9–1.2 m tall); a preparation of its leaves dried and cured by various processes; and a beverage made from these leaves. The plant, *Camellia sinensis* (or *Thea sinensis*), is an evergreen tree of the Theaceae family, native to southeastern Asia, and does best in a warm climate where the rainfall averages 90–200 in. (2200–5000 mm). The slower growth at higher altitudes improves the flavor. China, Japan, Taiwan, India, Sri Lanka, and Indonesia are among the leading tea-producing countries, with China contributing about one-half of the world's supply. According to Chinese folklore, the Chinese emperor Shen-mung discovered the use of tea about 2500 B.C. Tea leaves contain caffeine, various tannins, aromatic substances attributed to an essential oil, and other materials of a minor nature, including proteins, gums, and sugars. The tannins provide the astringency, the caffeine the stimulating properties.

Flowering branch of tea plant (*Thea sinensis*).

The tea plant normally produces a full crop of leaves (flush) about every 40 days. In China and Japan the leaves are plucked when the flush has mostly completed its growth, but in India and Indonesia the leaves are picked every week or two, resulting in most of the leaves being gathered at the best time. The leaves are named, beginning with the topmost, as pekoe tip, orange pekoe, pekoe, first souchong, second souchong, first congou, and second congou.

In its natural state the tea plant grows to a height of 15–30 ft (4.5–9 m). The leaves vary from $1\frac{1}{2}$ to 10 in. (4 to 25 cm) long and are thick, smooth, and leathery. The white, fragrant flowers are produced in the axils of the leaves (see **illus.**). The fruits are dark brown capsules. *See* THEALES. Earl L. Core

Diseases. Blister blight disease of tea was recognized in the midnineteenth century, but despite sporadic occurrence throughout the world, it was not considered a serious threat until it devastated plantings in Sri Lanka (Ceylon) and India during 1946. About 25,000 acres (10,000 hectares) of tea in Sumatra were destroyed 3 years later, and more recent reports of 30% losses attest to the destructive capability of the disease. Most serious losses attributed to blister blight have occurred in southeastern Asia.

The disease is almost entirely foliar, although young, green stems and buds may be attacked. Mostly round, translucent, sometimes pinkish spots lighter in color than surrounding leaf tissues enlarge to form noticeably concave depressions on the upper leaf surfaces. As seen from below, these convex "blister" surfaces colored light brown to white soon rupture to expose thousands of spores (with mucilaginous coverings) of the causal fungus *Exobasidium vexans*. These relatively fragile basidiospores, killed by exposure to 95°F (35°C) or 30 min of sunlight, are readily airborne and may germinate within 2 h at 80% relative humidity, but can remain dormant for as long

as 5 days. Infection follows germination within 24 h and, after 3–10 days incubation and 6–9 days for secondary lesion development, sporulation may occur again after 10–20 days. Production of basidiospores follows a diurnal cycle, with greatest concentration of spores in the air between midnight and 4 a.m. Disease development is retarded by low temperatures and the more intense solar insolation of higher altitudes, but frequent interruptions of full sunshine and prolonged periods of rainy weather at such elevations favor the disease.

High rainfall conditions are common to many tea-growing areas, and copper compounds (copper oxide or copper oxychloride) used either as sprays or dusts have been found most effective for control of blister blight. Disease forecasting, based on hours of sunshine, has been attempted experimentally.

Oil spot of tea is a condition known to occur only in Sri Lanka at elevations above 5000 ft (1500 m). There is presumptive evidence that the spotting symptoms seen on young leaves prior to defoliation result from a toxin produced by an unidentified fungus cultured from discolored wood below affected shoots. Early attempts at control with fungicides have been ineffective.

Curled leaves, shortened internodes, and zigzag growth of stems are often associated with phloem necrosis disease of tea. The only diagnostic symptom of the complex syndrome is necrosis of the phloem, seen as irregular or blotchy spotting on the inner phloem surface of bark removed by peeling. Known only from Sri Lanka, where incidence has ranged to 70%, phloem necrosis is almost completely restricted to elevations above 4000 ft (1200 m), and is more severe above 5500 ft (1650 m). The disease has been transmitted by grafting, but is not transmitted through soil. Plants placed at 82°F (28°C) have produced new shoots without symptoms, but the only practical control at present is to rogue diseased plants and replace them with healthy plants.

Chlorotic or brown leaves on dead branches or entire bushes are the obvious symptoms of collar and branch canker caused by *Phomopsis theae*. Bark cankers that usually develop on young plants result from infection through pruning or inadvertent injury to stems or branches. Application of fungicide to injuries has been recommended for control.

Symptoms of mottle scab (*Elsinoe theae*) include irregular to angular, dark lesions found only on mature leaves. Lesions caused by white scab (*E. leucospila*), also dark in color, occur on both surfaces of young leaves and later lighten to nearly white, enlarge to 0.04–0.12 in. (1–3 mm) in diameter, and become circular in shape. Both diseases are relatively uncommon.

Root diseases of tea, caused by *Poria hypobrunnea* (red root), *Rosellinia necatrix* (white root rot), *Helicobasidium compactum* (purple root rot), *Armillariella mellea* (Armillaria root rot), and *Ustulina zonata*, often remain undetected until the infection has spread to adjacent plants. No treatment adequate to rejuvenate infected plants has been devised, and control presently consists of roguing infected plants, fumigating of the planting site, and replanting with healthy stocks. *See* PLANT PATHOLOGY.

A. A. Cook

Tea production. Tea production technology depends on whether black, green, or semifermented tea is to be manufactured. The characteristic appearance and flavor of each arise more from different processing and manufacturing methods than from type of leaf.

Black tea. This tea requires quick transportation of the green leaf from garden to factory, where it is carefully spread on withering tats. The leaf loses moisture and becomes flaccid within 24 h. Circulation of conditioned air during withering shortens the time substantially. The next process, called rolling, ruptures the leaf cells and releases their juices. This is accomplished in a brass roller box, equipped with an adjustable pressure cap and an open base. The box is filled with withered leaf and rotated eccentrically over a brass table fitted with curved battens. During a 30-min period the leaf cells rupture under the pressure and the rubbing action. In a second rolling period the leaf is twisted and coated with juices, leading to balling. The leaf changes to a bright copper color, and a characteristic tea aroma develops. A roll breaker is subsequently employed to disintegrate the balls for fermentation.

In fermentation, enzyme action and oxidation occur. The rolled leaf is spread to a 2–3-in. (5–7.5-cm) depth on fermenting beds and held under high humidity at 75–80°F (24–27°C). Since the body, or strength, of the tea depends on this fermentation, careful attention to its progress is required. Approximately $3\frac{1}{2}$ h are needed for the rolling and fermentation operations.

Firing or drying of fermented tea leaf conventionally is carried out by passing it through heated ovenlike chambers. Temperatures of 170–180°F (77–82°C) arrest fermentation and develop the familiar blackish color. Drying is practically completed in one pass, but further heat is applied to produce the "case hardening" that protects quality. The leaf is sieved, graded, and packed into chests lined with aluminum foil. Each chest contains about 100 lb (45 kg) of finished tea.

Green tea. Green tea manufacture requires that the plucked leaves be steamed as quickly as possible. Such processing at 160°F (71°C) makes them soft and pliable and, by inactivating the natural enzymes, prevents fermentation. After steaming, the leaf is alternately rolled and dried until it becomes too stiff for further manipulation. For export, green tea is refired in pans with mechanical stirring to produce a luster.

Oolong teas. Oolong teas are midway between black and green teas in that they are semifermented. After a short sun-withering in the garden, the leaf is gently rolled in the plucker's hands, so a slight fermentation is initiated. After a short period this leaf is sent to the factory to be fired and packed for shipment.

Instant tea. This type of tea is obtained by spray-drying of a black tea extract, with or without the admixture of maltodextrins. The technology is

patterned closely after that followed in manufacturing of instant coffee. Concentrated tea extract also may be combined with a heavy sugar syrup and marketed as a liquid. Solubility of instant tea powder can be varied by manipulations in processing. *See* FOOD ENGINEERING. John H. Nair

Bibliography. A. Stella, *The Book of Tea*, 1992.

Technetium

A chemical element, Tc, atomic number 43, discovered by Carlo Perrier and Emilio Segrè in 1937. They separated and isolated it from molybdenum (Mo; atomic number 42), which had been bombarded with deuterons in a cyclotron. Technetium does not occur naturally. *See* ATOMIC NUMBER; ELEMENT (CHEMISTRY).

In the periodic table, technetium is located in the middle of the second-row transition series, situated between manganese and rhenium. Because of the lanthanide contraction, the chemistry of technetium is much more like that of rhenium, its third row congener, than it is like that of manganese. Its location in the center of the periodic table gives technetium a rich and diverse chemistry. Oxidation states −1 to +7 are known, and complexes with a wide variety of coordination numbers and geometries have been reported. *See* LANTHANIDE CONTRACTION; PERIODIC TABLE; RHENIUM; TRANSITION ELEMENTS.

1																	18
1 H	2											13	14	15	16	17	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

The most readily available chemical form of technetium is the ion pertechnetate (TcO_4^-), the starting point for all of its chemistry. In its higher oxidation states (+4 to +7), technetium is dominated by oxo chemistry, which is dominated by complexes containing one or two multiply bonded oxygen (oxo) groups.

Since about 1979, inorganic chemists have been very interested in understanding and developing the fundamental chemistry of technetium, utilizing the isotope ^{99g}Tc. Most of this interest has arisen from the utility of another isotope of technetium, ^{99m}Tc (where m designates metastable), to diagnostic nuclear medicine. In fact, ^{99m}Tc in some chemical form is used in about 90% of all diagnostic scans performed in hospitals in the United States. The nuclear properties of ^{99m}Tc make it an ideal radionuclide for diagnostic imaging. This technetium isotope has a 6-h half-life, emits a 140-keV gamma ray which is ideal for detection by the gamma cameras used in hospitals, and emits no alpha or beta particles. Silvia S. Jurisson

Bibliography. R. Alberto, *Technetium, Rhenium: Topics in Current Chemistry*, vol. 176, Springer-Verlag, 1995; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; K. Schwochau, *Technetium: Chemistry and Radiopharmaceutical Applications*, Wiley-VCH, 2000.

Technology

Systematic knowledge and action, usually of industrial processes but applicable to any recurrent activity. Technology is closely related to science and to engineering. Science deals with humans' understanding of the real world about them—the inherent properties of space, matter, energy, and their interactions. Engineering is the application of objective knowledge to the creation of plans, designs, and means for achieving desired objectives. Technology deals with the tools and techniques for carrying out the plans.

For example, certain manufactured parts may need to be thoroughly clean. The technological approach is to use more detergent and softener in the wash water, to use more wash cycles, to rinse and rerinse, and to blow the parts dry with a stronger, warmer air blast. Often such refinements provide an adequate action. However, if they do not suffice, the basic technique may need to be changed. Thus, in this example, science might contribute the knowledge that ultrasonically produced cavitation counteracts surface tension between immiscible liquids and adhesion between clinging dirt and the surface to be cleaned, and thereby produces emulsions. Engineering could then plan an ultrasonic generator and a conveyor to carry the parts through a bath tank in which the ultrasonic energy could clean them. The scientist may use ultrasonic techniques to determine properties of materials. The engineer may design other types of devices that employ ultrasonics to perform other functions. These specialists enlarge their knowledge of ultrasonics and their skill in using this technique not for its own sake but for its value in their work.

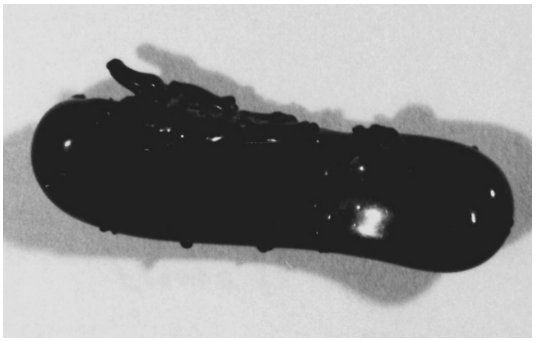
The technologist is the specialist who carries out the technique for the purpose of accomplishing a specified function, and extends knowledge and skill of ultrasonic cleaning by refinement and perfection of the technique for use on various materials soiled in different ways. Technological advances improve and extend the application to cleaning other parts under other conditions. *See* AERONAUTICAL ENGINEERING; CHEMICAL ENGINEERING; CIVIL ENGINEERING; ELECTRICAL ENGINEERING; ENGINEERING; ENGINEERING, SOCIAL IMPLICATIONS OF; ENGINEERING DESIGN; FOOD ENGINEERING; HUMAN-FACTORS ENGINEERING; MECHANICAL ENGINEERING; NUCLEAR ENGINEERING; PRODUCTION ENGINEERING; SCIENCE; SPACE TECHNOLOGY. Robert S. Sherwood; Harold B. Maynard

Tektite

A member of one of several groups of objects that are composed almost entirely of natural glass formed from the melting and rapid cooling of terrestrial rocks by the energy accompanying impacts of large extraterrestrial bodies. Tektites are dark brown to green, show laminar to highly contorted flow structure on weathered surfaces and in thin slices, are brittle with excellent conchoidal fracture, and occur in masses ranging to as much as tens of kilograms but are mostly much smaller to microscopic in size. The shapes of tektites are those of common fluid splash and rotational forms including drops, spheres, and dumbbells, unless they have been abraded together with surface gravels. A few tektites have shapes that are caused by two different heating events: the impact that melted the parent rock to form the glass, and a second event apparently due to reentry aerodynamic heating.

Occurrences, groups, and ages. With the discovery of tektites in the Soviet Union, there are now five major groups known: (1) North American, 3.4×10^7 years old, found in Texas (bediasites) and the Georgia Coastal Plains, with a single specimen reported from Martha's Vineyard, Massachussetts; (2) Czechoslovakian (moldavites), 1.5×10^7 years old, found both in Bohemia (green and transparent) and in Moravia (brown and turbid); (3) Ivory Coast, 1.3×10^7 years old; (4) Russian (irgizites; see **illus.**), 1.1×10^7 years old, found in the Northern Aral Region; and (5) Australasian, 700,000 years old, occurring notably in Australia, the Philippines, Belitung, Thailand, and numerous other localities. The North American, Ivory Coast, and Australasian tekites also occur as microtektites in oceanic sediment cores near the areas of their land occurrences. In the land occurrences, virtually all of the tektites are found mixed with surface gravels and recent sediments that are younger than their formation ages.

Tektites, in general, are not rare objects. Millions of tektites have been recovered from the Australasian occurrence. However, some of the more pleasingly colored specimens from Bohemia and Georgia are prized by collectors and command substantial prices, either as specimens or potential gem material.



Dumbbell-shaped tektite (irgizite) from the Zhamanshin meteoritic impact crater, Northern Aral Region, Russia. The glass is so dark brown that it appears black in reflected light.

Three of the tektite groups are associated with known large impact craters. The irgizites occur in and immediately around the Zhamanshin meteoritic impact crater [10–15 km (6–9 mi) in diameter] in the Northern Aral Region of Russia. Both the Ivory Coast tektites and the moldavites occur close to probable source impact craters of identical age, the Bosumtwi Crater, Ghana [10.5 km (6.5 mi) in diameter] and the Ries Crater, Germany [24 km (15 mi) in diameter], respectively. The source craters for the North American and Australasian tektites have not yet been identified positively.

Composition. The chemical compositions of tektites differ from those of ordinary terrestrial rocks principally in that they contain less water and have a greater ratio of ferrous to ferric iron, both of which are almost certainly a result of their very high-temperature history. The extreme ranges of major element compositions, expressed as oxides in weight percent, are as follows: SiO_2, 48–85 wt %; Al_2O_3, 8–18 wt %; FeO, 1.4–11 wt %; MgO, 0.4–28 wt %; CaO, 0.3–10 wt %; Na_2O, 0.3–3.9 wt %; K_2O, 1.3–3.8 wt %; TiO_2, 0.3–1.1 wt %. However, portions of the foregoing ranges are based on analyses of glass particles from oceanic sediment cores that may or may not be tektites, as this identification is difficult for these very small particles. If other glass occurrences from impact craters were included as tektites, such as splash form glass from the Wabar Craters in Saudi Arabia, and the Lonar Crater in India, the ranges would expand significantly. As indicated by the wide range of chemical compositions, there are corresponding wide ranges of properties such as specific gravity and refractive index.

Inclusions. Spherical vesicles ranging in size from microscopic to as much as several centimeters are common in most tektites, as are small lechatelierite (silica glass) particles. The presence of coesite, a high-pressure polymorph of silica, in some tektites from Southeast Asia is additional evidence of the impact genesis of tektites. Coesite is known to form at the Earth's surface only from the very high pressures of transient shock waves caused by large hypervelocity meteorite impacts and cratering events. Baddeleyite, monoclinic ZrO_2, is present in some tektites as a high-temperature decomposition product of the mineral zircon. Meteoritic nickel-iron also has been observed in a few tektites, which is further evidence of the origin of tektites as meteoritic impact melts. Such inclusions of coesite, baddeleyite, and nickel-iron are common in the glassy fusion products in and around a number of terrestrial impact craters. *See* COESITE; METEORITE.

Historical perspective. Tektites have been recognized as unusual objects for more than a millennium, and they were the objects of intensive scientific research from the turn of the century until a few years before the return of lunar samples by the Apollo missions. This activity was generated by the possibility that tektites originated from the Moon as secondary ejecta from impact craters on the lunar surface. This

point of view was effectively advocated by a number of scientists until the recognition of the associated impact craters with some tektite groups on the Earth. Later analyses of lunar samples demonstrated that they are not suitable parent rocks for tektites. Unfortunately, the discovery of the Russian tektites, in a clear relation to an impact crater on the Earth, came too late to influence these arguments significantly.

Unanswered questions. The source crater of the Australasian tektites has not yet been identified, although this is the youngest and most widespread tektite group. Also, present understanding of the mechanics and sequence of events in very large impact crater formation does not permit specifying unambiguously the mode by which some tektites escape the Earth's atmosphere to reenter and form the remelted layer as a result of aerodynamic heating. Elbert A. King

Bibliography. G. B. Dalrymple, *P40 SAr: P39 SAr Age Spectra and Total-Fusion Ages of Tektites from Cretaceous-Tertiary Boundary Sedimentary Rocks in the Beloc Formation, Haiti*, U.S. Geological Survey, 1993; E. A. King, The origin of tektites: A brief review, *Amer. Sci.*, 65:212–218, 1977; E. A. King, *Space Geology: An Introduction*, 1976.

Telecast

A television broadcast, involving the transmission of the picture and sound portions of the program by separate transmitters at assigned carrier frequencies within the channel assigned to a television station. A telecast is intended for reception by the general public, just as is a radio broadcast. The picture may be either in black and white or in full color, using amplitude modulation, while the sound portion (in the United States) uses frequency modulation. The channels assigned for telecasts by the Federal Communications Commission, each 6 megahertz (MHz) wide, cover frequencies as follows: 54–72 MHz (channels 2 through 4), 76–88 MHz (channels 5 and 6), 174–216 MHz (channels 7 through 13), and 470–890 MHz (channels 14 through 83). *See* RADIO SPECTRUM ALLOCATION; TELEVISION; TELEVISION NETWORKS; TELEVISION RECEIVER; TELEVISION STANDARDS; TELEVISION TRANSMITTER. John Markus

Telecommunications civil defense system

When any possible emergency threatens the security or socioeconomic structure of the United States or any political subdivision, an appropriate warning must be made in order for a reaction to occur within an appropriate period of time. Any such reactions cannot operate without a means of communication to prewarn of impending disaster or attack and to assist the civil defense and military effort after a disaster has occurred. To meet these requirements in an integrated manner, the Federal Emergency Management Agency (FEMA) and the National Communications System (NCS), working with the telecommunications industry, provide warning centers and private-line telecommunications systems to carry vital emergency information and alerts to both the military and the general public.

Federal Emergency Management Agency. The Federal Emergency Management Agency has primary responsibility for financial and technical support of the two basic communications systems that meet this need during a crisis situation: the National Warning System (NAWAS) and the Emergency Broadcast System (EBS).

National Warning System. Operating 24 h a day, the National Warning System is a voice-grade service that consists of over 65,000 channel miles (100,000 km) connecting 2300 warning points in the United States with a national primary warning center located with the North American Air Defense (NORAD) Command. Should the primary warning center be rendered inoperative, alternate centers can cover all warning points that were dependent on the inoperative center. Similarly, modern switching facilities are available to replace any portion of the network that may become inoperative. The warning centers are located at key defense installations so that full advantage can be taken of all available information.

This warning system (originally called the Civil Defense Warning System) has been in existence since the late 1950s. The system is designed to bypass, insofar as practicable, major cities that are likely targets. Warning information can be sent to all points on the network in approximately 15 s. From the 2300 warning points on the network, warning information is relayed through state and local systems to more than 5000 local points in an average time of 7 min.

The National Warning System network consists of a control circuit and various warning (ring-down) circuits. The control circuit is used to pass administrative information and to coordinate operation. This circuit passes both voice and tone signals to activate switching devices that control national and regional circuit configurations. A selective signaling arrangement with predetermined codes is used to set up the various regional configurations. State circuits interconnect with regional circuits through a unique arrangement that permits state or local extensions to receive voice communication directly from regional circuits. Operation of a switching device at the regional-state interface transfers control of the state circuit to the state operator for two-way voice operation with other state and local-extension users. Many of the local users are fire stations, police, sheriffs, and other civilian agencies with an emergency control mission. Most regional, state, and local-extension National Warning System stations consist of a four-wire telephone instrument, a push-to-talk handset, a loudspeaker, and a bell which operates independent of the telephone instrument switchhook.

Emergency Broadcast System. The mission of the Emergency Broadcast System is to provide the President of the United States with a means of communicating with the general public through the use of

nongovernment broadcast stations during the period preceding, during, and following an enemy attack, natural disaster, or other national emergency. The Broadcast System consists of four multipoint private-line services routed to 31 locations at 2400 bits per second. Either of the two point-of-origin, send-receive stations (a primary location and its alternate) can transmit the activation, termination, and National Industry Advisory Council (NIAC) order messages to all other stations. Each point-of-origin send-receive station is equipped with a line controller unit, a 4420 terminal, diagnostic modems, and floppy disk storage units. All other receive-only stations are equipped with 2024 modems and a Model 43 buffered terminal which is programmed to provide an answer with the depression of a key. Activation of a switch at these locations and at the point of origin allows voice confirmation and emergency coordination.

To activate the national-level Emergency Broadcast System, the President gives an order to the White House Communications Agency duty officer, who notifies one of the two origination points by telephone or radio. At the origination point, the systems controller or a communications center assistant releases a message over the EBS Net. This message, containing the Emergency Activation notice and National Industry Advisory Council order number, reaches key offices of the broadcast networks, press agencies, and American Telephone and Telegraph offices. Calls among the White House Communications Agency, origination points, and industry can be authenticated. Authentication lists are furnished by the National Security Agency. The broadcast networks send the Emergency Activation Notice to their affiliated stations by internal network alerting systems. The press agencies, after confirmation, transmit a teletype message over their Radio Wire Teletype Networks. Broadcast stations transmit the Federal Communications Commission attention signal, and an emergency action broadcast announcement is made over the air.

National Communications System. The National Communications System was established in 1963 as a confederation in which federal departments and agencies participate with their telecommunications assets to provide essential communications under all conditions ranging from normal day-to-day situations to national and international crises, including nuclear attack. The principal assets of the National Communications System comprise the telecommunications networks of 22 federal agencies including the Federal Emergency Management Agency and the Department of Defense. The Secretary of Defense serves as the executive agent for the National Communications System with the director of the Defense Communications Agency designated as the National Communications System manager. The National Communications System manager performs various functions to support the mission of the National Communications System. Among these are directing a National Communications System/Defense Communications Agency Operations Center (NCS/DCAOC) and the National Coordinating Center (NCC).

National Communications System/Defense Communications Agency Operations Center. The Defense Communications Agency systems control officers and their staff at the center provide the National Communications System a 24-h daily point of contact for status on essential government networks and communications services. The center is connected to commercial telecommunications carriers' network operations centers, which provide status on a 24-h basis to the center regarding any significant failures or threatening conditions that might affect any Defense Communications service provided over their facilities.

Of the systems provided by commercial carriers, two of the most vital to the nation's defense are the Strategic Air Command's Primary Alert System (PAS) and the NORAD Alert System (NAS). The Primary Alert System is a voice alerting network specifically designed to serve the Strategic Air Command. Tests are conducted frequently to assure the viability of this system. Presently made up of two networks, the NORAD Alert System consists of the Primary Voice Alert System (PVAS), which connects principal military locations, and the Weapons Alert System (WAS) which is mainly used to alert fighter interceptor squadrons. For utmost in service reliability, hardened cable facilities, diversity, and redundancy are used in these systems. In addition, constant upgrades are made to the systems as new concepts and technology become available, which will further enhance reliability and survivability.

National Coordinating Center. Another major function that is under the National Communications System manager is the operation of the National Coordinating Center. This center was established in January 1984 and is the vehicle by which the federal government and the telecommunications industry jointly respond to National Security and Emergency Preparedness (NSEP) telecommunications requirements. Representatives from the telecommunications industry and federal government staff the National Coordinating Center. They provide for a rapid exchange of information and expedite responses during crisis situations.

The National Coordinating Center has the capability to suggest responses to a broad spectrum of emergency or crisis situations. The extent of its authority is governed by the nature of the emergency and the following specific legal conditions: after Section 606 of the Communications Act of 1934 is invoked; during disasters declared by the President; and during any emergency or crisis defined in the National Security Emergency Preparedness Procedure Manual distributed by the Defense Communications Agency. Activation of the National Coordinating Center is brought about by one of these legal conditions and any one of the following operational situations: (1) requests for assistance or coordination during an emergency are received from a telecommunications industry; (2) the federal government agency requesting an emergency expedite declares that existing procedures will not result in sufficient service;

or, (3) the scope of the National Security and Emergency Preparedness service or support requirements involves coordination across several states, regions, companies, or organizations. In a wartime situation where the assets of the telecommunications industry become nationalized, the representative in the National Coordinating Center would be responsible for carrying out directives, related to their specific company, consistent with federal government authority.

James P. McQuillan

Teleconferencing

Broadly, the various ways and means by which people communicate with one another over some distance. In a narrow sense, a teleconference is a two-way, interactive meeting, between relatively small groups of people (approximately 1 to 10 at each end), who usually use permanent teleconferencing facilities. A teleconference involves audio communication between the locations, but may also involve video, graphics, or facsimile.

Teleseminar. A teleseminar is utilized for educational purposes; it is primarily one-way communication to many destinations from one source. A teleseminar almost always uses audio communication, and may also use video and some form of graphics. The source location usually has just lecturers, and the destination locations are usually classrooms or meeting rooms with no more than 20–30 people in each location. There are typically 2–10 receiving locations. A means is provided for the receiving locations to ask questions of the instructor via microphones or telephone handsets. The facilities are usually built or converted especially for teleseminars. Many large corporations utilize teleseminars for continuing education purposes to keep a widely distributed work force equally informed on the latest topics.

More sophisticated training operations utilize viewer response systems. These small desktop devices let students respond to questions presented by the lecturer. At each classroom location, all such devices connect to a computer, which immediately tabulates the responses to the question and sends them back over a separate circuit to the lecturer. Such systems can be used to give tests or to provide feedback from students to teachers to see if they are really learning the material.

Telemeeting. A telemeeting is often called an ad hoc teleconference, with the ad hoc referring to places, times, participants, and purpose. A telemeeting is similar to a teleseminar in that it is primarily a one-way communication, usually staged or prepared by video program professionals. It may be set up to order, using temporary equipment or circuits, and typically utilizing hotel facilities. Events include stockholders' meetings, new product introductions, sales meetings, promotions, and press conferences, and thus involve large numbers of people at all locations. Audio and video are almost exclusively used to communicate and, as with a teleseminar, a means is usually provided at the receiving locations for questions to be asked of those at the transmitting location. Corporations that have many locations spread out across the country often lease satellite transponders and use them for telemeetings to stay in touch with their employees, disseminating information and answering questions.

Computer conferencing. Computer conferencing is a method for people to communicate by using computers. The medium is quite flexible, as it can be used between just two people, between one and many people, or among many people. Basically, computer conferencing involves typing a message on a computer terminal and transmitting it to one or more destinations electronically.

Networks such as the Internet or corporate "intranets" (wide area networks) or building networks (local area networks, or LANs) are required to accomplish computer conferencing between many users; the other extreme, a simple data modem and telephone circuit, can allow two people at a time to conference. Networks allow both real-time and non-real-time conferencing. Non-real-time conferencing is like an electronic mailbox, commonly called e-mail, whereby, for example, a paper is distributed to certain people via computer with a request for comments; each person could transmit comments via computer to the originator. Real-time conferencing is fairly common today between many users; it occurs via chat rooms hosted by Internet service providers, schools, companies, and many other organizations. Another form of real-time conferencing between two or more people is exemplified by the AOL instant messaging type of service. Computer conferencing is widely used within offices that contain networks of personal computers (LAN). Telecommunications links can extend such networks, and therefore computer conferencing, outside of one building to a number of geographically diverse locations (WANs). The building of the Internet has greatly facilitated computer conferencing. The spread of personal computers into households and the growth of on-line services, which have enabled computers to connect to the Internet, have stimulated other forms of computer conferencing, such as bulletin boards, list servers, and the sharing of digitized music and pictures. *See* DATA COMMUNICATIONS; ELECTRONIC MAIL; LOCAL-AREA NETWORKS; MODEM; WIDE-AREA NETWORKS.

Audio communication. As a minimum, teleconferencing audio systems utilize a device such as a speakerphone to allow hands-free operation. More typically, microphones are either clipped to people or attached to a table or ceiling, and loudspeakers are mounted in a wall, cabinet, or ceiling. Depending on the circuits utilized, the audio performance can vary greatly in quality with corresponding variations in cost and complexity. Scrambling is used in some systems to assure communications privacy. *See* COMMUNICATIONS SCRAMBLING; LOUDSPEAKER; MICROPHONE; TELEPHONE.

Video and videographic communication. Video systems use one or more cameras as image sources, and

television receivers, monitors, or large-screen projectors as a means of display. Cameras may be used to view people, objects, chalkboards or easels, slides, transparencies, or text on paper; video tape players and computers are also used as sources. Again, as a function of the communications equipment and circuits used, the nature and quality of the video images may range from that of broadcast television to that of slow scan or freeze frame. Slow scan paints a new picture on the screen in 10 to 60 s, but nevertheless succeeds in providing a perception of presence of people and good-quality videographics. High-speed digital compression codecs are also used to transmit images, which are near full motion, over telephone trunks. Audio and graphic information may also be sent through these codecs, thus allowing one circuit to be used for all forms of communication. *See* MAGNETIC RECORDING; TELEPHONE SERVICE; TELEVISION; TELEVISION CAMERA; TELEVISION RECEIVER; VIDEOTELEPHONY.

Multimedia conferencing. Video codecs have been reduced in size to fit inside personal computers. Low-resolution video cameras have been reduced to smaller than a computer mouse. Many personal computers contain high-fidelity sound cards that drive stereo speakers and allow microphone inputs. Computer modems provide the capability to store and dial telephone numbers. Personal-computer-based software has been written to integrate all of these functions, thereby turning a personal computer into a personal teleconferencing terminal. Such software typically divides a computer screen into quadrants, two of which are allocated for incoming and outgoing video and two of which are allocated for applications that are shared and worked on jointly during the teleconference; these applications may be computer based, such as word processing or spread sheets, or may be video graphics or facsimile. *See* MICROCOMPUTER; MULTIMEDIA TECHNOLOGY; SOFTWARE.

Graphic communications. Numerous devices are available that allow nonvideo graphics to be used in conferencing. These include electronic blackboards and writing tablets which digitize, transmit, and reproduce an exact replica of the source at the destination. Computers are also used to generate graphic material for conferences. The material can be created before a meeting, and with the computer connected into the communications system, the material can be transmitted to a similar computer and displayed at the distant end. Devices are available which allow the computer's screen to be projected onto a large screen, suitable for viewing by a conference room full of people. Computer graphics can thereby directly replace artist-generated graphics that must be turned into overhead transparencies. *See* COMPUTER GRAPHICS.

Facsimile communications. Equipment of this type basically consists of office facsimile units that are used in teleconferences to send standard-size pages of textual material between sites, either over telephone circuits or through codecs over high-speed circuits. There are more elaborate versions of the equipment available which also allow the transmitted page to be displayed via a monitor or projector at both the transmitting and receiving locations. *See* FACSIMILE.

Popular usage. Currently, the most popular form of teleconferencing by far is the audio conference. Using plain speaker-telephones, special speaker phones, corporate private branch exchange (PBX) systems, or special dial-in, conferencing services, most business people use this form of teleconference regularly. Probably next in popularity is the Teleseminar, used for "distance learning," or formal education and training. After that, the chat room and instant messenger service are the ones frequently used. The Internet has become a popular medium to facilitate all of these forms of teleconferencing, and will become more popular as high-speed connections such as digital subscriber line (DSL), cable modem, and satellite become more prevalent. *See* COMMUNICATIONS SATELLITE; INTERNET; PRIVATE BRANCH EXCHANGE.

John J. Bleiweis

Bibliography. J. L. Creighton and J. W. R. Adams, *Cybermeeting: How To Link People and Technology in Your Organization*, 1997, 2003; K. Kelleher and T. B. Cross, *Teleconferencing: Linking People Together*, 1985; B. C. Lindberg, *Future of Communications Teleconferencing—Audio, Graphics, Data and Video*, 1998; E. Margulies, *Audio Teleconferencing—The Complete Handbook*, September 1997.

Telegraphy

A method of communication employing electrical signaling impulses produced and received manually or by machines. Telegraph signals are transmitted over open wire or cable land lines, submarine cables, or radio. Telegraphy as a communication technique uses essentially a narrow frequency band and a transmission rate adapted to machine operations. *See* ELECTRICAL COMMUNICATIONS.

Early equipment devised by Samuel F. B. Morse consisted of a mechanical transmitter and receiver or register. Operators soon learned to handle messages faster by using simple manual keys and audible sounders. Subsequently, telegraph transmission and reception became mechanized. Telegraphy may also be used in other ways.

Telegraph facilities for use by the general public to transmit messages both domestically and internationally are provided by communication companies and government administrations. Special telegraph facilities include those for news services, distribution of market prices of securities and commodities, and private lines between such points as the factories and offices of a company for the exchange of messages, orders, payroll data, and inventories. Fire and police alarms are a special form of telegraphy. The armed forces have extensive fixed and mobile telegraph systems.

Telegraph codes. For manual operation, the code consists of short dot and long dash signals (**Fig. 1**). The original Morse code also used various-length

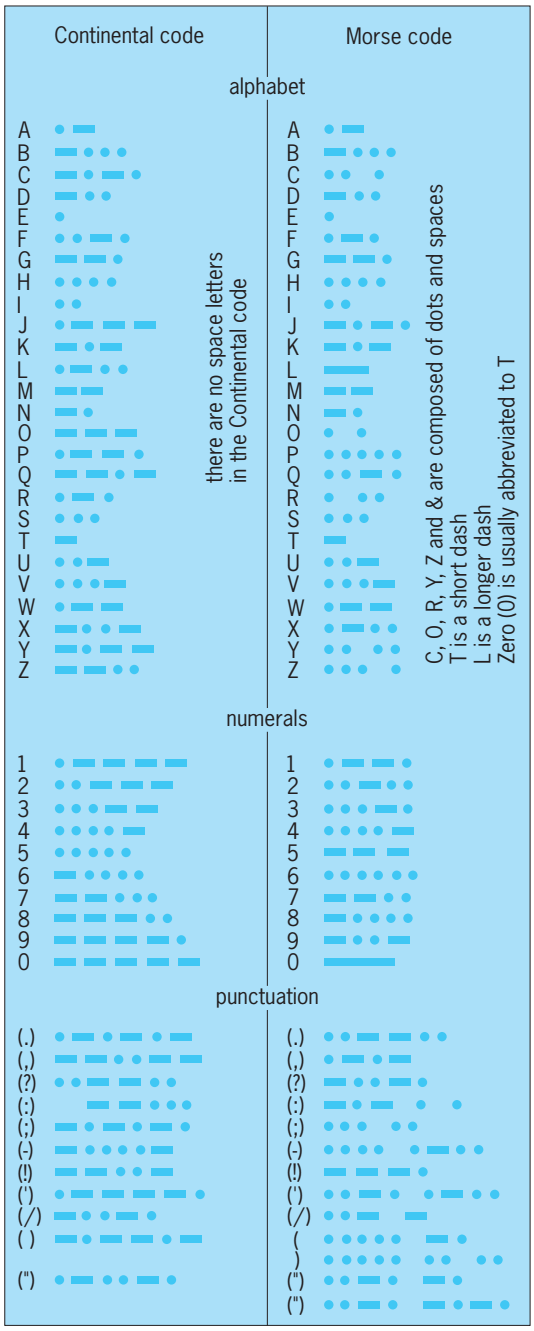


Fig. 1. Continental code is commonly used for telegraph communication. Morse code continues in use on a few land lines in United States and Canada.

spaces; the Continental code avoids them. On submarine cables, the dots and dashes are of equal length for most efficient use of transmission characteristics and are distinguished by being of opposite electrical polarity.

Most automatic printing telegraph circuits, including American cable operation, use a code of five equally spaced signals or units per letter or other symbol perforated into a paper tape (**Fig. 2**). The presence or absence of current or current reversals during these intervals constitutes the distinguishing feature. When a perforated tape is used, it usually is driven by a sprocket running in holes between the second and third code unit. Machines translate automatic teleprinter code to cable code, cable code to automatic code, or, for special applications, make other translations. For stock quotation systems and teletypesetter operation, a six-unit code is used to provide control of machine action. For data transmission or machine control, a seven- or eight-unit code may be used.

Telegraph circuits and equipment. A single circuit provides transmission in only one direction at a time. For transmission in both directions simultaneously, a duplex circuit is used. Multiplex (frequency-division) apparatus provides two, three, or more channels operable in both directions simultaneously over a single circuit. Carrier-current and frequency division techniques enable several circuits, each comprising one or more communication channels, to operate through the same wide-band wire, cable, or radio facility. *See* ELECTRICAL COMMUNICATIONS; MULTIPLEXING AND MULTIPLE ACCESS.

In automatic transmission, an operator at a manual keyboard, operated like a typewriter, perforates a tape. The tape is fed through an electromechanical or photocell tape reader that drives the tape at a rapid and uniform rate and transmits the electrical code. Interconnecting wire lines or other communication channels carry these code pulses to the receiver. The received impulses may automatically actuate a reperforator to produce a duplicate punched tape for retransmission or later transcription; the impulses may actuate a teletypewriter (also called a teleprinter) to retype the original message; or they may actuate other terminal equipment, such as an accounting machine.

Alternatively, the equipment at both terminals may be teletypewriters, in which keyboard and printing mechanisms are combined in one machine. Such machines use a five-unit code but operate without perforated tape. Business firms that originate and receive numerous messages have such equipment installed at their offices for direct service. Also widely used are facsimile instruments that transmit or reproduce a typed or handwritten message as a picture on electrosensitive paper. *See* FACSIMILE; TELETYPEWRITER.

Grosvenor Hotchkiss; William R. Webster

Message service. Individual circuits are established as needed from a nationwide network of central and branch telegraph offices and interconnecting wires, coaxial cables, and radio relay stations. The total traffic capacity of each of these interconnecting facilities may be divided by frequency-division multiplex or time-division multiplex into individual telegraph channels. Such networks of national telegraph companies and government administrations are interconnected for worldwide service.

Automatic switching. To achieve speed and accuracy in handling messages en route, direct circuits may be set up from the originating keyboard to the distant teleprinter. This is done in telex services. This mode of system operation is known as circuit switching. In previous times, if direct channels were not immediately available, or if the volume of traffic required a storage interval, the message

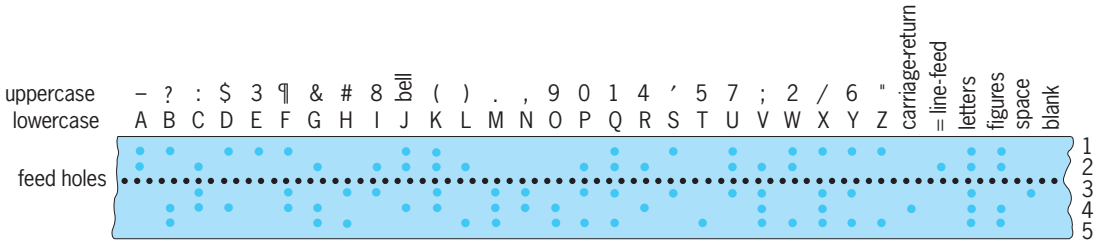


Fig. 2. Code for automatic telegraphy as it appears on a punched tape. (*Western Union Telegraph Co.*)

could be transferred by means of perforated tape at switching centers. This mode is known as message switching or store-and-forward switching. Message switching has the advantage of using the channel capacities of trunk lines more fully than do direct connections. This function is now performed by computers at central sites.

An example of message switching is the route followed by a telegram in previous times from Providence, Rhode Island, to Walla Walla, Washington. In the Western Union telegraph system, the United States had 11 area switching centers, all directly and continuously interconnected. The message from Providence went to the area switching center at Boston, Massachusetts, preceded by the director signal code PR perforated by the Providence operator. An automatic sensor at Boston received this code (meaning Portland, Oregon), actuated equipment to find an idle reperforator in the Portland trunk group, and released the fully prepared telegram from Providence into the Boston storage for Portland. After the message was completely stored in Boston, its reperforator signaled its availability to a seeker switch, the function of which was to connect reperforators with idle trunks. In this way the message was transmitted to Portland, where it was again perforated and punched on tape. The message then appeared before the Portland operator who, seeing the Walla Walla address, routed it by push button to its destination, where it was received on a teleprinter.

The reperforator switching network just described is based upon electromechanical implementation of the route-selection and transmission functions. Modern general-purpose digital computers have replaced this electromechanical switching equipment. Suitably programmed computers perform the switching function with more speed and versatility. When they are not busy routing messages, they are used to do billing, generate traffic statistics, or perform any other off-line task related to the communication operation. *See* DATA COMMUNICATIONS; DIGITAL COMPUTER.

W. E. Girardin; William R. Webster

Overseas communications. In overseas communications, telegraph messages or telegrams are transmitted over high-frequency radio (3–30 MHz), transoceanic submarine cables, or satellite communication channels. A decreasing number of messages are carried on high-frequency radio. These types of overseas channels are also used for private-line service, which is referred to as leased-channel service, and customer-to-customer teletypewriter service, which is known as Telex in the international service. *See* COMMUNICATIONS SATELLITE; RADIO-WAVE PROPAGATION; SUBMARINE CABLE.

Eugene D. Becken; William R. Webster

Bibliography. N. S. Biswas, *Principles of Telegraphy*, 1964, reprint 1968; R. L. Freeman, *Telecommunications Transmission Handbook*, 4th ed., 1998; J. Lehnert, *Introduction to Telegraph Engineering*, 1977; J. D. Ralphs, J. E. Flood, and C. J. Hughes (eds.), *Principles and Practice of Multi-Frequency Telegraphy*, 1985; D. A. Tugal and O. Tugal, *Data Transmission*, 2d ed., 1989; L. Wiesner, *Telegraph and Data Transmission over Shortwave Radio*, 3d ed., 1984.

Telemetering

The branch of engineering, also called telemetry, which is concerned with collection of measurement data at a distant or inconvenient location, and display of the data at a convenient location. One example of a complex telemetering system is used to measure temperature, pressure, and electrical systems on board a space vehicle in flight, radio the data to a station on Earth, and present the measurements to one or several users in a useful format. A simpler form of telemetering would be measuring temperature, fuel level, battery voltage, and speed in a car, and then via a hard-wire link displaying that information on the instrument panel. Telemetering involves movement of data over great distances, as in the above example, or over just a few meters, as in monitoring activity on the rotating shaft of a gas turbine. It may involve less than 10 measurement points or more than 10,000.

Telemetering involves a number of separate functions: (1) generating an electrical variable which is proportional to each of several physical measurements; (2) converting each electrical variable to a proportional voltage in a common range; (3) combining all measurements into a common stream; (4) moving the combined measurements to the desired receiving location, as by radio link; (5) separating the measurements and identifying each one; (6) processing selected measurements to aid in mission analysis; (7) displaying selected measurements in a useful form for analysis; and (8) storing all measurements for future analysis.

Aerospace systems. The largest category is commonly called aerospace telemetry, used in testing developmental aircraft and in monitoring low-orbit space vehicles. Other applications include missile

and rocket testing, automobile testing, and testing of other moving vehicles. Because of the broad range of applications for this type of system, military test ranges in the United States, through the guidance of the Range Commanders Council-Telemetry Group (RCC-TG), maintain ever-evolving Inter-Range Instrumentation Group (IRIG) standards. These standards include the IRIG 106, which deals with transmitters, modulation techniques, data formatting, data recording, and other related topics; as well as the IRIG 118 family of standards, which define test procedures for certifying telemetry equipment such as radio-frequency (RF) receivers. These documents are the standard for aerospace telemetry in the United States, and have become the de facto standard for aerospace telemetry manufacturers and users throughout most of the world.

Transducers. An aerospace telemetry system (**Fig. 1**) may send a very small number of measurements or as many as several thousand. It may send just a few measurement points per second or more than a million. In general, transducers accept various types of physical measurement variables, which may include temperature, pressure, strain, vibration, acceleration, shock, liquid or gas flow, liquid level, torque, and radiation. Each physical variable is converted into an electrical variable (current, voltage, resistance, or capacitance, for example) whose magnitude is proportional to the magnitude of the physical measurement.

Signal conditioners. For each electrical variable, a signal conditioner generates an output voltage with a standard range (−5 V to +5 V, for example), where the output represents the magnitude of the related physical measurement variable and follows it; the lowest expected physical magnitude may be represented by a −5-V output, the highest by a +5-V output, and other magnitudes in direct proportion.

Data multiplexer. All conditioned voltages are applied to a data mutliplexer, which combines all measurement voltages into a common link for movement to the receiving station. The commonly used methods of data combining are frequency-division multiplexing (where measurements are placed on individual frequency-modulated subcarriers located at different frequencies for transmission) and time-division multiplexing (where measurement voltages are sampled or commutated and placed in a specific time sequence for transmission). Since measurements occupy unique frequency or time slots, they do not interfere with each other and can be separated and identified at the receiving station. *See* MULTIPLEXING AND MULTIPLE ACCESS.

Pulse-code modulation. Most time-division multiplexing systems use pulse-code modulation (PCM), where each sampled measurement point is converted into a binary number, typically with 10 or 12 bits of resolution, and shifted out one bit at a time in a continuous data stream. In pulse-code modulation, a unique pattern signals the beginning of each data sequence or frame. *See* PULSE MODULATION.

Generally, where pulse-code modulation is used as the modulation scheme, the system not only accepts encoded analog data but also accepts information in digital formats. The most common source of such information is an on-board digital computer, but other sources may include on-off occurrences or other types of data. Generally, a system is programmed to sample each measurement point five or more times for each cycle of the highest expected data frequency.

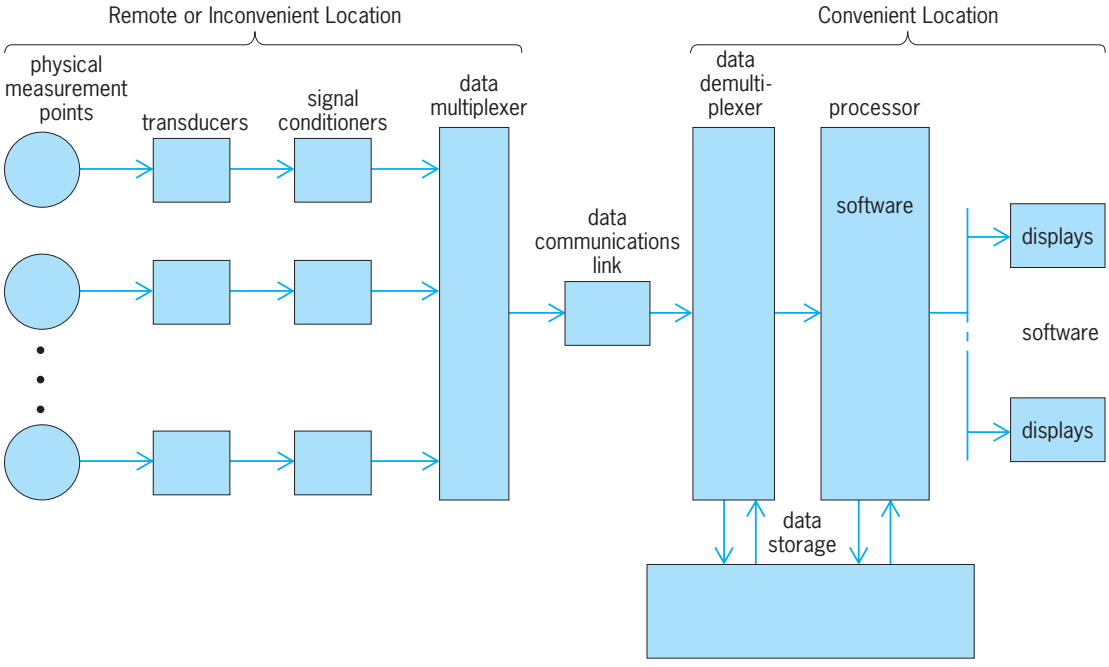


Fig. 1. Functions of a generic aerospace telemetry system.

Modulating a frequency-modulated (FM) carrier with non-return-to-zero (NRZ) PCM has been the standard for a number of years. With trends toward higher bit rates and less allowable bandwidth for telemetry users, alternate, more efficient modulation techniques have been developed. IRIG 106 now specifies Feher Patented Quadrature Phase Shift Keying-B (FQPSK-B), Feher Patented Quadrature Phase Shift Keying-JR (FQPSK-JR), Single Offset Quadrature Phase Shift Keying (SOQPSK), and Advanced Range Telemetry Continuous Phase Modulation (ARTM CPM). A 5-megabit-per-second (Mbps) data stream, for example, would require a transmission bandwidth of 12.5 MHz using NRZPCM, whereas it would require only 6.5 MHz using FQPSK-B, FQPSK-JR, or SOQPSK, and only 5 MHz using ARTM CPM. *See* FREQUENCY MODULATION; PHASE MODULATION.

Data communications link. The data communications link moves frequency-division-multiplexed or time-division-multiplexed data from the remote or inconvenient location to the convenient location by one of several methods. The most common link uses radio, where a transmitter's output is frequency-modulated or phase-modulated by the multiplexed data. Most aerospace users throughout the world operate in the 1435–1540-MHz band, the 2200–2290-MHz band, or the 2310–2390-Mhz band. While telemetry signals may be transmitted within these bands, telemetry is the primary user only through 1435-1525 MHz, 2200-2290 MHz, and 2360-2390 MHz. Telemetry signals may be transmitted in the other portions of these bands on a noninterfering basis. Alternatively, the data communications link may use fiber optics, telephone, or other communications media.

Where radio is used, generally the electrical power at the data collection site is limited, so the radio transmission power must be as low as practical (typically 2–5 W on a small missile, and 5–10 W on an airplane). To compensate for this relatively low radiated power, often a high-gain parabolic dish antenna is used at the receiving station. Further, when the vehicle under test is moving, it is necessary to steer the receiving antenna automatically or manually to follow the vehicle and thus receive data continuously as the test progresses. *See* DATA COMMUNICATIONS; RADIO SPECTRUM ALLOCATION.

Data demultiplexer. At the receiving station, a data demultiplexer identifies each measurement from the data communications link. For frequency-division multiplexes, band-pass filters are tuned to select the subcarriers, and each one is demodulated to yield measurement voltages for all channels. (The equipment is called a discriminator.) For time-division multiplexes, a frame synchronizer recognizes the unique pattern which signals the beginning of each data sequence, and counts samples to identify all measurements. (The equipment is called a decommutator.) *See* ELECTRIC FILTER; FREQUENCY-MODULATION DETECTOR.

Examination, processing, and display. In almost every aerospace telemetry system, data are routed to a processor for examination, processing, and display. Examination may consist of comparing each measurement value with operator-defined upper and lower limits and generating an alarm when any measured value goes outside those limits. Processing usually includes converting each received measurement value into a new number which denotes its actual physical value, such as degrees Celsius. Often it involves also frequency analysis or other unique calculations. Measurement displays may consist of numerical or graphical representations as selected by the individual users.

In many systems the major processing load is in a separate hardware device known as a preprocessor. This high-speed special-purpose device is under control of the computer with its unique telemetry software to ensure operation with minimum operator error. It is equipped to perform the 100 or more processing functions which are used most often in telemetry system applications, thus relieving the system computer of routine time-consuming examination and processing. With the advent of personal computers with clock speeds over 2 GHz, a lot of functions formerly performed in hardware are now being performed in software. The use of software allows for changing processing algorithms quickly as the need arises.

The system's processing power can be contained in one device, but in medium and large systems it is usually distributed between a general-purpose computer and one or more computer-powered workstations, each controlled independently to meet the needs of its user. Interconnection of the computer and display stations is by an industry-standard local-area network, such that display sites can be added or removed easily as requirements change. *See* LOCAL-AREA NETWORKS.

System software is built on a standard operating system, with special telemetry applications programs to handle the data rates and versatile processing requirements. A system uses two special databases: one defines all characteristics of the incoming data format, such as the modulation type, data rate, and unique synchronization pattern; and the other defines virtually all aspects of each measurement being received and processed, such as the measurement name (such as temperature), units of measurement (such as degrees Celsius), relationship between the physical units of measurement and the related telemetry codes being received, and other details which may be necessary for high-speed processing and display. The software also performs system diagnostics and reports on the operating condition of each component whenever electrical power is applied or on special command. *See* OPERATING SYSTEM; SOFTWARE.

Data storage. To archive data for further analysis, most systems include data storage, possibly consisting of a wide-band magnetic tape recorder at the data demultiplexer section and of a magnetic disk recorder at the processor section. When data are stored, they are annotated with the time of day so that they can be analyzed with respect to time. *See* COMPUTER STORAGE TECHNOLOGY.

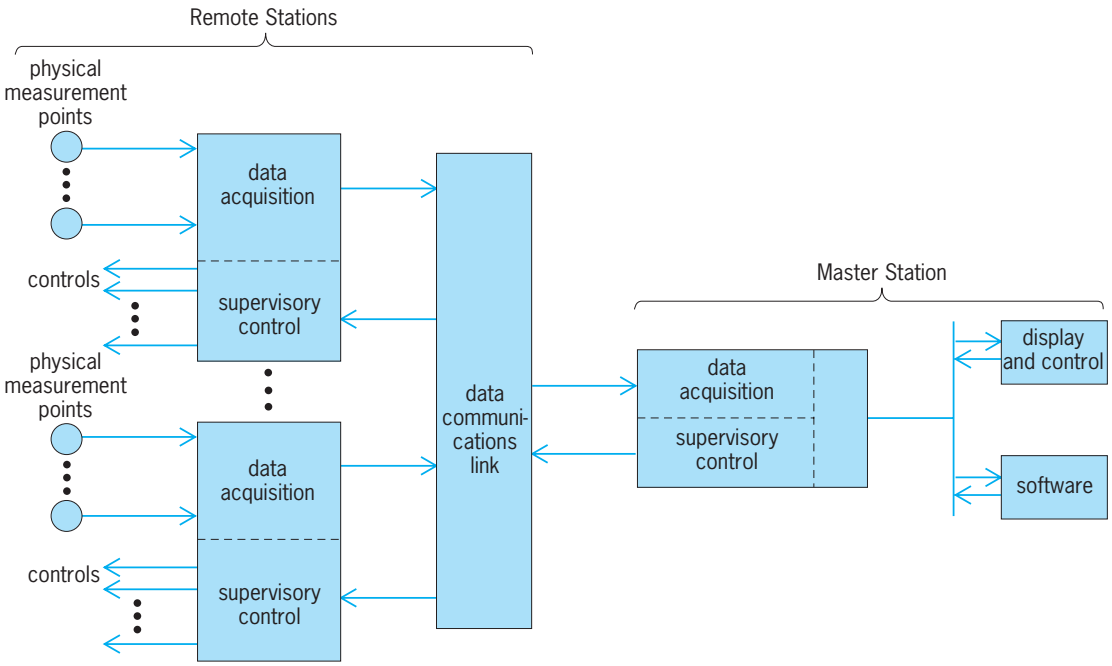


Fig. 2. Functions of a supervisory control and data acquisition (SCADA) system.

Industrial systems. The version of telemetry commonly used in an industrial application includes supervisory control of remote stations as well as data acquisition from those stations over a bidirectional communications link. The generic term is supervisory control and data acquisition (SCADA); the technology is normally used in electrical power generation and distribution, water distribution, and other wide-area industrial applications. There is no authoritative SCADA standards committee; nevertheless, these systems have many characteristics in common. *See* ELECTRIC POWER SYSTEMS; WATER SUPPLY ENGINEERING.

The number of remote stations in a SCADA system (**Fig. 2**) may be one, several, or even 100 or more. The number of data measurements per station varies over a wide range, as does the number of unique commands which can be delivered. Data or command transfer rates are quite low as compared to aerospace telemetry; generally they are 10–1000 per second.

In a SCADA system, data acquisition at a remote station is quite similar to the same function in an aerospace telemetry system, except that measurements are usually encoded into American Standard Code for Information Interchange (ASCII) characters for transfer, rather than frequency-division multiplex or time-division multiplex formats. These ASCII characters can be transferred as individual messages or grouped as a sequential message.

Data acquisition equipment at the master station brings these characters in and prepares them for the display stations, where each message is processed and displayed for the user in meaningful units for analysis. A user at any display station can initiate control messages for any remote station. These messages are relayed through the supervisory control equipment at the master station, and received by the equipment at the appropriate remote station to close or open switches, move rotary devices, initiate data acquisition, or perform other supervisory tasks. Since ASCII characters are compatible with a wide range of data communications equipment, the two-way data communications link can be commercial-grade telephone line, electric utility line, or low speed radio link. *See* REMOTE-CONTROL SYSTEM.

Each SCADA system is powered by one or more computers with versatile software. In many systems, the remote stations are computer powered also. SCADA technology is quite similar to aerospace telemetry technology in many areas. The major differences are in formatting and in the use of two-way communications.

One SCADA application involves monitoring wind direction and velocity as indicated by anemometers located on the approach and departure paths near an airport, so that air-traffic controllers can make pilots aware of dangerous differences in wind direction and velocity, known as wind shear. This type of system is operated over a radio communications link, with the appropriate anemometers being interrogated as their measurements are needed by the computer for wind analysis. *See* AERONAUTICAL METEOROLOGY; WIND MEASUREMENT.

Somewhat similar systems are used in oceanographic data collection and analysis, where instrumented buoys send water temperature and other data on command. *See* INSTRUMENTED BUOYS.

Space systems. Because two-way communication with a complex and distant Earth synchronous satellite or other spacecraft presents a unique challenge, a technology called packet telemetry is in widespread use for these applications. Here, messages between the Earth station and the spacecraft are formed

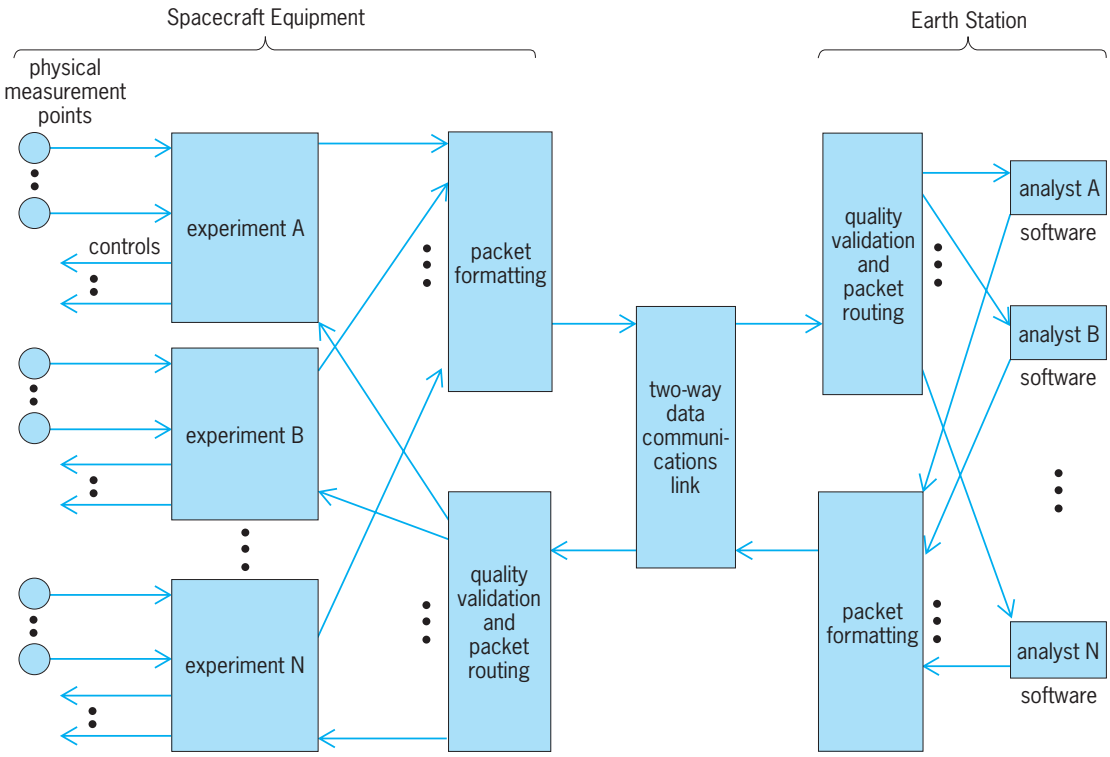


Fig. 3. Functions of a spacecraft telemetry and command system, using packet telemetry.

into groups of measurements or commands called packets to facilitate routing and indentification at each end of the link (**Fig. 3**). Each packet begins with a definitive preamble and ends with an error-correcting code for data quality validation. Packet technology is defined by an international committee, the Consultative Committee for Space Data Systems (CCSDA). *See* PACKET SWITCHING.

A spacecraft may have several on-board devices (labeled experiment A, experiment B, and so forth), each of which can accept commands from and send data to the Earth station. Several analysts are at the Earth station. Typically analyst A is concerned with experiment A, analyst B with experiment B, and so forth. When analyst A sends a command, it is put into a standard packet and sent to the spacecraft via a two-way communications link. The quality of the received packet is validated, and it is routed to the proper experiment for execution. Responses generated by the command are sent through the spacecraft's packet formatter and communications link to the Earth station. After quality validation, the response is routed to the proper analyst for processing and display. The packet technique provides optimum efficiency in use of the two-way communications link, as each analyst has access to the related experiment, subject only to occasional delays in link access.

Much of the technology in a space telemetering system is similar to that used in aerospace or industrial systems. The major differences relate to formatting, validation, and routing of packets.

Spacecraft telemetry must operate with limited electrical power. To overcome such limitations, a spacecraft may have a steerable antenna, typically 1–5 m (3–15 ft) in diameter, on board to boost the effective radio power. Often it sends data by phase-modulated rather than frequency-modulated radio, and typically it operates at a relatively low data rate in order to get the most efficiency in transmission. The ground-based station is likely to have an extremely large high-gain parabolic dish antenna, possibly as much as 30 m (100 ft) in diameter. *See* SPACE COMMUNICATIONS; SPACECRAFT GROUND INSTRUMENTATION.

Other systems. The three major families of telemetering systems (aerospace, industrial, and space) each use equipment in which simplicity is less important than standardization. Often a user is willing to use equipment which has higher performance than the application requires but which is readily available from one or more manufacturers. Certain special applications fall outside the three families, however, and some of those users have chosen to build special equipment which is optimum for their needs.

One unique example is the telemetering system used to monitor the condition of individuals in a cardiac care unit of a hospital. A single channel data-acquisition unit sends heartbeat data from a given individual over a low-powered radio transmitter (operating in the 460–470-MHz band) to a central receiving station nearby. There, a specialist receives and monitors the heartbeats of all individuals simultaneously and is able to respond quickly if a problem occurs.

A special multichannel medical telemetry system is used on some ambulances to monitor and radio vital signs from a person being transported to a

hospital, so that medical staff can prepare to treat the specific condition which caused the emergency.

Another unique system is possibly the oldest user of radio telemetry, the radiosonde. A data collection and transmission system is lifted by a balloon to measure and transmit pressure, temperature, and humidity measurements from various altitudes as an aid to weather prediction. *See* BALLOON; METEOROLOGICAL INSTRUMENTATION.

A system using telemetry multiplexing and recording technology is utilized on most operational commercial aircraft for recording selected data measurement points. In case of crash, the recording medium from this flight data recorder can be recovered from the crash scene and analyzed to help determine the cause.

O. J. Strock; Tomas C. Chavez

Bibliography. F. Carden, R. Jedlicka, and R. Henry, *Telemetry Systems Engineering*, 2002; S. Horan, *Introduction to PCM Telemetering Systems*, 2d ed., 2002; International Telemetering Conference, *Proceedings*, annually; E. M Rueger and O. J. Strock, *Telemetry System Architecture*, 3d ed., 1995.

Teleostei

The largest, youngest (first appearing in the Upper Triassic), and most successful group of the actinopterygians (rayfin fishes), and a sister group of the Amiidae (bow fins). The 23,600 species of teleosts make up more than half of all recognized species of living vertebrates, and over 96% of all living fish species. There are 494 families, of which 69 are extinct leaving 425 extant families, 43% of which have no fossil record. However, the fish beds at Monte Bolca in Verona, Italy (lowermost Eocene), provide the earliest records for a large number of teleost higher taxa. *See* ACTINOPTERYGII; AMIIFORMES.

Morphology. Much of the evidence for teleost monophyly and relationships comes from the caudal skeleton and concomitant acquisition of a homocercal tail (upper and lower lobes symmetrical). This type of tail primitively results from an ontogenetic fusion of centra (bodies of vertebra) and the possession of paired bracing bones located bilaterally along the dorsal region of the caudal skeleton, derived ontogenetically from the neural arches (urolneurals) of the ural (tail) centra. The presence of uroneurals is a synapomorphy (homology) for all teleosts and is recognizable in the fossil groups, as is another synapomorphy—the articulation of the first two hypurals with the same vertebral centrum (ural centrum 1, which results from an ontogenetic fusion of two centra). Other characters of the teleosts include the mobile premaxilla bone, the extension of the posterior myodome (the eye muscle canal) into the basioccipital bone, and the development of the swim bladder.

Phylogeny. By the end of the Jurassic, several of the taxa with Recent representatives had evolved; these include the Osteoglossomorpha (bonytongues; *Arapaima*, mormyrids and allies; 217 species), Elopomorpha (tarpons, eels, halosaurs; 801 species), and Clupeomorpha (herrings; 360 species), but it was in the succeeding Cretaceous Period that the great radiation of the group occurred, resulting in some 20 or more orders (see **illus.**). *See* GEOLOGIC TIME SCALE; OSTEICHTHYES.

Since the Late Cretaceous, the teleosts have been by far the largest and most diverse group of vertebrates, occupying a plethora of habitats from the deepest ocean abyss to the highest mountain lakes—all of which makes the unraveling of their evolutionary history a daunting task. Consequently, ichthyologists are still actively involved in reconstructing teleost phylogeny, particularly that of the euteleosteans.

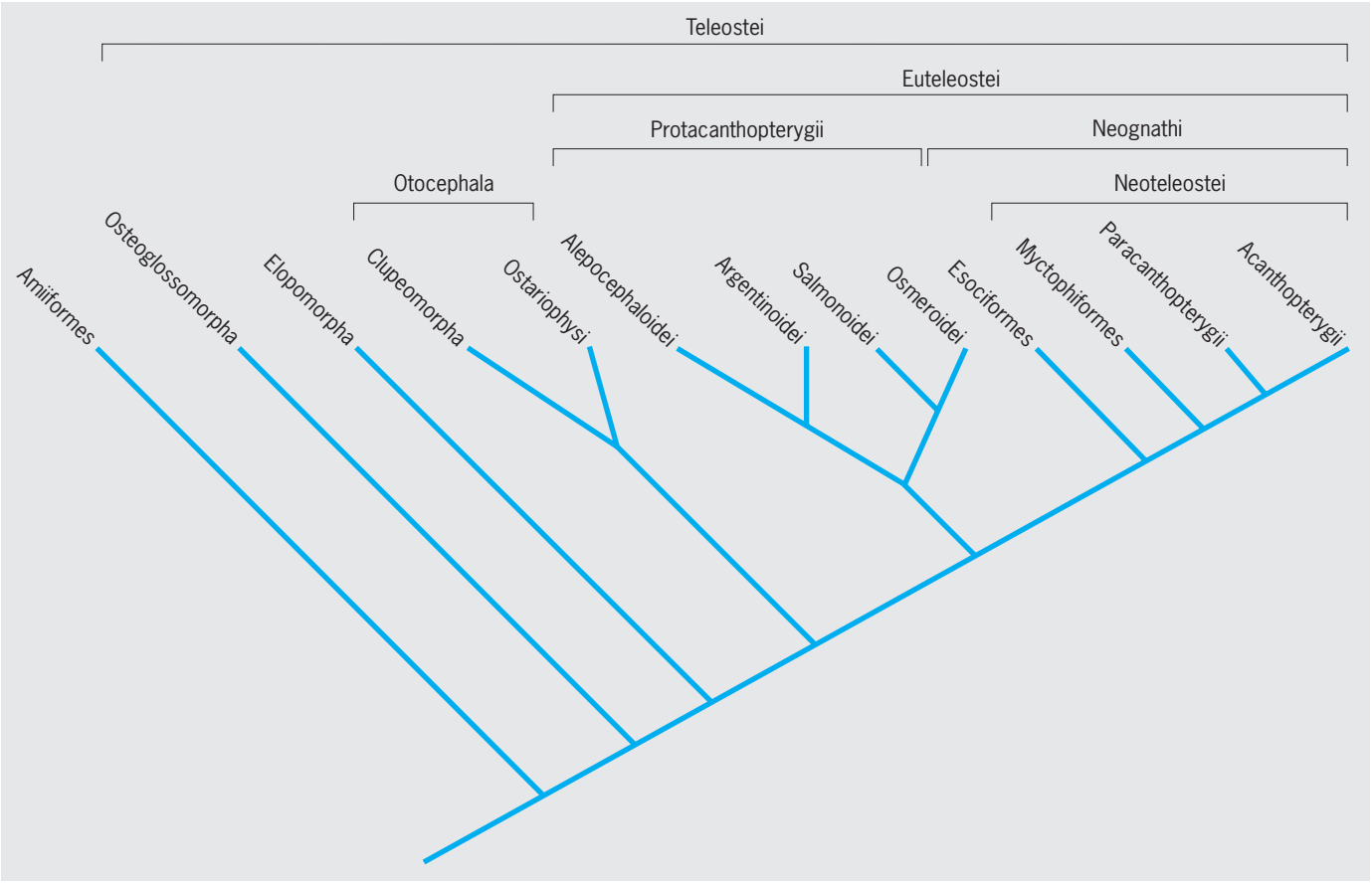
Classification. The Osteoglossomorpha, Elopomorpha, and Clupeomorpha are now generally regarded as successive clades (groups) above the level of the fossil, paraphyletic pholidophorids. The Clupeomorpha is considered to be the sister group of the Ostariophysi.

Clupeomorpha and Ostariophysi. The Clupeomorpha includes almost 80 genera and some 360 Recent species in three main families: Engraulidae (anchovies), Dussumieridae (round herrings), Clupeidae (herrings), as well as Pristigasteridae. The Ostariophysi make up nearly 75% of the fresh-water fishes of the world (over 6000 species) and include the Cypriniformes (carps, loaches, and relatives), Siluriformes (catfishes), Gymnotiformes (knife fishes, electric eel), and *Chanos*. They are characterized by the development of a bony chain of ossicles and associated structures connecting the swim bladder with the inner ear, the so-called Weberian apparatus. Interestingly, in the clupeomorphs, there is a direct connection between the diverticulum of the swim bladder and the labyrinth of the ear, with the diverticulum passing through an opening in the exoccipital bone and terminating in two distinct vesicles (one in the prootic, the other in the pterotic). Thus, in both ostariophysans and clupeomorphs, stimuli are transmitted from the swim bladder to the utriculus. However, the nature of the two otophysic (ear–swim bladder) connections are not regarded as homologous. *See* CYPRINIFORMES; OSTEOGLOSSIFORMES; PHOLIDOPHORIFORMES.

It has been shown that the Clupeomorpha plus Ostariophysi shared an ancestry with all the remaining teleosts. In other words, the Clupeomorpha plus Ostariophysi are the sister group of the Euteleostei.

Euteleostei. The Euteleostei are by far the largest teleost taxon with more than 22,000 species arranged in some 340 families. They comprise two major lineages, the Protacanthopterygii and the Neognathi (see illus.).

The Protacanthopterygii (often regarded as lower euteleosteans) include four groups: the salmonoids (salmon and allies; 66 species) and the osmeroids (smelts, salangrids, *Lepidogalaxius*; 72 species), which together are the sister group of the alepocephaloids (slickheads; 60+ species) plus the argentinoids (argentines or herring smelts; 60+ species). The Neognathi, on the other hand, are



Phylogeny of the Teleostei.

made up of the esocoids (pike and mudminnows; 10+ species) plus the sister group Neoteleostei (see illus.). *See* SALMONIFORMES.

The Neoteleostei, with more than 15,319 species, comprise four main clades, the Stomiiformes, Aulopiformes, Myctophiformes, and Acanthomorpha (paracanthopterygians plus acanthopterygians); the Myctophiformes are the sister group of the Acanthomorpha. The basal clade of the Neoteleostei is the Stomiiformes, a morphologically diverse group (321 species) of deep-water oceanic fishes with unique photophores living down to 1000 m. In contrast, the Aulopiformes (219 species) are a morphologically diverse group of benthic (deep-water) and pelagic (surface-water) fishes that range in habitat from the abyss to estuaries.

The Neoteleostei include three groups of very advanced fishes, the Myctophiformes (242 species), paracanthopterygians (1212 species), and Acanthopterygians (13,414 species). The paracanthopterygians, like the Aulopiformes, contain various deep-sea groups, including the Lophiiformes (angler fishes; 297 species), and the Gadiformes with about 482 species, many of which are important commercial fishes that constitute over one-quarter of the world's marine fish catch. Other paracanthopterygian groups include the fresh-water Percopsiformes (trout and pirate perches and cave fishes; 9 species) and the Batrachoidiformes (coastal, benthic toadfishes; 69 species). *See* BATRACHOIDIFORMES; GADIFORMES; LOPHIIFORMES; PERCOPSIFORMES.

The acanthopterygians are the largest subgroup of the Euteleostei, distributed among 12 orders and 218 families (whose interrelationships are mostly unknown). The Perciformes form the largest of these orders with 9293 species (many members of which bear spines on their fin). Of the more advanced members, the Scorpaeniformes (mail-cheeked fishes), with 25 families and about 1271 species, are second to the Perciformes in species richness, whereas the Pleuronectiformes (flat fishes) with 570 species and the Tetradontiformes (trigger fish and puffer fish) with 339 species are a distant third and fourth. Members of a lower clade include the Gasterosteiformes (sticklebacks, sea horses, pipe fishes, and so on; 257 species) and the Atherinomorpha (rainbow fishes, silver sides, blue eyes; 285 species). *See* GASTEROSTEIFORMES; PERCIFORMES; PLEURONECTIFORMES.

Biology. An important characteristic of the Teleostei is the presence of a swim bladder, a gas-filled hollow organ lying between the gut and kidneys, which develops as a diverticulum from the dorsal wall of the esophagus. In the basal groups, the swim bladder usually remains in contact with the esophagus via the pneumatic duct. In more advanced groups, such as the myctophoids, this connection may be lost, and a new structure—the

gas gland—develops. Thus in the myctophoids, the principal function of the swim bladder is hydrostatic and gas can be secreted into or removed from it. Consequently, myctophoids can undergo diurnal migrations of several hundred meters. *See* SWIM BLADDER.

In some osteoglossomorphs, clupeomorphs, and euteleosts there is a direct otophysic connection between the labyrinth of the ear and the swim bladder, whereas in ostariophysans there is an indirect connection via a paired chain of bones. There can be little doubt that these Weberian ossicles allow the swim bladder to function as a manometer or pressure receptor, since the behavior of the loaches is used to predict weather changes. In the Clupeomorpha there is also a connection between the swim bladder and the lateral line, and it has been shown that these swim bladder–ear connections in the herrings are important in shoaling behavior. Finally, the swim bladder can also function as a resonator when beaten by the ribs, as in certain toadfishes. Brian Gardiner

Bibliography. J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, New York, 1994; J. R. Paxton and W. N. Eschmeyer (eds.), *Encyclopaedia of Fishes*, 2d ed., Academic Press, San Diego, 1998; M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson (eds.), *Interrelationships of Fishes*, Academic Press, San Diego, 1996.

Teleostomi

A monophyletic grade of craniates comprising Acanthodii, Actinopterygii (ray-finned fishes), and Sarcopterygii (coelacanths, lungfishes, and tetrapods). [Monophyletic refers to any form evolved from a single interbreeding population. Craniates are vertebrates distinguished by a cranium.] The Acanthodii (all fossil species) is the sister group to the actinopterygians and sarcopterygians, a relationship based on otoliths and certain details of the vertebral column and associated elements. The term Teleostomi, coined by C. L. Bonaparte in 1836, has changed in meaning, as have Acanthodii and Actinopterygii, which originated with E. D. Cope in 1871. The term Osteichthyes, first used by T. H. Huxley in 1880, is not used by J. S. Nelson in a formal taxonomic sense; however, the term is conveniently used for the vast number of fishes with a bony skeleton. The term Sarcopterygii was used by A. S. Romer to include the lobed-finned fishes, that is, crossopterygians and dipnoans; however, E. O. Wiley, and D. E. Rosen and collaborators include the tetrapods in Sarcopterygii to form a monophyletic group. The grade Teleostomi contains about 53,633 extant valid species, of which 26,891 are actinopterygians and 26,742 are sarcopterygians. *See* ACANTHODII; ACTINOPTERYGII; OSTEICHTHYES; SARCOPTERYGII. Herbert Boschung

Bibliography. J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006; D. E. Rosen et al., Lungfishes, tetrapods, paleontology, and plesiomorphy, *Bull. Amer. Mus. Nat. Hist.*, 167(4):159–276, 1981; E. O. Wiley, Ventral gill arch muscles and the interrelationships of gnathostomes, with a new classification of the Vertebrata, *Zool. J. Linn. Soc.*, 67:149–179, 1979.

Telephone

An instrument containing a transmitter for converting the acoustic signals of a person's voice to electrical signals, a receiver for reconverting electrical signals to acoustic signals, and associated signaling devices (the dial and alerter) for communicating with other persons using similar instruments connected to a network. The term "telephone" also refers to the complicated system of transmission paths and switching points, called the Public Switched Network (PSN), connected to this instrument. This article discusses only the instrument (also called a set); for a discussion of the telephone system *see* TELEPHONE SERVICE.

Telephone set design underwent a steady evolution for a century from the first instruments of A. G. Bell and T. A. Edison. During this period a standard interface was maintained between the telephone instrument and the telephone system network. A basic telephone instrument provides plain old telephone service (POTS). The interface between the instrument and the network is two copper wire conductors brought into the subscriber's premises. These are called ring and tip. Sometimes a third conductor (ground) and a fourth (sleeve) are added to provide additional capabilities such as selective ringing on party lines and coin collection on coin telephones. The invention of the transistor and developments in digital electronic circuits have allowed substantial changes in this connection method, which are exploited in wireless and some business telephones, but the copper tip and ring arrangement is still used for basic telephones. Bell's original invention, the creation of an electrical analog to the acoustic signal of a person's voice, continues to be the basis for the speech-related circuitry of a POTS telephone.

A POTS or "standard desk" telephone set contains eight basic parts: transmitter, receiver, antisidetone network, dial, alerter, switchhook, connecting cords (wires), and chassis. Added to these may be other feature-oriented hardware (and computer programs) such as electronic memory (for dialing preprogrammed numbers), loudspeakers and microphones (for hands-free use), a voice synthesis unit, a voice recorder (for storing messages), and a video screen (for displaying text, graphs, or an image of the person at the other end of the connection). The set's complexity, depending on its intended usage, can approach that of a small computer. The telephone's original function, that of permitting voice communication with another person, may therefore be enhanced to permit communication with computers, automatic reporting of emergencies, and use of written or graphic information.

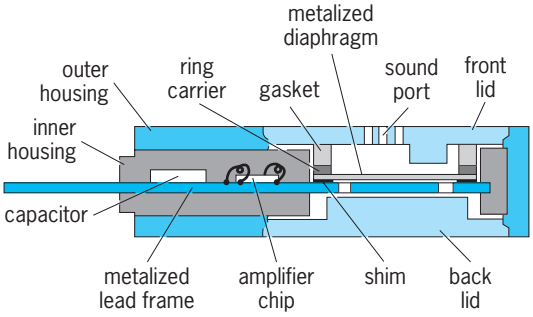
Transmitter. The transmitter is a transducer that converts acoustic energy into electric energy. E. Berliner's invention of the carbon transmitter was

the key to practical telephony because it amplified the power of the speech signal, making it possible to communicate over distances of many miles. In the carbon transmitter, sound pressure on the diaphragm varies the pressure on the carbon. Changes in pressure on the carbon granules vary the resistance of the transmitter, causing the magnitude of the direct current supplied by the central office to change in proportion to the sound, a process called modulation. *See* MODULATION.

Current designs are based on the charged electret (a condenser microphone) or on electrodynamic principles. Both the electret and electrodynamic transmitters use transistors to provide needed power gain; they introduce less distortion than the carbon transmitter. *See* ELECTRET TRANSDUCER; MICROPHONE; TRANSDUCER; TRANSISTOR.

Transmitters in a basic set have a frequency-response range intentionally limited to 250–5000 Hz to improve voice clarity on the Public Switched Network. Sets that use another network for transmission may have a wider frequency response. Other components in the Public Switched Network may have even less bandwidth than the transmitter provides. The frequency response of the transmitter rises uniformly to a broad maximum in the region of 2500 Hz. Even though normal human hearing has a much broader frequency response, speech heard on the telephone resembles closely that heard by a listener a few feet from the person speaking. Confusion is possible between "f" and "s" sounds, but the listener usually subconsciously resolves this by the context of the speech. *See* HEARING (HUMAN).

The heart of an electret transmitter (see **illustration**) is an electrical capacitor formed by the metal on the diaphragm, a conductive coating on top of the metalized lead frame, and the plastic and air between the metal layers. The diaphragm is made of a special plastic that can be given a permanent electrostatic charge (analogous to the magnetization of a permanent magnet). As sound waves entering the sound port cause pressure changes, the diaphragm moves closer to and farther away from the metalized lead frame. This changes the value of capacitance and produces a varying electric voltage which is the analog of the impinging sound-pressure wave. The signal is amplified by the internal amplifier chip to a level that is suitable for transmission on the telephone network.



Cross section of an electret transmitter.

Receiver. The receiver transducer operates on the relatively low power used in the telephone circuit; it converts electric energy back into acoustic energy. As in the transmitter, careful design of the relationship of the acoustical and electrical elements produces a desired response-frequency characteristic. Usually a rising response over the frequency range 350–3500 Hz is desirable to match the transmission characteristics of the traditional telephone network, optimize intelligibility, and minimize noise. However, sets intended for use with networks other than the traditional network may have a wider and flatter frequency response.

Historically, there were two common types of receiver units with fixed-coil windings, the ring armature receiver and the bipolar receiver. Fixed-coil receivers were designed for close coupling to the ear and were placed in a cuplike enclosure shaped to fit closely to the ear. Moving-coil designs, similar to loudspeakers, have replaced these in most telephones. This change can provide hands-free (speakerphone) capability in an ordinary handset but, because an ear-shaped cup is often not used, results in poorer performance in noisy environments. *See* LOUDSPEAKER.

Anti-sidetone network. The anti-sidetone network has two basic functions: to reduce sidetone (explained below), and sometimes to provide equalization (reduce the variation in loudness that is caused by varying lengths of cable between the central office and different subscribers). However, the incorporation of volume controls in many sets makes equalization unnecessary. This network may also provide more efficient coupling of the transducers to the telephone line if it incorporates a transformer.

Sidetone is the sound of the speaker's voice reaching his or her own ear. Normally this occurs through the air. However, when one is using a telephone handset, the acoustic sidetone path to one ear is replaced by an electric sidetone path through the telephone's transmitter, network, and receiver. Since the transmitter amplifies power impinging on its diaphragm, this new path may result in a louder sound in the ear coupled to the receiver than would exist if the receiver were not being used. This sound is unpleasant and can also cause the telephone user to talk more softly, thereby making it harder for the person at the far end to understand the conversation. Anti-sidetone networks may consist entirely of passive electric components such as transformers, capacitors, and resistors; or may include active components such as transistors embedded in integrated circuits. *See* ELECTRIC FILTER; INTEGRATED CIRCUITS.

Dial. The dial was originally designed to permit selection of a particular far-end station by operation of switches in the central office. The all-mechanical rotary dial used for many years contains a pair of electric contacts that periodically interrupt the power flowing from the central office. The resulting direct-current pulses are generated at the rate of 10 per second. Dual-tone multifrequency (DTMF) codes are used for the same purpose in Touch Tone telephones.

In this system, two audio tones selected from a 3×4 matrix are used to represent each digit. The tones are selected by a corresponding 3×4 matrix of pushbuttons. Each time a button is pushed, the two tones are sent simultaneously. The tone pairs are selected so that they are unlikely to occur in natural speech, thereby avoiding the inadvertent simulation of digits by speech during the normal dialing interval. These tones have an advantage over the low-frequency direct-current pulses in that they can be transmitted over the transmission network to the far-end station (as is the electrical analog of the voice signal), whereas low-frequency pulses normally will not carry beyond the local central office. Used this way, the multifrequency tone signals can convey information to a so-called intelligent telephone such as an answering machine or to a computer at the far end of the connection. *See* SWITCHING SYSTEMS (COMMUNICATIONS).

Alerter. The alerter, also called the ringer, which normally is powered by a 20-Hz 88-V alternating current, serves to alert the subscriber to an incoming call. In some older party-line systems, selective ringing of different sets on the common loop (one pair of wires connecting several different subscribers to the central office) was accomplished by using several different frequencies, one for each subscriber or by using the ground conductor mentioned previously. Mechanical ringers consist of a clapper that periodically strikes a gong, the bell. Both the electrical circuit that powers the clapper and the mechanical elements of the clapper are made resonant at the right frequencies to maximize sound output and minimize sensitivity to signals of other frequencies. A nonlinear magnetic circuit serves to present a higher impedance to speech signals on the telephone line than to the ringing signal. Electronic (or tone) ringers make use of electrical oscillator design techniques to produce their sound. They also contain nonlinear circuitry to emulate the characteristics of the mechanical ringer's nonlinear magnetic circuit. The frequency selectivity of the ringer helps prevent ringing when dial pulse or noise signals are present on the telephone line. *See* OSCILLATOR.

Switchhook. The switchhook contains a set of electrical contacts that interrupt the flow of current from the central office whenever the set is "on hook." The closing of these contacts when the handset is lifted signals the central office that the set's user wants to initiate or answer a call.

Connecting cords. Telephone sets may include two cords: a line cord that connects the instrument to tip and ring, and a handset cord that connects the telephone handset to the chassis.

Line cords. Basic telephone sets require two-conductor line cords, while more complicated sets, with capability of two or more lines and other special services, use four, six, eight, or more conductors in their cords. Specialized service requirements may require twisted-pair construction to reduce crosstalk problems between multiple-voice circuits, or between data and voice circuits. Line cords are typically constructed with individually insulated stranded copper conductors covered with an outer jacket. Material selection for both the conductor insulation and the outer jacket is critical in determining mechanical strength, chemical resistance, protection against voltage hazards, appearance, and cost. *See* CROSSTALK; DIELECTRIC MATERIALS; ELECTRICAL INSULATION.

Handset cords. Handset cords typically use four conductors. Two conductors connect to the telephone receiver, and the other two to the telephone transmitter. Telephone handset cords need to be extremely flexible and to withstand a great deal of physical abuse. The handset cord conductors are of a special construction, called tinsel, that provides superior mechanical strength, flexibility, and fatigue properties. A tinsel conductor consists of a nylon or polyester thread, or center core, with three or four thin copper-alloy ribbons wrapped helically around it. These thin, springlike conductive ribbons are extremely flexible and are able to withstand the pulling, bending, and twisting loads to which handset cords are subjected. During manufacture, handset cords are coiled, or made retractile, by wrapping the cord around a heated rod called a mandrel. After cooling, the cord retains its coiled shape.

Modular connectors. Usually, line cords are connected to the telephone network with standardized modular plugs and jacks (connectors). This replaces the former practice of hard-wiring instruments to their network connection. The mating surfaces in the plug and jack are usually gold-plated to ensure good electrical contact. A flexible locking tab on the modular jack secures the modular plug and is depressed to disconnect the cord. This system permits the telephone to be moved easily from one location to another. Modular connectors are also used on the handset cord, making cord replacement easy, an important consideration since cords are subjected to heavy wear and abuse.

Chassis. The chassis provides a mechanical support to hold all the other parts together. It must be capable of withstanding considerable abuse as telephones are often inadvertently dropped. The chassis must also present the components in a way that is convenient for the user of the telephone. The dial in desk sets, for example, should be mounted at an angle easy to use.

Special features. Many attempts were made in early telephones to add memory to the telephone dial. This feature was never popular, however, until digital circuits were used instead of mechanical devices to store telephone numbers. A display, included in some sets, allows the user to see what telephone number has been stored. This same display can also give the time and date when not showing telephone numbers or can display the number of the calling party. Another popular feature is "hands-free" telephony, which substitutes a microphone and loudspeaker for a handset-mounted transmitter and receiver.

Communication with machines. Applications of telephony in which one of the two parties communicating is a machine have become common. One

example is telephone banking, a service using the DTMF dial to send numerical information such as account numbers, passwords, transaction codes, and dollar amounts from a human subscriber to the service provider's computer on the other end of the connection. Another example is the automated receptionist function where the caller is asked to dial a digit corresponding to the service desired (make a doctor appointment, speak to a nurse, fill a prescription, and so forth).

Features for hearing-impaired users. Special features are provided for hearing-impaired users. These include small amplifiers for either the transmitter or the receiver that are built into the handset, devices analogous to ringers that provide visual signals, and receivers having a leaky magnetic field that can be picked up by specially equipped hearing aids. For the totally deaf, a machine with a keyboard and alphanumeric display may be used in place of an acoustic instrument.

Cordless telephones. Cordless telephones use a two-way radio link between one or more handsets and a base unit; this replaces the handset cord. The base unit is connected through a normal line cord to the telephone network. Typically, the dial, alerter, and switchhook functions are built into the handset, providing full calling and answering capability on a portable basis within a radius of several hundred feet of the base unit. The base unit normally requires a connection to alternating-current power, as well as to the telephone network, while the handset uses batteries that are recharged when the handset is replaced on the base unit. Allowable radio power between the handset and the base unit is severely limited by law, resulting in a telephone suitable for use around a residence but not suitable for longer distances. *See* RADIO SPECTRUM ALLOCATIONS.

Wireless telephones. The cellular radio system, a form of mobile radio telephony, serves the need for a telephone with which the user can wander over a large geographical area while making a call. Cellular telephones use a two-way radio link to replace the line cord; thus tip and ring are not brought to the telephone. The telephone communicates with one of a number of base stations spread in cells throughout the service area. As the telephone user changes location, the link is automatically switched from cell to cell to maintain a good connection. The transmitted radio power for a cellular telephone is larger than that for a cordless telephone, resulting in the larger service range. Cellular telephones operate on different radio frequencies from those used by cordless telephones. Both analog and digital radio links are used; analog links are being replaced by digital to provide more channels and longer operational time on a battery charge. Several different protocols are used by providers between the telephone and base station, so the telephone used must be chosen to be compatible with the service provider's network. Also, enhanced features, such as text messaging and transmission of photographs or video, are available in some protocols and require telephones having the necessary hardware and software to support them. *See* MOBILE COMMUNICATIONS. Richard M. Rickert

Business terminals. Business telephones designed for use on a customer's premises may connect to an on-premises switching system known as a private branch exchange (PBX), or a central office switch. They generally offer the ability to support multiple-line appearances without requiring more than a single port on the switch. Multiple-line appearances provide the ability to receive or originate additional voice or data calls at the terminal while still being engaged in a primary voice call, and the ability to bring additional parties onto a primary call (forming a conference call) or to transfer the primary call to a second party. *See* MICROCOMPUTER; PRIVATE BRANCH EXCHANGE; TELECONFERENCING. Peter S. Warwick; Wesley L. Shanks

Digital transmission and VoIP. Digital transmission of speech is used in long-distance telephony because digital signals can be regenerated precisely and thus are much less subject to degradation as a function of increasing transmission distance than are analog signals. The cost of converting the analog signal back and forth from the digital format usually makes this mode of operation uneconomic for local loops (those within a single central office). However, when all or a major part of the information flow from a set to a central point is already in digital format, as in the wireless telephones and business terminals discussed above, it is economical to digitize the speech at the set. In this arrangement, all the information flowing from the instrument (including the voice signals) is converted to the form of binary digits before leaving the set. In this case the local loop and serving central office or private branch exchange must also be digital. These offices are part of the existing Public Switch Network. Since an all-digital set cannot communicate directly with standard analog instruments, equipment is needed at some central point in the transmission system to convert the digital bit stream to an analog signal, and vice versa. *See* PULSE MODULATION.

The wide availability of high-speed connections to the Internet has made possible the introduction of a new type of telephone design, network connections, and usage patterns referred to as voice over the Internet Protocol (VoIP). VoIP telephones may retain some analog circuits near the transmitter and receiver interfaces, but the interface from the set to the outside world is digital and does not involve a tip or ring. VoIP calls are not carried on the Public Switched Network. Special interfaces are needed to complete a call from a telephone using VoIP technology and one connected to the Public Switched Network. These interfaces are provided by the company selling VoIP service. Sometimes VoIP service is provided by using a standard telephone as described in this article and a converter between the telephone and a high-speed Internet connection. *See* VOICE OVER IP. Richard M. Rickert

Bibliography. J. C. Bellamy, *Digital Telephony*, 3d ed., 2000; J. Carr, S. Winder, and S. Bigelow, *Understanding Telephone Electronics*, 4th ed., 2001;

T. Dean, *Guide to Telecommunications Technology*, 2002; D. G. Fink and D. Christiansen, *Electronics Engineers' Handbook*, 4th ed., 1997; Hill Associates Inc., *Telecommunications: A Beginner's Guide*, 2001; H. Newton, *Newton's Telecom Dictionary*, 22d ed., 2006; A. M. Noll, *Introduction to Telephones and Telephone Systems*, 3d ed., 1999.

Telephone service

The technology of providing many types of communications services via networks that transmit voice, data, image facsimile, and video by using both analog and digital encoding formats.

Component sectors and core technologies. Telephone services involve three distinct sectors of components: (1) customer premises equipment (CPE), such as telephones, fax machines, personal and mainframe computers, and systems private branch exchanges (PBXs); (2) transmission systems, such as copper wires, coaxial cables, fiber-optic cables, satellites, point-to-point microwave routes, and wireless radio links, plus their associated components; and (3) switching systems that often can access associated databases, which can add new intelligent controls for the network's users. *See* COAXIAL CABLE; COMMUNICATIONS CABLE; COMMUNICATIONS SATELLITE; COMPUTER; DIGITAL COMPUTER; MICROWAVE; OPTICAL COMMUNICATIONS; OPTICAL FIBERS; RADIO; SWITCHING SYSTEMS (COMMUNICATIONS); TELEPHONE.

Advances in these components are driven by three core technologies: microelectronics, especially microprocessors; photonics, particularly fiber optics and lasers; and software, which is especially important as a means of introducing, operating, and maintaining telephone services. *See* LASER; MICROPROCESSOR; SOFTWARE.

Controlled by systems integration techniques, the interaction of these three component sectors and these three core technologies has expanded both the capacities and capabilities of the domestic and international communications systems. At the same time, the synergy of the core technologies has reduced significantly the costs of installation and operation of telecommunications systems. This means more sophisticated telephone services can be made more affordable to a broader range of customers. *See* SYSTEMS INTEGRATION.

Private networks. Most telephone service customers share the network facilities with other customers (also known as subscribers). However, some customers, especially large businesses and government agencies, lease private lines that are available only to them. Numerous private networks have been formed within the facilities of the public switched networks. A growing number of these systems employ transmission and switching concepts that create temporary virtual networks as needed, using whatever network facilities are available at the moment. Some of these network applications are discussed below.

A commonly used private network is based on a small on-site switchboard called the private branch exchange (PBX), which serves facilities ranging from a small law office to a multibuilding manufacturing corporation. Although earlier private branch exchange switchboards, such as in hospitals and hotels, required one or more operators, modern versions of private branch exchanges are completely automatic. Calls can be made among telephone stations in the network without the circuit going outside the facility. Some phone companies offer a substitute service, called centrex, in which a section of the central office switch serves the customer like a private branch exchange. *See* PRIVATE BRANCH EXCHANGE.

One widely used private branch exchange feature is voice mail. In essence, it is a centralized answering machine in which a prerecorded message asks the caller to leave a spoken message. The person assigned to that number can listen to the voice-mail messages either upon return to the office or hotel room or by calling from a remote location. Each private branch exchange phone number has an assigned "mailbox" in which message recordings are stored.

Voice and video. Communication, by definition, involves the exchange of information. In telephone systems, this is done by establishing and maintaining a connection between two or more communications terminal devices such as telephones, fax machines, computers or computer modems, video transmitters and receivers, teletypewriters or telex machines, scanners, and voice-mail systems.

A connection between subscribers may be local, regional, transcontinental, or international, and may last for only a few seconds or for years, depending on the needs of the subscribers. A simultaneous two-way connection is known as a duplex channel or circuit, the capacity of which is measured in the range of frequencies, or hertz (Hz), needed to transmit the information. This capacity is called the circuit's bandwidth.

For a simple analog voice-grade phone line, the bandwidth is specified as 4000 Hz, of which about 300–400 Hz are used for supervisory signaling (dial tone, line-busy tone, network-busy tone, and so forth). The distinctive narrow-range sound spectrum of a telephone voice is the result of transmitting only about one-fourth of the actual frequency range used in speaking. Providing the full range of sound would require far more transmission capacity or retrofitting existing networks with more complex components, increasing the cost of service.

In digital systems, the analog speech sounds are encoded into digital signals by a sampling process, using a method known as pulse-code modulation (PCM). This is based on two concepts: first, a continuously varying signal wave, formed by a rapidly changing series of different frequencies (that is, the sounds of speech), can be sliced into a series of samples that can be reconstructed to reproduce the signal wave; and second, the samples can be adequately approximated by discrete numbers. *See* DIGITAL-TO-ANALOG CONVERTER; PULSE MODULATION.

To achieve a natural speech sound (within the voice-channel spectrum), the analog 4000-Hz signal must be sampled digitally at a relatively high rate (8000 times per second, using eight bits to describe the amplitude of the waveform at each sample), resulting in transmission at a nominal rate 64,000 bits per second (64 kb/s). However, advances in coding techniques have lowered the number of bits needed to maintain the quality of telephone voice transmission. Of course, the digitally encoded speech signals must be decoded into analog sounds before a human can hear them. *See* DATA COMPRESSION.

Use of a telephone channel for ordinary analog voice conversations, often referred to as POTS (plain old telephone service), continues to be the dominant mode of service in both domestic and international networks. Although most telephone communication is based on speech (or, increasingly, data exchanges between desktop or portable computer terminals), limited-motion video transmission is now possible on ordinary analog phone lines. There are even videophones designed for encrypted transmission between compatible units to prevent unauthorized interceptions. In addition, digital videophones, employing integrated services digital network (ISDN) technology, serve a variety of businesses. ISDN systems have been particularly successful in Europe, where they have been used to "leapfrog" aging analog systems. Full-motion video currently requires more bandwidth than is available on an ordinary analog phone line. *See* INTEGRATED SERVICES DIGITAL NETWORK (ISDN); VIDEOTELEPHONY.

Significant numbers of telephone customers have special physical needs. They are physically challenged by deafness, blindness, or speech impairment. Special devices are designed to overcome such problems. These include the TTY (teletypewriter) or TDD (telecommunications device for the deaf) systems that display text on small screens, as well as volume controls on receivers, large-button or Braille-button keypads on telephone sets, and electronic larynxes.

Infrastructure. Made over a web of circuits known as a network, a connection often involves several different telephone companies or carriers. In the United States, there are more than 1300 local exchange companies (LECs) providing switched local service, plus more than 500 large and small interexchange companies (IECs), that provide switched long-distance services.

In addition, wireless mobile telephone services are provided on a city-by-city basis by cellular systems operated by two different companies in each metropolitan area. Each cellular system is connected to the wire networks of the local exchange companies and interexchange companies. *See* MOBILE COMMUNICATIONS.

Local exchange company operations of wired systems are divided geographically into 164 local access and transport areas (LATAs), each containing a number of cities and towns. Within a LATA, the telephone company operates many local switches, installed in facilities known as central offices or exchanges.

Transmission. Each central office or exchange switch in connected to local telephone customer premises by a system of twisted-pair wires, coaxial cables, and fiber-optic lines called the loop plant. Direct current (dc) electricity that carries signals through the wires (or powers the lasers and photodetectors in the glass-fiber lines) is provided by a large 48-V stationary battery in the central office which is constantly recharged to maintain its power output. If a utility power outage occurs, the battery keeps the transmission system and the office operating for a number of hours, depending on the battery's size.

The transmission is usually analog to and from the customer site, especially for residences. However, in many systems a collection point between the customer and the switch, known as a subscriber loop carrier (SLC), provides a conversion interface to and from one or more digital cable facilities called T1 carriers (each with 24 channels) leading to the switch. In turn, most switches are interconnected by digital fiber-optic, coaxial, or copper-wire cables or by microwave transmission systems either directly or via intermediate facilities called tandem switches between local switches (**Fig. 1**) or between a group of central offices and a long-distance toll switch. *See* TELEPHONE SYSTEMS CONSTRUCTION.

Domestic dialing. This matrix of local, regional, and national networks is said to be transparent to telephone users, since the connections between different networks usually are made automatically by the caller simply dialing the number. For domestic calls within the United States and Canada, the telephone number for a local call consists of seven digits. The first three digits identify the exchange or central-office switch serving the customer's

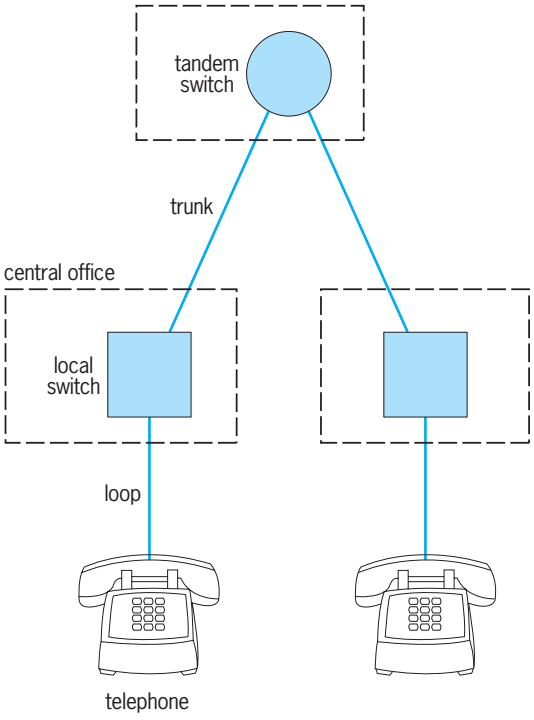


Fig. 1. Elements of the local telephone network.

premises, and the next four digits identify the line number.

If a call is placed beyond the exchange or switch that serves a given phone line, the dialed number often must be proceded by 1 for intra-LATA toll calls (involving switches outside the local area but still inside the LATA), or a 1 plus a three-digit area code and the seven-digit number for a long distance call between LATAs. By the mid-1990s, the reservoir of 152 area codes was depleted; the restriction on the second digit (formerly limited to either 0 or 1) was removed, expanding the feasible selection to 792 area codes.

Although most long-distance calls are dialed directly, using an automatic switching technology called direct distance dialing (DDD), some may require operator assistance (for example, for collect calls charged to the number being called, or for calls to be charged to a telephone credit card). In these cases, the prefix 0 is dialed instead of 1, followed by the area code and number being called. An operator answers the call and processes it from that point on, or a recording instructs the caller to enter the card number and a personal identification number.

International dialing. For international direct-distance dialing (IDDD), the caller usually must dial about 15 digits, starting with 011 to reach an overseas circuit, followed by a country code as well as an area code and telephone number. The dialed numbers are processed automatically through an international gateway switching machine.

In 1963, an international commission developed a numbering plan for all member countries of the United Nations. Each nation was assigned a country code (of one, two, or three digits), followed by the national number of the subscriber's access line.

Switching and transmission designs. Telephone switching machines used analog technology from 1889 to 1974, when the first all-digital switch was introduced, starting with long-distance or toll service and expanded later to central offices or local exchanges. Digital switching machines are comparable to digital computers in their components and functions. Unlike the analog electromechanical switches of the past, digital switches have no moving parts and can operate at much faster speeds. In addition, digital switches can be modified easily by use of operations systems—special software programs loaded into the switch's computer memory to provide new services or perform operational tasks such as billing, collecting and formatting traffic data from switches, monitoring the status of transmission and switching facilities, testing trunk lines between end offices, and identifying loop troubles.

Digital transmission. Transmission systems also evolved from analog technology to digital technology. The first digital transmission system preceded the first digital switch by more than a decade, but use of such systems was limited to trunks between switches or other relatively short routes. The introduction of high-capacity fiber-optic transmission systems in the 1980s made the present-day digital network feasible. The first long-distance fiber (or light-wave) system linked Boston, New York City, Philadelphia, and Washington, D.C., in 1981. The first transatlantic fiber-optic submarine cable, which also introduced undersea cable branching for multiple landing sites, was activated in 1988, followed in 1989 by the first transpacific fiber-optic cable system. *See* SUBMARINE CABLE.

Although fiber-optic cables are much thinner than copper cables, they have far more capacity in terms of channels. The last copper transatlantic cable (installed in 1983 and known as TAT-7) was rated for 4000 simultaneous analog conversations, while the first transatlantic fiber cable (TAT-8) was rated for 40,000 simultaneous digital conversations. Subsequent fiber cables were rated for 80,000 digital conversations (TAT-9 and TAT-10).

The submarine light-wave cables provided the first true high-speed digital telephone circuits across the Atlantic and Pacific oceans, but they also encountered a disparity between the digital protocols of North American (where pulse-code modulation encoding is done with a mu-law algorithm) and Europe, Africa, South America, and most Asian countries (where such coding uses an incompatible A-law algorithm). By international agreement, the mu-law countries such as the United States are responsible for providing gateway transcoding facilities. The digital signals transmitted via the submarine cables are coded in A-law pulse-code-modulation signals.

Digital hierarchies. In North America, digital switching and transmission are conducted within a hierarchy of multiplexing levels. The single digital telephone line, rated at 64 kilobits per second (kb/s), is known as a digital signal 0 (DS-0) level. The lowest digital network transmission level is DS-1, equivalent to 24 voice channels multiplexed by time-division multiplexing (TDM) to operate at 1.544 megabits per second (Mb/s). DS-1 is the most common digital service to customer premises. Two interim levels, now seldom used, are followed by DS-3, perhaps the most widely used high-speed level. DS-3 operates at 44.736 Mb/s, often rounded out to 45 Mb/s, the highest digital signal rate conventionally provided to customer premises by the telephone network. Of course, groups of DS-3 trunks can be connected to a customer facility if needed. Outside the United States, other digital multiplexing schemes are used. This results, for example, in a DS-1 level having 30 voice channels rather than 24 channels. *See* MULTIPLEXING AND MULTIPLE ACCESS.

A second network hierarchy appeared during the early 1990s and was gradually implemented in the world's industrial nations. Known in North America as SONET (synchronized optical network) and in the rest of the world as SDH (synchronized digital hierarchy), these standards move voice, data, and video information over a fiber network at any of eight digital transmission rates. These range from OC-1 at 51.84 Mb/s to OC-48 at 2.488 gigabits per second (Gb/s), and the hierarchy can be extended to more than 13 Gb/s. The OC designation stands for optical carrier.

This concept enhances the power of the public

network by increasing the amount of information that can be carried, using international standards for digital transmission. SONET/SDH also brings network intelligence closer to end users by enlarging the capacity of the overhead channel embedded in the circuit's signal. This channel carries operations, administration, maintenance, and provisioning (OAM&P) information accessible by the user, in addition to the customer's data (called the payload portion of the signal). The built-in capabilities of microprocessors or firmware in network elements constitute the means to make the network self-aware, self-adapting, and self-managing. Problems are detected and corrected automatically at their root, a process sometimes known as a self-healing network.

In the standard digital hierarchy, customer networking applications are supported up to the DS-3 (45-Mb/s) transmission rate. Advanced technologies and services, such as the asynchronous transfer mode (ATM) and the broadband integrated services digital network (BISDN), define rates as high as 155 Mb/s and 600 Mb/s, equivalent to SONET's OC-3 and OC-12 levels. Such high-speed rates help to improve the performance of certain applications such as multimedia and distributed data processing that can strain the resources of conventional digital systems operating at lower bit rates. *See* MULTIMEDIA TECHNOLOGY.

Asynchronous transfer mode. This technology is based on a technique called cell-oriented switching and multiplexing. This is a packet-switching concept, but the packets are much shorter (always 53 bytes) and faster (for better response times) than are the packets in such applications as the global Internet and associated service networks, based on the X.25 standard, which can have up to 4096 bytes of data in one packet. The asynchronous transfer mode cell-relay system moves cells through the network at speeds measured in megabits per second instead of kilobits. With its tremendous speed and capacity, asynchronous transfer mode technology permits simultaneous switching of data, video, voice, and image signals over cell-relay networks. *See* DATA COMMUNICATIONS; PACKET SWITCHING.

Standards. Uniform technical standards have been adopted worldwide. For example, facsimile or fax machines using a global standard known as Group 3 can communicate with each other. *See* FACSIMILE.

Within the United States, technical standards for telephone services are promulgated by the American National Standards Institute (ANSI). The Federal Communications Commission (FCC) regulates all interstate and foreign radio and wire communications services originating in the United States.

Globally, the United Nations oversees the International Telecommunications Union (ITU). In 1992, the ITU was reorganized into three groups that develop and recommend standards: the Radio Communications Sector (formerly the Consultative Committee for International Radio or CCIR), the Telecommunications Standardization Sector (formerly the Consultative Committee for International Telephony and Telegraphy, or CCITT), and the Telecommunications Development Sector (BDT). Their standards are only

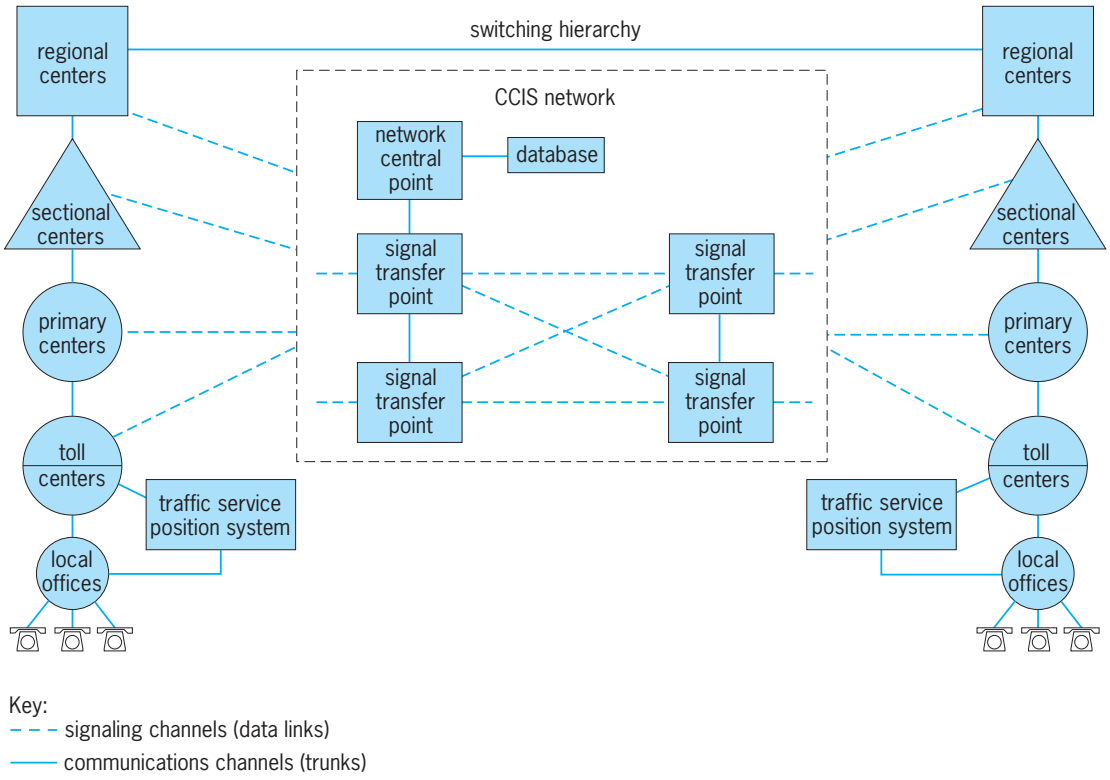


Fig. 2. National long-distance telephone network of the early 1980s, which introduced common-channel interoffice signaling (CCIS) to speed call completions, using 14 special packet switches called signal transfer points. Regional, sectional, primary, and toll centers are part of toll switching system. The traffic service position system is an operator services system. (*AT&T*)

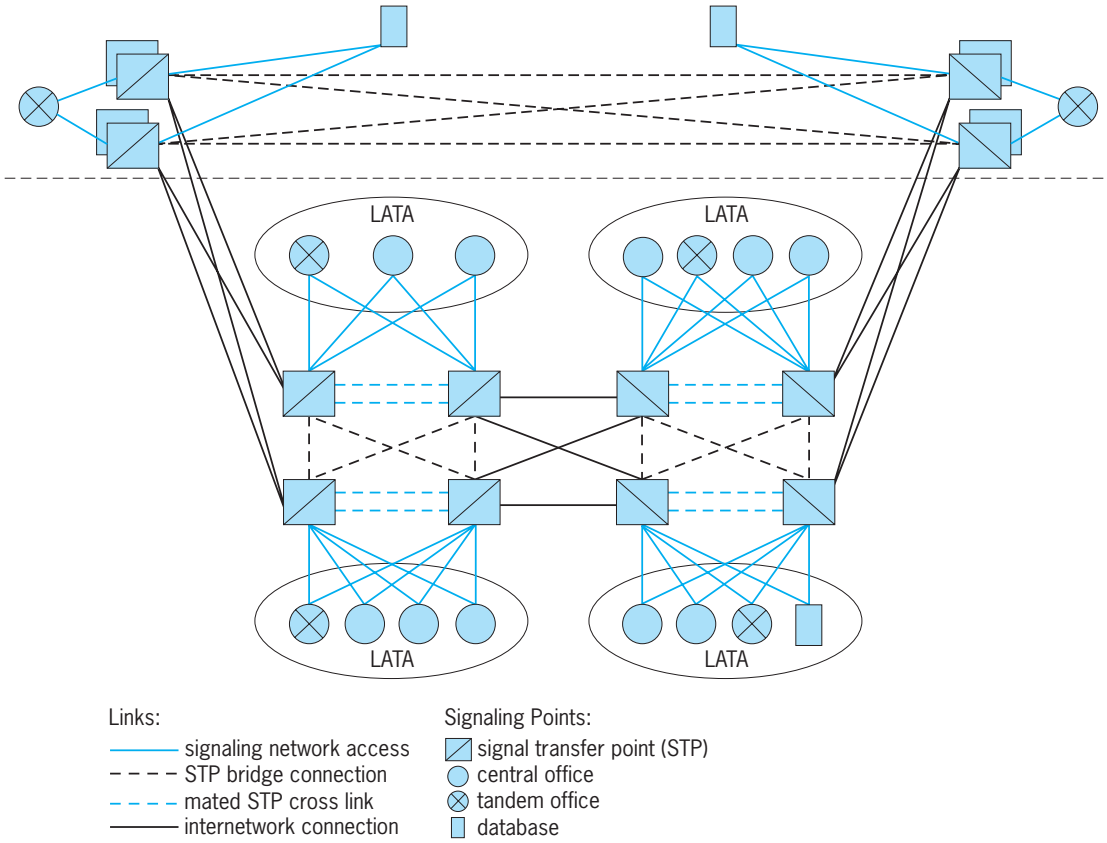


Fig. 3. National long-distance telephone network of the early 1990s, which uses common-channel Signaling System 7 technology. Signal transfer points serve local switches as well as the long-distance network with out-of-band access to various data banks, and intelligent networks can be set up. (*AT&T*)

recommendations, but most nations accept them verbatim, while others, including the United States, make minor modifications.

Out-of-band signaling systems. When the voice or communications information is encoded in a stream of bits, it becomes possible to use designated bits instead of analog tones for the supervisory signaling system. This simplifies the local switching equipment and allows the introduction of sophisticated data into the signaling, which significantly expands the potential range of telephone services.

Operations information now flows between switches as packet-switched signaling data via a connection that is independent of the channel being used for voice or data communications (a technique known as out-of-band signaling). Typically, a channel between two switching systems is known as a trunk. A trunk group between switches carries a number of channels whose combined signaling data are transmitted over a separate common channel as a signaling link operated at 56 kb/s. This technique is called common-channel interoffice signaling (CCIS).

SS6. In 1976, the first United States version of CCIS was introduced as a modification of an international standard, the CCITT No. 6 signaling system. The out-of-band signaling system was deployed in the long-distance (toll) network only, linking digital as well as analog switches via a network of packet-data switches called signal transfer points (STPs; **Fig. 2**). Also known as Signaling System 6 (SS6), it helped to make possible numerous customer-controlled services, such as conferencing, call storage, and call forwarding.

One key advantage of common-channel signaling technology was more efficient use of the national network; it typically reduced the routing time for a direct-dialed coast-to-coast call from around 10 s to only a second or two. The technology also enables the virtual elimination of a major means of fraudulent use of the network, since the common-channel signaling system is inaccessible to customers.

SS7. A major advance in common-channel signaling was introduced with the CCITT's Signaling System 7 (SS7), which operates at up to 64 kb/s and carries more than 10 times as much information as SS6. In addition, Signaling System 7 is used in local exchange service areas as well as in domestic and international toll networks (**Fig. 3**). The first trials took place in 1989; deployment in the United States long-distance networks soon followed, with local exchange companies deploying it at a somewhat slower rate.

In addition to increasing network call setup efficiency, Signaling System 7 enables and enhances services such as ISDN applications; automatic call distributor (ACD); and local-area signaling services (LASS), known as caller identification (**Fig. 4**). Other innovations include 800 or free-phone service, in

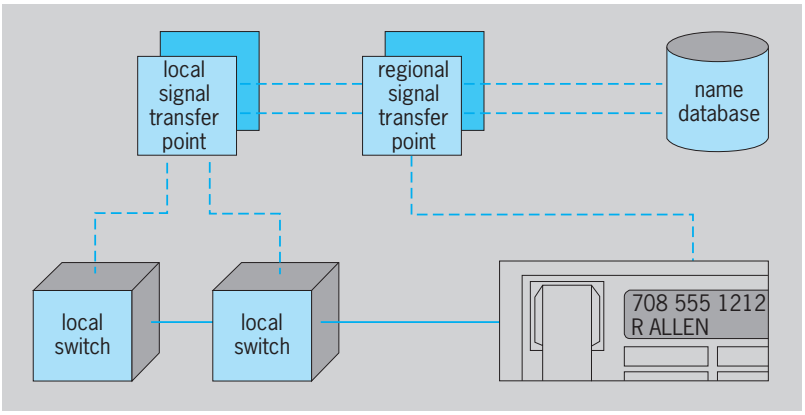


Fig. 4. Caller identification service in Signaling System 7, whereby call waiting identifies the caller's name and directory number for waiting calls. Broken lines indicate SS7 links. (*AT&T*)

which customers can place free long-distance calls to businesses and government offices; 900 services, in which customers pay for services from businesses; and enhanced 911 calls, in which customer names and addresses are automatically displayed in police, fire, or ambulance centers when calls are placed for emergency services.

The Signaling System 7 software also can include numerous other custom calling services for residential and business customers. Call waiting sends a special audio signal to a telephone that is being used, alerting the called party that another caller is waiting; the system allows the called party to place the first call on hold and answer the second call without hanging up on the first call. Selective call waiting gives the user a distinctive tone when a call is received from a number on a specified list of callers.

Other custom calling features include call forwarding, which transfers incoming calls to another number specified by the customer for a temporary period; three-way calling, which lets the customer set up conference calls without operator assistance; and speed calling, which enables the customer to program 8 or 30 frequently called numbers into the memory of the local exchange switch, so the customer need only press two or three buttons to reach these numbers.

Intelligent networks. The greatest potential for Signaling System 7 use is in emerging intelligent networks, both domestic and international. The intent is to provide customers with much greater control over a variety of network functions, yet to protect the network against misuse or disruption. Intelligent networks are evolving from the expanding use of digital switching and transmission, starting with the toll networks. The complexity of intelligent networks mandates the use of networked computers, programmed with advanced software.

An intelligent network allows the customer to setup and use a virtual private network as needed, and be charged only for the time that network is being used. A conventional private leased network, by contrast, reserves dedicated circuits on a full-time basis and charges for them whether they are used or not. A virtual private network enables a business to simultaneously reap the benefits of a dedicated network and the shared public-switched network by drawing on the infrastructure of intelligent networking (**Fig. 5**).

The introduction in 1964 of customer dialing for international calls is at the root of the intelligent network concept. The international direct distance dialing service has been expanded and made more flexible by implementing the SS7 together with new databases (such as 800 or free-phone services, emergency hot lines, capture of calling-party names and numbers, screening of incoming calls, automatic call-up and on-screen display of client order-billing records, computer-based voice messaging store-and-forward systems, and automated calling-card services. The databases are fundamentally software programs that interact with voice or data signals, either for business applications or for network operations and maintenance. *See* DATABASE MANAGEMENT SYSTEM.

For example, a retail chain can link its stores in the United States to factories operated by its vendors in Asia and Central and South America. Constant communication between computers at the sales points and the factory sources can provide savings in inventory space, better control of labor costs, and a quicker response to market conditions. Stores can be added to or dropped from the system as required, and factories can be canceled or appended as needed, by using the virtual network concept.

Domestic intelligent networks employ computers that function as signal transfer points. These signal transfer points route intelligent network messages (such as identifying 800 number destinations and

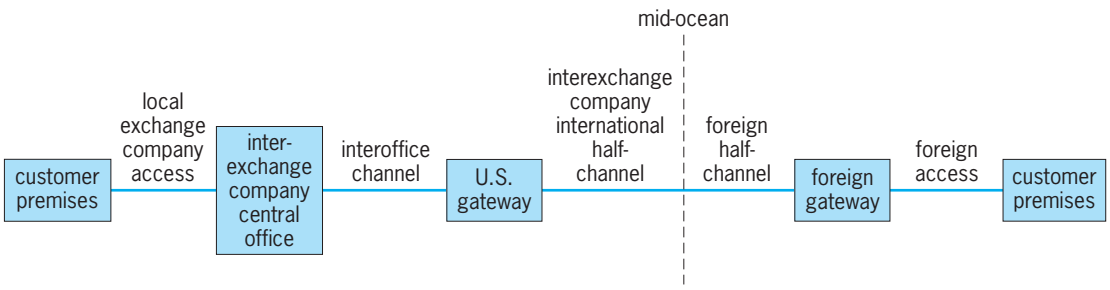


Fig. 5. Virtual private network with international links that must involve service providers of both the United States and other countries. (*AT&T*)

call screening) between switches and advanced network databases. They operate in pairs, with each signal transfer point in a pair capable of picking up the entire load if failure of the other occurs. They communicate constantly, checking on each other's health as well as finding clear paths for calls and providing their status to central operations-maintenance centers. Most major carriers, both local and long distance, use such network operations centers (NOCs) to monitor the health of their networks.

Government and business networks frequently use global services to control costs and to supplement the functionality of their private networks. Data calls, for example, can be funneled through packet-data switches that communicate with packet switches in other countries, so a data message can be transmitted through the telephone network over a variety of available channels at lower cost than if the data were sent over a single leased channel. This enables global credit-card validation, for example, on a real-time basis, with the card company computer located on one continent and the sales point on another.

Private networks are sometimes called overlay networks, because they are superimposed on the public switched telephone networks and share the same facilities. These networks can be enormous in scale, whether domestic or global.

Government networks. The United States government also employs private networks. The original Federal Telecommunications System (FTS) was activated in 1964, and in 1989 was replaced by a new digital voice-data network for government facilities. The FTS 2000 system supplies six communications services: switched voice service for transmission of voice and analog data at up to 4.8 kb/s; switched data service for circuit-switched digital data communications at 56 kb/s and 64 kb/s; switched digital integrated service for digital transmission of voice, data, image, and video signals at transmission rates of up to 1.544 Mb/s (also furnishing the integrated access lines and user-network interfaces needed for ISDN applications); packet-switched service for transmission of data in packet format, including electronic mail service; video transmission service for teleconferencing in both compressed (near-full-motion) and wideband (full-motion) video modes; and dedicated transmission service for point-to-point, dedicated voice and data lines at rates up to 45 Mb/s.

The Department of Defense sponsors the Defense Commercial Telecommunications Network (DCTN), a digital system designed to serve users in the Department of Defense and other federal agencies at about 150 locations. Another Department of Defense system, the Automated Voice Network (AUTOVON), was activated in April 1964, becoming the first worldwide switched network for private telephone and data transmission. AUTOVON is an analog network, divided into the continental United States portion and the overseas portion.

Certain federal government agencies require a high degree of protection against eavesdropping on telephone communications. For this purpose, a special encrypted telephone system called the STU III (Secure Telephone Unit Model III) was developed during the 1980s. It can be used as an ordinary voice telephone, or it can be activated for sending and receiving encrypted voice and data transmission (to and from another such terminal) by turning a special key in the side of the unit. Modified versions of these telephones are available for civilian use. *See* CRYPTOGRAPHY.

Internet. One federal government experiment that began in 1969 became unexpectedly popular in the scientific community and has proliferated into a global system of interconnected networks, now known as the Internet. The first component was the ARPANET, developed by the Defense Advanced Research Projects Agency (DARPA) to provide data communications links between major existing computer facilities and users in academic, industrial, and government research laboratories. ARPANET operates with 56-kb/s digital links using packet switching, and it served as a test for the development of advanced network protocols such as the TCP-IP (Transmission Control Protocol–Internet Protocol). During the 1980s, as desktop personal computers and workstations gained in use, ARPANET was expanded to include members of the NSFnet (sponsored by the National Science Foundation) as well as other data networks. Eventually, several thousand smaller data networks were linked into a global, but informal, Internet.

Telephone circuits form the transmission media for users of the Internet. The circuits are leased from various carriers by the network sponsors, who link into the public network through gateway packet switches. Most of the residential or small-business end users are equipped with desktop or portable computers that employ a modem (modulator-demodulator) to convert the computer's digital signals to analog tones acceptable for transmission over the analog loop lines, and to convert incoming analog tones to digital data accepted by the computer. Most messages are in the form of e-mail (electronic text), but graphics and still images also can be transmitted. Services provided over the Internet are made more accessible through browsers that use a protocol called hypertext linking that allows users to scan the World Wide Web. *See* ELECTRONIC MAIL; INTERNET; WORLD WIDE WEB.

Wireless. Another telephone technology that has gained popularity at unexpected speed is the wireless services.

Cordless telephones. In residences and offices, wireless transmission is handled by cordless phones, which are actually small radiotelephone receiver-transmitter systems. A base unit is plugged into the nearest telephone wall jack for linking to the loop facilities, and also is plugged into an electrical outlet for the necessary alternating current used to power its internal transmitter-receiver as well as to recharge the battery in the portable handset. The range of a cordless telephone is relatively limited and is usually less than 0.5 mi (0.8 km).

Originally, cordless phones were analog, and many were not protected against signal interception, so unauthorized listeners might hear one side of the conversation. More advanced models are based on digital technology and are far more secure during transmission.

Cellular service. A major innovation in wireless technology resulted from work on improving the service to mobile telephones in automobiles (advanced mobile phone service, or AMPS), which was introduced in 1983. This service was called cellular telecommunications because it uses a system of low-power radiotelephone transceivers serving a series of interlinked cells throughout a city or along a heavily traveled highway. As the car carrying the phone moves from one cell to another, the cellular system senses the change in location and transfers, or hands off, the call from one antenna site to another.

Although originally envisioned as an improved service for cars, the cellular service has grown dramatically, propelled more by the use of small, hand-held cellular telephones than by vehicular units. In addition to the convenience offered to telephone users, hand-held cellular phones represent a major advantage for emerging countries that lack the wired infrastructure that was previously required to establish communications.

The rapid growth of cellular-service subscribers has led to congestion in some cellular systems. The solution is to introduce digital cellular systems, which can hande up to 10 times more calls in the same frequency range. However, in countries such as the United States, which already have an analog infrastructure, the FCC has mandated that any digital system must be compatible with the existing analog system. The initial digital designs, based on a technology called time-division multiple access (TDMA), provide a three-times growth factor. Another digital technology, based on code-division multiple access (CDMA), a spread-spectrum technique, could increase the growth factor to 10 or even 20. *See* RADIO SPECTRUM ALLOCATION; SPREAD SPECTRUM COMMUNICATION.

Personal communications service (PCS). Another evolving concept, this is a sort of cellular system, in which a pocket-size telephone is carried by the user. A series of small transmitter-receiver antennas operating at lower power than cellular antennas are installed throughout a city or community (mounted on lampposts and building walls, for example). All the antennas of the personal communications network (PCN) are linked to a master telephone switch that is connected to the main telephone network.

Services to ships and aircraft. Perhaps the oldest form of mobile radiotelephone service is the long-range high-frequency (HF) service centered in Florida, New Jersey, and California to furnish ship-to-shore links. Initiated in 1929, this service uses various radio channels in the bands between 2 and 23 MHz, providing communications links between commercial and private vessels as well as aircraft. The route and progress of a ship or plane must be known in advance, so that the radio frequencies and antennas most suitable to its location can be used. Modern systems employ geostationary satellites operated by the International Maritime Satellite (INMARSAT) organization. That system has eight satellites covering the Atlantic, Pacific, and Indian oceans and is responsible for mobile communications on land, sea, and air.

Services to remote areas. Not all satellite communications links serve mobile telephones. In remote areas, digital packet-switching transceivers are now used to serve as many as 120 telephones each. The conversations are relayed via geostationary satellite to and from metropolitan switching centers linked by cable to the global telephone networks. Similar systems were used in the Persian Gulf War of 1990–1991 to link troops in the Saudi Arabian desert with their families in the United States, and to reestablish the Kuwait telephone links with the rest of the world until the regular facilities could be reconstructed. *See* ELECTRICAL COMMUNICATIONS. John H. Davis

Bibliography. J. Bellamy, *Digital Telephony*, 3d ed., 2000; R. J. Chapuis and A. E. Joel, Jr., *Electronics, Computers and Telephone Switching*, 1990; B. B. Lee, M. Kang, and J. Lee, *Broadband Telecommunications Technology*, 2d ed., 1996; D. E. McDysan and D. L. Spohn, *ATM: Theory and Application*, 1995; A. M. Noll, *Introduction to Telephones and Telephone Systems*, 3d ed., 1999; M. Sexton and A. Reid, *Transmission Networking: SONET and the Synchronous Digital Hierarchy*, 1992; G. A. Silver and M. L. Silver, *Data Communications for Business*, 3d ed., 1994; T. H. Wu, *Fiber Network Service Survivability*, 1992.

Telephone systems construction

Construction of the physical components of the telephone system from the central office to the end user. The telephone system can be roughly divided into three major components: outside plant, switching equipment, and interoffice facilities. The outside plant can be thought of as the local cables that bring the phone service to the subscribers' premises. The switching equipment resides in a building called a central office and serves to route the call to the correct location. The interoffice facilities connect one central office to another. When a call is initiated, the outside plant carries the signal to the local central office. The switching equipment in the local central office then determines the location to which the call needs to be routed. The call is then routed over the interoffice facilities to the central office serving the person being called, where the switching equipment then connects the call to the local outside plant. This article describes the outside plant portion of the telephone system.

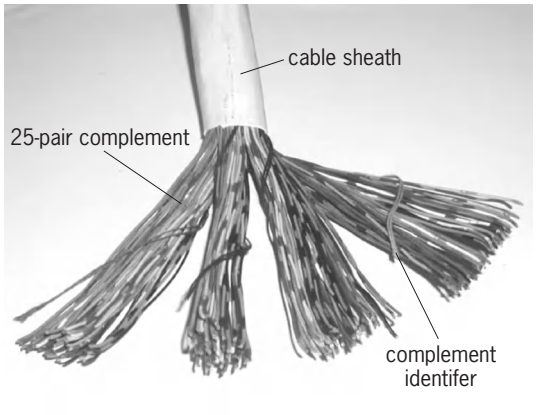
Outside plant consists of cables and their supporting facilities. This cable can be aerial (supported by poles), underground in conduit, or buried directly in the ground. There are two main types of cable, metallic and optical.

Metallic cable. Metallic cable consists of usually copper, but at times aluminum, conductors twisted

together in pairs. These conductors are commonly 22, 24, or 26 AWG (American Wire Gauge). A POTS (plain old telephone service) circuit uses a single pair of conductors. Additional services may require more than one pair of wires.

Initial cables consisted of copper conductors, insulated with wood pulp. These conductors were twisted together in pairs, and then the pairs were twisted together in groups. The entire bundle was encased in a lead sheath to create a telephone cable. This type of cable was referred to as lead or pulp cable. The number of individual pairs in a cable depended on the application for which the cable was designed, but could range up to 4200 pairs. In order to facilitate identification of individual pairs, the pulp insulation was color-coded. These cables presented a number of problems: Lead was found to be a hazardous material; the lead sheath was prone to cracking when flexed, especially in cold weather; the pulp insulation wicked moisture easily; and the color coding on the conductors would fade with time

Modern telecommunications cable consists of copper conductors coated with a plastic color-coded insulation. There are a number of different plastic insulations available, depending on the application, manufacturer, and local practice. The pairs are then twisted together and grouped into 25-pair complements, which are further twisted together to make a cable. The cable is then covered in a protective sheath, also called a turnplate, made of aluminum or steel (see **illustration**). The turnplate provides mechanical support to the cable and also provides electromagnetic shielding to the conductors in the cable. Sometimes additional armoring materials are added to the sheath, typically for resistance to abrasion or to prevent damage from burrowing animals in buried applications. The turnplate is then covered in a sheath made of polyethylene. Cable of this type is referred to as PIC (pronounced "pick") for plastic insulated cable. The center of the cable can be left as is, and is referred to as air core; or it can be flooded with a moisture-repelling compound, and is referred to as jelly-filled cable. Because of the sticky nature of this compound, people who work with this type of cable often call it "icky-PIC."



A 100-pair plastic insulated cable. (*From J. Clayton, McGraw-Hill Illustrated Telecom Dictionary, 4th ed., McGraw-Hill, 2002*)

Fiber cable. The second major type of cable deployed in the outside plant is optical fiber. Optical fiber consists of a glass core, coated with a glass cladding, and then coated with plastic. Light is transmitted along the core of the fiber. The interface between the core and the cladding causes the light to reflect back into the core, and in effect "bounce" back and forth as it travels along the core. Two major types of fiber are used in telephony, single-mode and multimode. A single-mode fiber has a core that ranges 8.3–10 micrometers in diameter. A multimode fiber has a core that is 50–100 μm in diameter. In a single-mode fiber, there is only enough "room" for the light to take one path along the core. In a multimode fiber, the light can reflect in a number of ways and take multiple paths along the fiber. Single-mode fiber provides higher transmission speeds over longer distances; however, it is more expensive and requires a more precise light source. *See* OPTICAL COMMUNICATIONS; OPTICAL FIBERS.

Cables are made up of multiple fibers. In a loose tube configuration, individual fibers are placed in plastic tubes. These tubes are larger then the fibers contained within, and may sometimes be flooded with a water-repelling compound. This oversized tube serves to prevent stresses due to handling the cable from being transferred to the delicate fibers. The tubes are gathered together around a central strength member, either metal or plastic. The whole bundle is then wrapped with an aramid yarn for strength and covered with a plastic sheath. In buffered cables, the individual fibers are covered with a protective plastic coating, and directly covered in the supporting yarns and sheath. Fibers can also be ganged together in a flat ribbon consisting of typically 6, 12, or 24 fibers held together by a plastic ribbon material. These fibers are then "stacked" and placed in a single tube, around which the cable is constructed. Each method has different applications. The cables vary in degrees of cost, toughness, ability to access individual fibers, and ease of splicing. As with metallic cable, fiber cable is also available with an armored sheath. *See* COMMUNICATIONS CABLE.

Aerial plant. Aerial plant consists of telephone cable that is suspended from structures, typically telephone poles. The majority of telephone poles are made of wood. Typical woods used are southern pine, western red cedar, and Douglas-fir. Other woods are used as well, including jack pine, red pine, and eastern cedar. The choice of wood depends mainly on the location of the manufacturer's wood lots. Utility poles are usually treated with chemicals to increase their rot resistance; the treatment method used depends on the manufacturer and local practice. While the majority of poles are wood, alternative products are available, including fiberglass, concrete, metal, and laminated wood poles. All of these materials have certain benefits and drawbacks, and are usually used only in specific situations.

A telephone pole is set into a hole in the ground. The depth of the hole depends on the size of the pole and how much weight it is designed to carry, but typically ranges 5–10 ft (1.5–3 m). In the past, these holes were dug with hand tools, and the poles were carried and tipped up into the holes by hand with the assistance of long poles with sharp tips called pike poles. Currently, the majority of poles are placed using a digger derrick machine. This machine consists of a truck with a powered auger that is used to dig the hole, and a powered boom with a winch and clamping jaws that is used to set the pole. However, in some situations such as in a backyard, the digger derrick cannot be maneuvered into position, and the poles must still be carried and set by hand.

The telephone poles are connected by a steel strand that is anchored to each pole. Cable is placed by hanging rollers, called blocks, on the strand. The cable is then pulled through the blocks. There are a variety of methods for performing this step. The difference lies in where the cable reel is placed and how the truck is moved. The decision on what method is used depends on the size of the job, the physical layout of the area, and personal preference.

Once the cable is suspended in the blocks, a machine called a cable lasher is placed over the cable and strand. Pulled along the cable, it wraps a wire around the cable and strand, lashing the cable snugly to the strand, and it also pushes the cable blocks along the strand ahead of it.

Underground plant. Underground plant consists of facilities that are placed in conduit and vaults. Vaults are typically constructed of concrete. They either can be precast and set into an excavation by crane or can be cast in place using forms. Precast vaults are available in a variety of sizes, and cast-in-place ones can be any size desired. However, a typical telecommunications vault is approximately 7 ft (2.5 m) deep, 6 ft (2 m) wide, and 12 ft (4 m) long. Most telecommunication vaults have two lids and many have four, depending on size. The metal cover and frame sit on top of a cement or brick chimney that extends down to the roof of the vault. The height of this chimney can be adjusted to accommodate changes in road height.

Vaults are connected by conduit, which is also known as ducts. At present, conduit is typically 4-in. (10-cm) polyvinyl chloride (PVC) pipe. In the past a variety of materials, including clay tile, wood, cement, and steel, were used as duct. A trench is excavated between vaults, and a number of conduits are placed there before the trench is backfilled. Conduit can also be run from a vault to a building basement, or to the base of a pole where the cable can emerge, run up the pole, and continue as an aerial cable.

When the conduit system is constructed, a pulling line is installed in each duct for the installation of cable. If the pulling line is missing, the duct can be rodded from vault to vault using a series of stiff steel rods and a new rope can be pulled in. Rodding can also be used to remove obstructions that may occur from a crack in a duct.

Once a rope is in place, it is used to pull in a cable. The cable is prepared with a special lubricant, and it is winched through the duct. A variety of rollers and shoes are available to prevent the cable from kinking and binding as it is pulled down into the vault and into the duct. Once in place, the cable is racked onto metal framework along the sides of the vault.

Buried plant. Buried plant consists of cables that are buried directly in the ground. This method is much cheaper than constructing a conduit system, but lacks flexibility. In a conduit system, if a cable needs to be replaced or upgraded, it can be pulled out and a new one pulled in. With direct-buried cable, the entire trench must be reexcavated. Direct-buried cable is also susceptible to damage from rocks. Cable should be bedded in sand or some other suitable fill, but rocks can work their way to the cable and eventually puncture the sheath. Direct-buried cables are usually used in limited situations in garden apartments, planned communities, and complexes. Such cables are usually made specifically for being buried and are armored and jelly-filled.

Splicing. Once cable is placed in the outside plant, it must be connected by splicing. There are a number of methods used to join copper conductors. In the past, conductors were bared, twisted together, and soldered. In today's outside plant, special connectors are used. Some connectors permanently join together a single conductor. If such a connector needs to be removed, it must be cut out and a new connector used. Other connectors are designed to be pluggable. Pluggable connectors make use of the fact that cables are built in 25-pair complements. Pluggable connectors are designed to take all of the conductors in a 25-pair complement and terminate the ends of the conductors into a large plug module. The modules can then be snapped together and apart to reconfigure the cable as needed. All modern connectors are insulation displacement connectors (IDC). The wire does not need to be stripped. The connector is applied with a pliers-like tool and pierces the insulation to make contact with the conductor inside.

Fiber cable is also spliced in a number of ways. Individual fibers can be joined using either mechanical or fusion methods. With mechanical methods, the ends of the fiber are cleaned and cleaved to create a flat clean end, and then placed in a device that holds them precisely together. There are also pluggable connectors that can be applied to the ends of fibers, enabling them to be connected and disconnected. These are often used when connecting fibers to equipment. The fusion method involves use of an electric arc. The ends of the fibers are prepared by cleaning and cleaving, and then placed into special holders. These holders are placed into a fusion splicing machine, which precisely aligns the ends of the fiber and applies a brief high-voltage arc. This arc fuses the ends of the two fibers together. Machines are available that can splice entire ribbons of fiber together at once. Mechanical splicing requires a much lower investment in tools, but costs more per splice and produces an inferior splice that causes more loss in the fiber. Tools for fusion splicing are

quite expensive, but the cost per splice is much less and the quality of the splice is higher.

Once a splice is created, a splice closure needs to be placed to protect it. In the past, splices were encased in muslin and then covered in molten lead in a technique known as wiping. Modern splice closures are plastic cases that snap around the open cable. They are available in a number of sizes and manufacturing methods.

Air pressure. If moisture gets into a copper cable, it will begin to act as a path for electricity to flow between conductors. At first, this will create noise or static on a POTS line or degrade a data circuit. If the cable becomes wet enough, the conductors in the cable will short and the circuit stops working. Telecommunications cables are pressurized with air to help prevent this failure.

An air system consists of dryers, air pipes, and transducers. An air dryer provides a source of pressurized dry air. Most air dryers are located in a central office, but on longer cable runs they can be located in the field as well. The compressed dry air flows from the dryer to a variety of manifolds and pipe panels, which break the single air feed from the dryer into a number of smaller air pipes for each cable. Transducers are mounted on the pipe panel to monitor air pressure and flow.

The air pipes are run next to the cables, and at regular intervals are tapped into the cable via a valve installed through the cable sheath. This valve provides pressurized air to the air core of a cable. Any breach in a cable's sheath that would let water in instead forces air out. Thus, this pressure difference protects the cable against water.

Pressure and flow transducers are also placed at regular intervals along the cable run. These transducers are monitored. An increase in flow or a decrease in pressure serves as an early warning that a cable may be damaged.

Cables placed underground in a conduit system are pressurized, as are some sections of cables in the air. When a cable changes from pressurized to nonpressurized, a plug of epoxy is poured in the cable to stop the airflow. When splice cases are placed on pressurized sections of cables, the cases must also be airtight.

Additional equipment. Besides the basic construction of the outside plant portion of a telephone network, there are many additional pieces of equipment. Various electronics can be placed in the field. They can serve to extend the distance a signal can travel, increase the number of voice circuits that can be carried on a particular copper pair, or provide higher-speed and -bandwidth data circuits. These electronics can be placed in small above-ground buildings called huts; in buried vaults called controlled environment vaults (CEV) that have heating, cooling, lighting, and dehumidification; or in above-ground mounted cabinets. New equipment is always being developed, and as it is deployed old equipment must be removed, due to either space limitations or incompatibility with the new equipment.

David F. Dolch

Bibliography. J. Clayton, *McGraw-Hill Illustrated Telecom Dictionary*, 4th ed., 2002; R. A. Meyers, *Encyclopedia of Telecommunications*, 1989; A. M. Noll, *Introduction to Telephones and Telephone Systems*, 3d ed., 1999.

Telescope

An instrument used to collect, measure, or analyze electromagnetic radiation from distant objects. A telescope overcomes the limitations of the eye by increasing the ability to see faint objects and discern fine details. In addition, when used in conjunction with modern detectors, a telescope can "see" light that is otherwise invisible. The wavelength of the light of interest can have a profound effect on the design of a telescope. *See* ELECTROMAGNETIC RADIATION; LIGHT.

Parameters. The utility of a telescope depends on its ability to collect large quantities of light and to resolve fine details. The brightness of an image is proportional to the area of the light-gathering element, which is proportional to the square of that element's aperture. The brightness also depends on the area over which the image is spread. This area is inversely proportional to the square of the focal length (f) of the lens. The brightness of the image therefore depends on the square of the f/ratio, just as in an ordinary camera. The resolving power of a telescope depends on the diameter of the aperture and the wavelength observed; the larger the diameter, the smaller the detail that can be resolved. An 11-cm-aperture (4.5-in.) telescope will resolve approximately 1-arcsecond details. On Palomar Mountain, California, the 5-m (200-in.) Hale telescope has an aperture that is roughly 1000 times the diameter of the dark-adapted eye. Thus, it gathers 1 million times more light than the eye. Theoretically, it can resolve details as small as 0.02 arcsecond, but because it must look at celestial objects through the Earth's turbulent atmosphere, it cannot resolve details as fine as the theoretical limit.

In a telescope, the angular separation A of two points on the object is magnified to the angle A' as seen by the eye. The magnification (A'/A) is numerically equal to the ratio of the focal length of the main optical element, f_1, to the focal length of the eyepiece, f_2; that is, the magnification is f_1/f_2. This is true whether the main optical element is a lens, a mirror, or a combination of lenses and mirrors. *See* FOCAL LENGTH; LENS (OPTICS).

Limitations. For many applications, the Earth's atmosphere limits the effectiveness of larger telescopes. The most obvious deleterious effect is image scintillation and motion, collectively known as poor seeing. Atmospheric turbulence produces an extremely rapid motion of the image resulting in a smearing. On the very best nights at ideal observing sites, the image of a star will be spread out over a 0.25-arcsecond seeing disk; on an average night, the seeing disk may be between 0.5 and 2.0 arcseconds. It has been demonstrated that most of the air currents

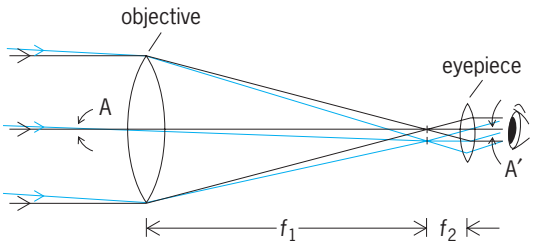


Fig. 1. Simplified optical diagram of a refracting telescope.

that cause poor seeing occur within the observatory buildings themselves. Substantial improvements in seeing have been achieved by modern design of observatory structures.

The upper atmosphere glows faintly because of the constant influx of charged particles from the Sun. This airglow adds a background exposure or fog to photographic plates that depends on the length of the exposure and the speed (*f*/ratio) of the telescope. The combination of the finite size of the seeing disk of stars and the presence of airglow limits the telescope's ability to see faint objects. One solution is placing a large telescope in orbit above the atmosphere. In practice, the effects of air and light pollution outweigh those of airglow at most observatories in the United States. *See* AIRGLOW.

Optical telescopes. There are basically three types of optical systems in use in astronomical telescopes: refracting systems whose main optical elements are lenses which focus light by refraction; reflecting systems, whose main imaging elements are mirrors which focus light by reflection; and catadioptric systems, whose main elements are a combination of a lens and a mirror. The most notable example of the last type is the Schmidt camera. In each case, the main optical element, or objective, collects the light from a distant object and focuses it into an image that can then be examined by some means.

Refracting telescopes. Small refracting telescopes are used in binoculars, cameras, gunsights, galvanometers, periscopes, surveying instruments, rangefinders, astronomical telescopes, and a great variety of other devices. Parallel or nearly parallel light from the distant object enters from the left, and the objective lens forms an inverted image of it (**Fig. 1**). The inverted image is viewed with the aid of a second lens, called the eyepiece. The eyepiece is adjusted (focused) to form a parallel bundle of rays so that the image of the object may be viewed by the eye without strain. The objective lens is typically compound; that is, it is made up of two or more pieces of glass, of different types, designed to correct for aberrations such as chromatic aberration. To construct a visual refractor, a lens is placed beyond the images formed by the objective and viewed with the eye. To construct a photographic refractor or simply a camera, a photographic plate is placed at the position of the image (**Fig. 2**).

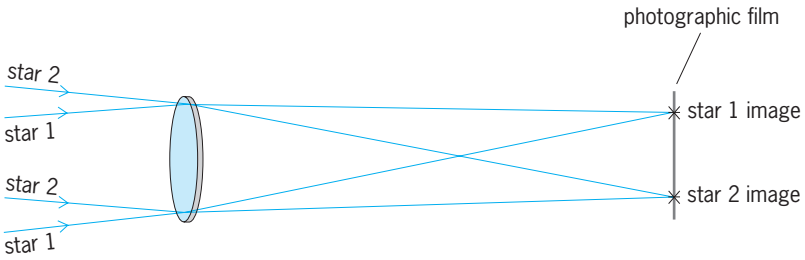


Fig. 2. Refracting optical system used to photograph a star field.

Generally, refracting telescopes are used in applications where great magnification is required, namely, in planetary studies and in astrometry, the measurement of star positions and motions. However, this practice is changing, and the traditional roles of refractors are being carried out effectively by a few reflecting telescopes, in part because of effective limitations on the size of refracting telescopes.

A refractor lens must be relatively thin to avoid excessive absorption of light in the glass. On the other hand, the lens can be supported only around its edge and thus is subject to sagging distortions that change as the telescope is pointed from the horizon to the zenith; thus its thickness must be great enough to give it mechanical rigidity. An effective compromise between these two demands is extremely difficult, making larger refractors unfeasible.

The largest refracting telescope is the 1-m (40-in.) telescope—built over a century ago—at Yerkes Observatory. This size is about the limit for optical glass lenses.

Reflecting telescopes. The principal optical element, or objective, of a reflecting telescope is a mirror. The mirror forms an image of a celestial object (**Fig. 3**) which is then examined with an eyepiece, photographed, or studied in some other manner.

Reflecting telescopes generally do not suffer from the size limitations of refracting telescopes. The mirrors in these telescopes can be as thick as necessary and can be supported by mechanisms that prevent sagging and thus inhibit excessive distortion. In addition, mirror materials having vanishingly small expansion coefficients, together with ribbing techniques that allow rapid equalization of thermal gradients in a mirror, have eliminated the major thermal problems plaguing telescope mirrors. Some advanced reflecting telescopes use segmented mirrors, composed of many separate pieces.

By using a second mirror (and even a third one, in some telescopes), the optical path in a reflector can be folded back on itself (**Fig. 4***a*), permitting a long focal length to be attained with an instrument housed in a short tube. A short tube can be held by a smaller mounting system and can be housed in a smaller dome than a long-tube refractor.

A variety of optical arrangements are possible in large reflecting telescopes, including the prime focus, the newtonian focus, the Cassegrain focus, and the coudé focus.

The newtonian focus is probably most widely used by amateur astronomers in reflectors having apertures on the order of 6 in. (15 cm; Fig. 3*b*). A flat mirror placed at 45° to the optical axis of the primary mirror diverts the focused beam to the side of the telescope, the image being formed by the paraboloidal primary mirror alone. An eyepiece,

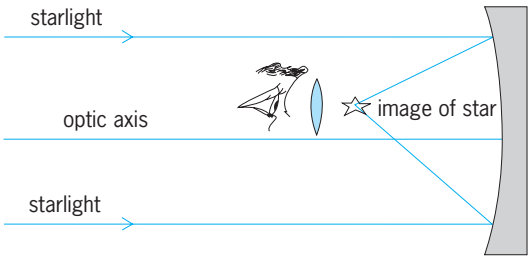


Fig. 3. Viewing a star with a reflecting telescope. In this configuration, the observer may block the mirror unless it is a very large telescope.

camera, or other accessories can be attached to the side of the telescope tube to study the image. In the largest telescopes, with apertures over 100 in. (2.5 m), provision is not usually made for a newtonian focus. Instead, an observing cage is placed inside the tube, where the observer can take accessories to observe the image formed by the primary mirror. This prime focus is identical to the newtonian focus optically, since the newtonian flat does nothing more than divert the light beam. The modern reflectors have fast primary mirrors, so that the focal ratio at the prime focus is *f*/2.5 to *f*/6. Lower focal ratios permit shorter exposures on extended objects such as comets and nebulae.

A Cassegrain system consists of a primary mirror with a hole bored through its center, and a convex secondary mirror that reflects the light beam back through the central hole to be observed behind the primary mirror (Fig. 3*a*). Since the secondary mirror is convex, it decreases the convergence of the light beam and increases the focal length of the system as a whole. The higher focal ratios (*f*/8 to *f*/13) of Cassegrain systems permit the astronomer to observe extended objects, like planets, at higher spatial resolution and to isolate individual stars from their neighbors for detailed studies.

The classical Cassegrain system consists of a paraboloidal primary mirror and a hyperboloidal secondary mirror. The newtonian focus, prime focus, and Cassegrain focus are not affected by spherical aberration. However, all of the systems are plagued by coma, an optical aberration of the paraboloidal primary. Coma causes a point source off the center of the field of view to be spread out into a comet-shaped image. To correct for the effect of coma, a corrector lens is often used in front of the photographic plate. The design of prime-focus corrector lenses is a major consideration in large telescope design, since the corrector lens itself can introduce additional aberrations. *See* ABERRATION (OPTICS); ASTRONOMICAL PHOTOGRAPHY; MIRROR OPTICS.

To avoid complicated corrector lenses at the Cassegrain focus, the Ritchey-Chrétien system, an alternate design with both a hyperboloidal primary and a hyperboloidal secondary, is used in modern telescopes. This arrangement is not affected by coma or spherical aberration, so it has a wider field of view than the classical Cassegrain.

Reflecting telescopes are often used in applications where great light-gathering capabilities are required. Since 1990, a significant number of large astronomical telescopes have been built with mirrors over 8 m (315 in.) in diameter.

Catadioptric telescopes. Catadioptric telescopes combine both mirrors and lenses. This combination is generally used to image wide fields. Catadioptric telescopes with apertures of 15–18 cm (5–6 in.) are popular with amateur astronomers. One variant of the catadioptric system is the Schmidt design.

The Schmidt is an optical system used almost exclusively for photographic applications such as sky surveys, monitoring of galaxies for supernova explosions, and studies of comet tails. The primary mirror of a Schmidt camera has a spherical shape, and therefore suffers from spherical aberration. To correct for this problem, the light passes through a thin corrector plate as it enters the tube. Schmidt camera correcting plates are among the largest lenses made for astronomical applications. The special features of this system combine to produce good images over a far larger angular field than can be obtained in a Cassegrain. *See* SCHMIDT CAMERA.

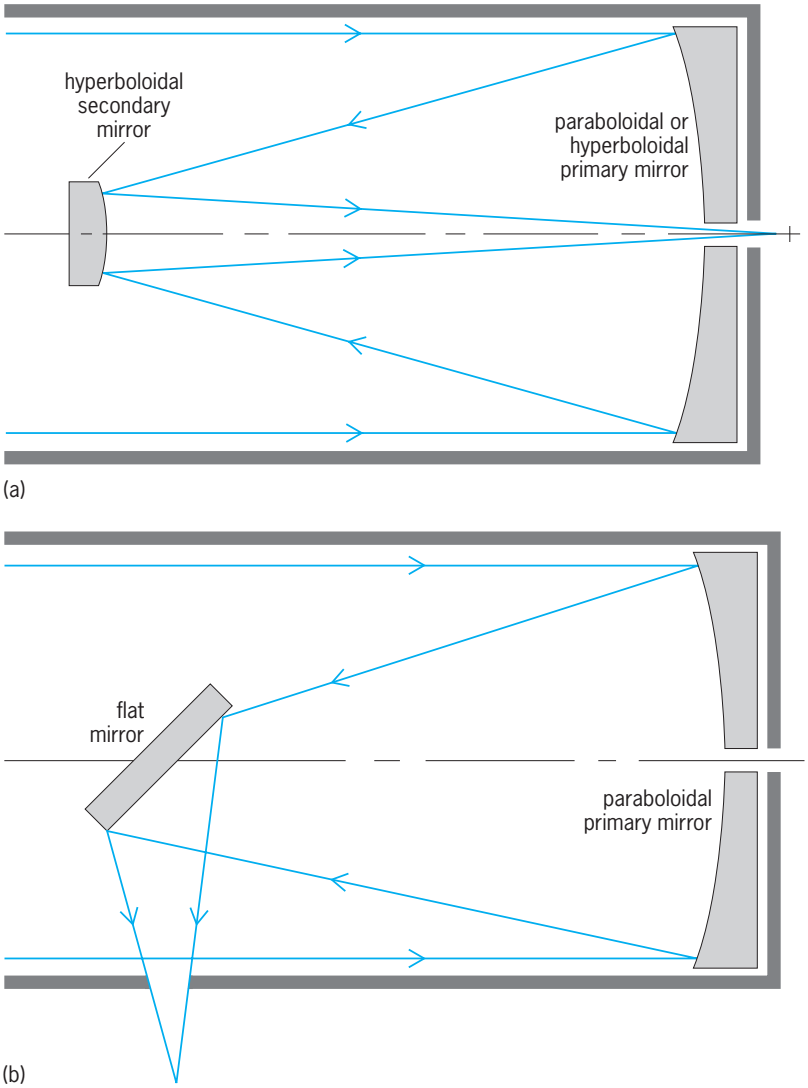


Fig. 4. Diagrams of reflecting telescopes. (*a*) Cassegrain telescope, with either classical or Ritchey-Chrétien optics. (*b*) Newtonian telescope.

Tools. Astronomers seldom use large telescopes for visual observations. Instead, they record their data for future study. Modern developments in photoelectric imaging devices are supplanting photographic techniques for many applications. The great advantages of detectors such as charge-coupled devices is their high sensitivity, and the images can be read out in a computer-compatible format for immediate analysis. *See* CHARGE-COUPLED DEVICES.

Light received from most astronomical objects is made up of radiation of all wavelengths. The spectral characteristics of this radiation may be extracted by special instruments called spectrographs. Wide field coverage is not critical in spectroscopy of stellar objects, so spectrographs are mounted at the Cassegrain and coudé foci. In the coudé configuration, flat mirrors divert the focused beam so that the focal point is fixed in position, regardless of where the telescope points in the sky. Thus a spectrograph or other instrument that is too heavy or too delicate to be mounted on the moving telescope tube can be placed at the coudé focus. *See* ASTRONOMICAL SPECTROSCOPY.

Photoelectric imaging devices may be used in conjunction with spectrographs to record spectral information. Photoelectric detectors are useful tools for classifying stars, monitoring variable stars, and quantitatively measuring the light flux from any astronomical object. Photometry carried out with different filters yields basic information about the source with shorter observing time than that required for a complete spectroscopic analysis. *See* PHOTOMETRY.

Collectors of radiant energy. As collectors of radiation from a specific direction, telescopes may be classified as focusing and nonfocusing. Nonfocusing telescopes are used for radiation with energies of x-rays and above. Focusing telescopes, intended for nonvisible wavelengths, are similar to optical ones, but they differ in the details of construction.

Solar telescopes. Solar instrumentation differs from that designed to study other celestial objects, since the Sun emits great amounts of light energy. In one design (**Fig. 5**), a large heliostar reflects sunlight down the fixed telescope tube to a spectrograph. This spectrograph is evacuated to avoid problems that would be created by hot air currents. In an alternative design, the heliostat feeds light into a vertical telescope that is evacuated. *See* SUN.

Radio telescopes. Radio telescopes utilize mirrors of very large size that, because of the long wavelength for which they are used, may consist of only an open wire mesh. Because of the limitation by diffraction, arrays of large reflectors have been constructed. The output of each mirror of the array is combined in a process called aperture synthesis to yield a resolution roughly equivalent to that provided by a telescope the size of the array. An example of this type is the Very Large Array (VLA) near Socorro, New Mexico. *See* RADIO TELESCOPE.

Infrared telescopes. In infrared telescopes the secondary mirror (that near *A* in illus. *b*) is caused to oscillate rotationally about an axis through a diameter. This motion causes an infrared detector at *B* to see alternately the sky and the sky plus the desired object. The signals received at these two mirror positions are subtracted and, as a consequence, the large background radiation received both from the atmosphere and from the telescope is canceled. However, because of the random nature of thermal radiation, the fluctuations of the background emission are not canceled. Thus infrared telescopes are additionally designed to reduce the telescope background radiation and its fluctuations. Two large infrared telescopes are the 3.8-m (150-in.) United Kingdom Infrared Telescope and the 3-m (120-in.) NASA Infrared Telescope Facility on Mauna Kea. *See* INFRARED ASTRONOMY.

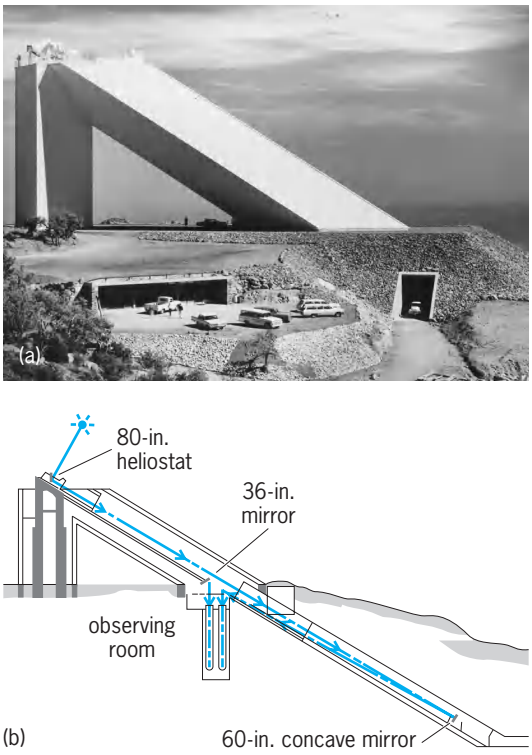


Fig. 5. Solar telescope configuration. (*a*) McMath 1.5-m (60-in.) solar telescope, Kitt Peak, Arizona. (*b*) Diagram of the optical elements in the telescope. 1 in. = 2.5 cm.

Ultraviolet telescopes. Ultraviolet telescopes have special mirror coatings with high ultraviolet reflectivity. Since the atmosphere is not transparent below 300 nanometers, ultraviolet telescopes are usually flown above the atmosphere either in rockets or in orbiting spacecraft. An example is the *International Ultraviolet Explorer*. *See* ULTRAVIOLET ASTRONOMY.

X-ray telescopes. X-ray telescopes must be used above the atmosphere and are flown in rockets or satellites. The focusing type uses an unusual optical design in which the reflection from the surfaces occurs at nearly grazing incidence. This is the only way of achieving reflective optics for x-rays.

Nonfocusing x-ray telescopes use opaque heavy-metal (lead) channels or tubes in front of an x-ray detector to confine the directional sensitivity of the detector. The detectors may be proportional

counters or scintillation detectors. *See* X-RAY TELESCOPE.

Gamma-ray telescopes. Gamma-ray telescopes use coincidence and anticoincidence circuits with scintillation or semiconductor detectors to obtain directional discrimination. With coincidence counting, two or more detectors in a line must give a simultaneous detection for a gamma ray to be counted. Other detectors are often used to surround the telescope to reduce the unwanted background arising from undesired particles. A simultaneous count received in one or more of these shielding detectors nullifies (by anticoincidence) the detection otherwise registered in the coincidence detectors. Thus gamma rays that trigger only the coincidence circuits are detected. Since many gamma rays are produced within the atmosphere by other particles, the telescopes are usually flown in balloons, rockets, or satellites. However, at very high energies (more than 100 GeV), ground-based techniques have been used either to detect the Cerenkov light from the shower of electrons produced when a gamma ray hits the atmosphere or to detect directly the particles that penetrate to the ground. *See* GAMMA-RAY ASTRONOMY.

Cosmic-ray telescopes. Cosmic-ray telescopes are used to detect primary protons or heavier-element nuclei or to detect the products produced when these particles interact with the atmosphere. In its simplest form a cosmic-ray telescope may consist of nuclear track emulsions borne aloft in balloons or spacecraft. A very large cosmic-ray telescope deep in a mine in Utah detects penetrating mesons in the 1–100-teraelectronvolt range. This telescope uses Cerenkov detectors in combination with plane parallel arrays of cylindrical spark-tube counters. The Cerenkov counter detects the presence of a high-energy particle and triggers the acoustic sensing of sparks in the counters. A computer is used to analyze the spark data and determine the direction of the incoming meson. The arrival of high-energy cosmic rays appears to be isotropic, most likely because of scattering by magnetic fields of the galaxy. *See* CERENKOV RADIATION; COSMIC RAYS.

Neutrino telescopes. Although neutrinos are not electromagnetic radiation, devices to collect astrophysically generated neutrinos are, by analogy, often called neutrino telescopes. Neutrinos from several astrophysical sources have been recorded by omnidirectional detectors operating in deep mines. Neutrinos from the Sun have been observed since the 1970s by a detector in a mine in South Dakota. Several underground detectors registered neutrinos from supernova 1987A in February 1987. *See* NEUTRINO; SOLAR NEUTRINOS; SUPERNOVA.

The newest major neutrino detector is the Sudbury Neutrino Observatory (SNO), which was built 2100 m (6800 ft) belowground in a mine near Sudbury, Ontario, and is operated by the Queens University. SNO is designed to detect neutrinos produced by fusion reactions in the Sun and in supernovae in the Milky Way Galaxy. SNO consists of a 12-m-diameter (40-ft) acrylic sphere, filled with 1000 tons of heavy water (D_2O). When neutrinos react with the heavy water in the sphere, it produces flashes of Cerenkov radiation. This light is then detected by an array of 9600 photomultiplier tubes surrounding the heavy-water vessel. The observatory, located in the deep mine, is shielded from cosmic rays by the intervening layers of rock. The detector laboratory was designed to reduce background radiation signals, which might obscure the very weak neutrino signal. *See* NEUTRINO ASTRONOMY.

Notable telescopes. The definition of a large telescope depends on its type. For a refracting telescope to be considered a large telescope, its objective lens must be larger than about 0.6 m (24 in.), whereas a reflecting telescope will have to exceed 79 in. (2 m) to be considered large (see **table**). A few telescopes that are notable for their large size or innovative design will be discussed. *See* ASTRONOMICAL OBSERVATORY.

The 1.02-m (40-in.) refractor at the Yerkes Observatory has been a major contributor to astronomy. The 40-in. objective lens has a focal length of 19 m (62 ft) and is housed in a 23-m (90-ft) dome.

The 5-m (200-in.) Hale telescope at Palomar Mountain, California, was completed in 1950. The primary mirror is 5 m in diameter with a 1.02-m (40-in.) hole in the center. Its focal length is 14 m (660 in.) and it has a paraboloidal figure. The focal ratio of the prime focus is *f*/3.3, and of the Cassegrain focus *f*/16.

The 4-m (158-in.) Mayall reflector at the Kitt Peak National Observatory was dedicated in 1973. The 158-in. mirror has an *f*/2.7 hyperboloidal shape and is made from a 61-cm-thick (24-in.) fused quartz disk which is supported in an advanced-design mirror cell. The prime focus has a field of view six times greater than that of the Hale reflector. An identical telescope was subsequently installed at Cerro Tololo Inter-American Observatory, in Chile.

The mirrors for these traditional large telescopes were all produced using the same general methodology. A large, thick glass mirror blank was first cast; then the top surface of the mirror was laboriously ground and polished to the requisite shape. In the process, great quantities of glass were removed from the blank. The remaining glass helps to give the mirror the required structural rigidity as it is tilted at different angles as the telescope is pointed to different parts of the sky.

The practical and economical limit to the size of traditional mirror designs was nearly reached by the 6-m (236–in.) telescope in the Caucasus Mountains, Russia. Newer telescopes have been designed and built that use either a number of mirrors mounted such that the light collection by them is brought to a common focus, or lightweight mirrors in computer-controlled mounts.

In a multi-mirror telescope (MMT), all mirrors are of similar shape, but additional optical elements are introduced to bring the light to a common focus. An example was the MMT on Mount Hopkins in Arizona, in which six 1.8-m (72-in.) mirrors were combined to collect as much light as a single 4.5-m (177-in.)

Large telescopes of the world

Mirror diameter		Observatory	Year completed
m	in.		
		Some of the largest reflecting telescopes	
10.4	404	Gran Telescopio Canarias (GTC), La Palma, Canary Islands	2006
10.0 × 2	400 × 2	Keck Telescopes, Mauna Kea, Hawaii	1993–1996
10.0	394	South African Large Telescope (SALT), Sutherland, South Africa	2005
9.0	360	Hobby-Eberly Telescope, McDonald Observatory, Fort Davis, Texas	1997
8.4 × 2	331 × 2	Large Binocular Telescope (LBT), Mount Graham, Arizona	2005
8.2	319	Subaru Telescope, National Astronomical Observatory, Japan; Mauna Kea, Hawaii	2000
8.2 × 4	320 × 4	Very Large Telescope (VLT), ESO Paranal Observatory, Cerro Paranal, Chile	1998–2000
8.1	319	Gemini North (Gillett), National Optical Astronomy Observatory, Mauna Kea, Hawaii	1999
8.1	319	Gemini South, National Optical Astronomy Observatory, Cerro Pachon, Chile	2000
6.5	256	Multi-Mirror Telescope, Mount Hopkins, Arizona	1999
6.5 × 2	256 × 2	Magellan Telescopes, Cerro Las Campanas, Chile	2000–2001
6.0	236	Special Astrophysical Observatory, Zelenchukskaya, Caucasus, Russia	1976
5.1	200	California Institute of Technology/Palomar Observatory, Palomar Mountain, California	1950
4.2	165	William Herschel Telescope, La Palma, Canary Islands	1987
4.1	161	Southern Astrophysical Rearch (SOAR) Telescope, Cerro Pachon, Chile	2006
3.8	147	Mayall Telescope Kitt Peak National Observatory, Arizona	1973
4.0	158	Cerro Tololo Inter-American Observatory, Chile	1976
4.0	158	Visible and Infrared Survey Telescope for Astronomy (VISTA), ESO Paranal Observatory, Cerro Paranal, Chile	2007
3.9	153	Anglo-Australian Telescope, Siding Spring Observatory, Australia	1975
3.8	150	United Kingdom Infrared Telescope, Mauna Kea Observatory, Hawaii	1978
3.6	144	Canada-France-Hawaii Telescope, Mauna Kea Observatory, Hawaii	1979
3.6	142	Cerro La Silla European Southern Observatory, Chile	1976
3.6	141	New Technology Telescope, European Southern Observatory, Chile	1989
3.5	138	Calar Alto Observatory, Calar Alto, Spain	1984
3.5	138	Astrophysics Research Consortium, Apache Point, New Mexico	1993
3.5	138	Wisconsin-Indiana-Yale-NOAO (WIYN), Kitt Peak, Arizona	1994
3.6	140	Telescopio Nazionale Galileo (TNG), La Palma, Canary Islands	1998
3.2	126	NASA Infrared Telescope, Mauna Kea Observatory, Hawaii	1979
3.0	120	Lick Observatory, Mount Hamilton, California	1959
2.7	107	McDonald Observatory, Fort Davis, Texas	1968
2.6	102	Crimean Astrophysical Observatory, Ukraine	1960
2.6	102	Byurakan Observatory, Yerevan, Armenia	1976
2.6	102	VLT Survey Telescope, ESO Paranal Observatory, Cerro Paranal, Chile	2006
2.5	100	Nordic Optical Telescope, La Palma, Canary Islands	1989
2.5	100	Mount Wilson and Las Campanas, Mount Wilson, California	1917
2.5	100	Carnegie Southern Observatory, Cerro Las Campanas, Chile	1976
2.5	96	Isaac Newton Telescope, La Palma, Canary Islands	1984
2.5	98	Sloan Telescope, Apache Pont, New Mexico	1998
2.4	94	University of Michigan-Dartmouth College-Massachusetts Institute of Technology, Kitt Peak, Arizona	1986
2.4	94	Hubble Space Telescope, low Earth orbit	1990
		Largest refracting telescopes	
1.02	40	Yerkes Observatory, Williams Bay, Wisconsin	1897
0.91	36	Lick Observatory, Mount Hamilton, California	1888
0.83	33	Observatoire de Paris, Meudon, France	1893
0.80	32	Astrophysikalisches Observatory, Potsdam, Germany	1899
0.76	30	Allegheny Observatory, Pittsburgh, Pennsylvania	1914

mirror. The multi-mirror design for this particular telescope was abandoned in favor of a single mirror design. Segmented mirror telescopes (SMT) have also been built.

New generation telescopes. The Keck Telescopes on Mauna Kea, Hawaii, completed in 1993–1996, are the largest of the segmented mirror telescopes to be put into operation (**Fig. 6**). The telescopes are a fairly traditional design. However, their primary mirrors are made up of 36 individual hexagonal segments mosaiced together to form single 10-m (386-in.) mirrors. Electronic sensors built into the edges of the segments monitor their relative positions, and feed the results to a computer-controlled actuator system which maintains the relative positions of the segments to 5 nm (0.2×10^{-6} in.). The mosaic has four times the light-gathering power of the 5-m (200-in.) Hale Telescope, though its weight is the same as the 200-in. mirror. The two Keck telescopes are 85 m (280 ft) apart.

Work is under way to combine the light of the two 10-m telescopes to produce an optical interferometer. This requires a number of 2-m (80-in.) telescopes in addition to the Keck giants. By combining the light from the large telescopes and the 2-m telescopes, the interferometer will be capable of finding Jupiter-size planets orbiting nearby stars. *See* EXTRASOLAR PLANETS; INTERFEROMETRY.

In 1989, the European Southern Observatory put into operation their New Technology Telescope. The 3.58-m (141-in.) mirror was produced by a technique known as spin-casting, where molten glass is poured into a rotating mold. As the glass cools and solidifies, the surface of the relatively thin mirror takes

on a shape that is close to the desired one, reducing substantially the need for grinding away excess glass. In operation, computers monitor the mirror's shape, and direct a support system that will maintain the proper shape.

The Hobby-Eberly Telescope (HET) is a 9-m (360-in.) telescope built as an international collaboration between the University of Texas at Austin, Pennsylvania State University, and Stanford University in the United States and two universities in Germany. It uses a 91-segment spherical main mirror that is tipped at a constant 35° from the zenith. To move from observation to observation, the telescope moves only in azimuth. During a single observation, the telescope remains fixed in azimuth and objects are tracked by moving a spherical aberration corrector. Since the primary mirror remains at a fixed angle with respect to gravity, significant cost savings were realized in the mirror support system. The Hobby-Eberly Telescope was designed for astronomical spectroscopy. It began operations in 1998. The Southern African Large Telescope (SALT), completed in 2005, is based on the HET design.

Worldwide efforts are under way on a new generation of large, ground-based telescopes, using both the spin-casting method and the segmented method to produce large mirrors. The Gemini project of the National Optical Astronomy Observatory has built twin 8.1-m (319-in.) telescopes, one on Mauna Kea, Hawaii, and the other on Cerro Pachon in Chile. The Gemini North telescope (now called Gillett; **Fig. 7**) was dedicated in June 1999. Gemini South, in the high dry mountains of central Chile, went into operation in 2000. Each telescope is on an alt-azimuth mounting, requiring precise computer control of altitude, azimuth, and image rotation in order to track astronomical sources.

The Very Large Telescope (VLT), operated by the European Southern Observatory on Cerro Paranal, Chile, consists of four 8-m (315-in.) "unit" telescopes with spin-cast mirrors and several 1.8-m (71-in.) auxiliary telescopes. The light from the telescopes is combined to give the equivalent light-gathering power of a 16-m (630-in.) telescope. The last of the four unit telescopes began collecting scientific data in September 2000. In 2001, the telescopes of the VLT system were successfully used together as the VLT Interferometer (VLTI). This system has successfully measured the angular diameters of several bright stars. Cerro Paranal is in the Atacama Desert, where it almost never rains (the average precipitation is less than a centimeter per year). The clear, dry skies are ideal for infrared observations.

Very close to the Keck telescopes on the summit of Mauna Kea is the Subaru Telsecope, an 8.2-m. (319-in.) optical-infrared telescope that achieved first light in 1999 (**Fig. 8**). The Subaru Telescope is the main telescope of the National Astronomical Observatory of Japan. Its main mirror is a single 22.8-ton piece of low-expansion glass. The observatory dome of the Subaru was designed to minimize local atmospheric turbulence. The 8.2-m mirror is supported by an active system, consisting of 261 actuators, that maintains its surface accuracy with high precision.

Fig. 6. Keck Telescope, the world's largest optical-infrared telescope. The primary mirror, 10 m (386 in.) in diameter, is viewed through the W. M. Keck Observatory's open aperture. (*R. Ressmeyer-Starlight; California Association for Research in Astronomy*)

The Large Binocular Telescope (LBT), built by an international collaboration, has two 8.4-m (331-in.) telescopes on a common binocular mount. The pair has the light-gathering power of an 11.8-m (465-in.) telescope and the resolution of a 22.8-m (900-in.) instrument. The Large Binocular Telescope, located on Mount Graham in Arizona, achieved first light in October 2005.

Space-based telescopes. A number of space-based telescopes have been highly successful. The *International Ultraviolet Explorer* (*IUE*), launched into geosynchronous orbit in 1978, is not a giant telescope; its mirror is only 45 cm (18 in.) in diameter. The satellite, positioned over the Atlantic Ocean, was operated in real time for almost 19 years to obtain high-dispersion ultraviolet spectra of solar system objects, stars, nebulae, and extragalactic objects.

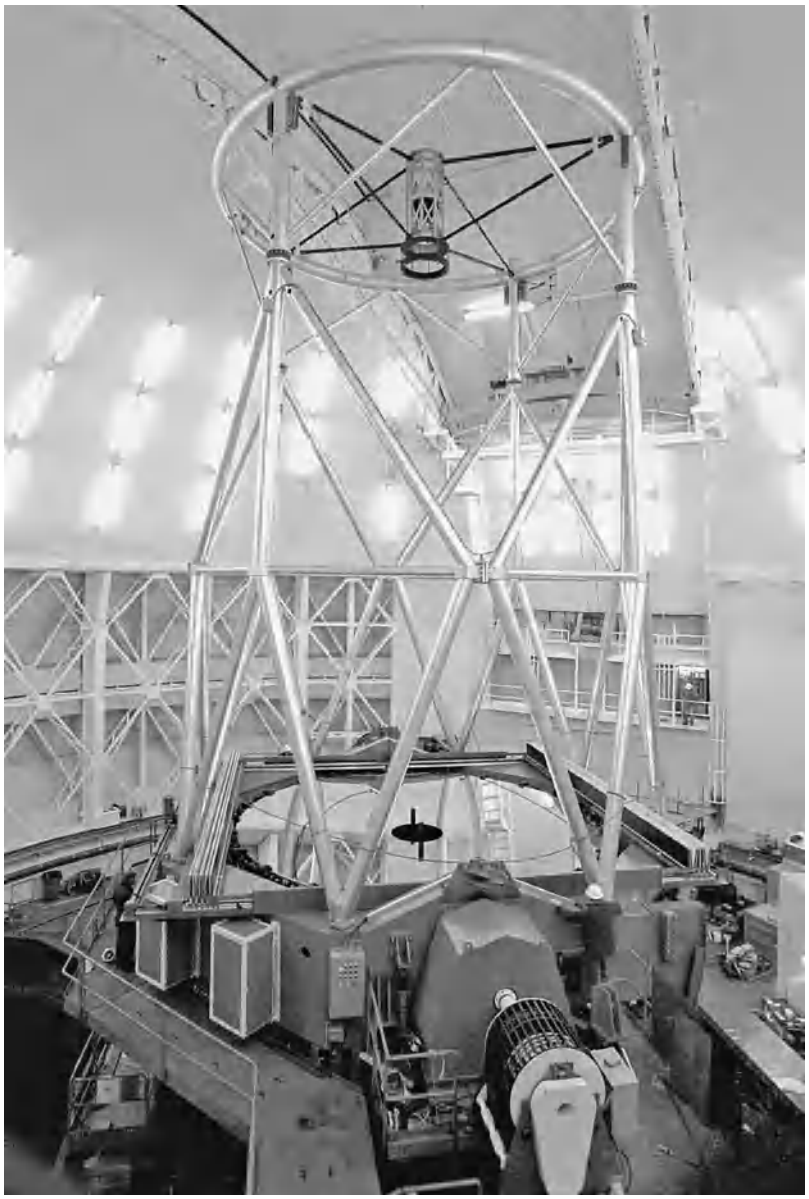


Fig. 7. Gemini North (Gillett) telescope in its dome. (*Copyright Gemini Observatory/AURA/NOAO/NSF, all rights reserved*)

The ability of large telescopes to resolve fine detail is limited by a number of factors. The ultimate limitation set by diffraction of light is rarely or never achieved by large telescopes. Instead, the finest resolved detail is limited by atmospheric inhomogeneities which set a limit of about 0.2 arcsecond. Distortion due to the mirror's own weight causes additional problems in astronomical telescopes. The Earth-orbiting *Hubble Space Telescope* (*HST*), with an aperture of 2.4 m (94 in.), was designed to eliminate these problems. The telescope operates in ultraviolet as well as visible light, resulting in a great improvement in resolution not only by the elimination of the aforementioned terrestrial effects but by the reduced blurring by diffraction in the ultraviolet. *See* DIFFRACTION; RESOLVING POWER (OPTICS).

The *Hubble Space Telescope* was launched into Earth orbit by the space shuttle in 1990. It has a Ritchie-Chrétien reflector which feeds light into a number of scientific instruments. Soon after the telescope was launched, it was discovered that the optical system was plagued with spherical aberration, which severely limited its spatial resolution. In December 1993, space-shuttle astronauts serviced and repaired the telescope, adding what amounts to eyeglasses for the scientific instruments. Since that first servicing mission, the *Hubble Space Telescope* has exceeded its prelaunch specifications for spatial resolution. A second servicing mission occurred in 1997, in which two new scientific instruments replaced first-generation instruments. One of the new instruments, the Near Infrared Camera Multi-Object Spectrography, extended the *Hubble Space Telescope's* capabilities into the infrared. A third servicing mission was flown in 1999 and a forrth in 2002. *See* HUBBLE SPACE TELESCOPE.

The *Einstein Observatory* (*High Energy Astronomy Observatory 2*), launched in 1978, and *ROSAT*, launched in 1990, carried telescopes that imaged soft x-rays. The *Chandra X-Ray Observatory* (one of NASA's Great Observatories) was launched in 1999. This space observatory (a follow-on to the *Einstein Observatory*) was named after the Indian-American scientist Subrahmanyan Chandrasekhar. The *Chandra X-Ray Observatory* is provides significantly better resolution than earlier x-ray

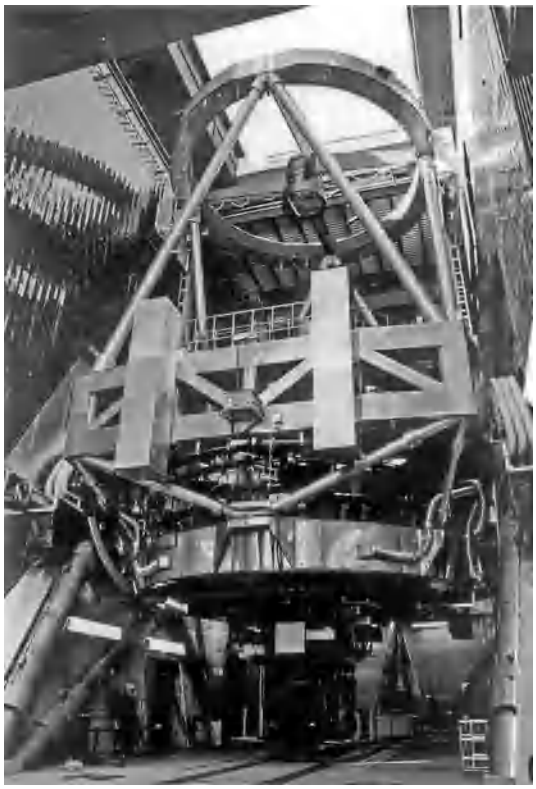


Fig. 8. Subaru Telescope of the National Astronomical Observatory of Japan (NAOJ). (© *The Subaru Telescope, National Astronomical Observatory of Japan*)

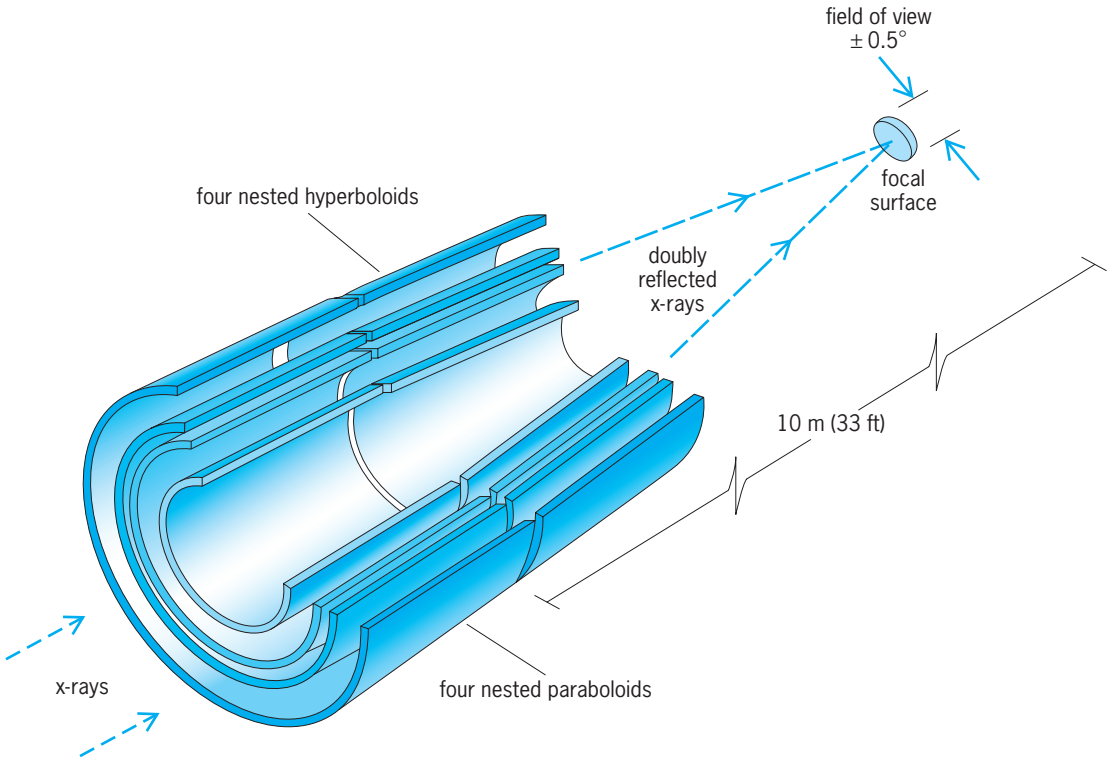


Fig. 9. The four nested Wolter Type I telescopes that are the x-ray focusing optics of the *Chandra X-Ray Observatory*. The mirror elements are 0.8 m (2.4 ft) long and 0.6–1.2 m (2.4 ft) in diameter. (*Raytheon Optical Systems, TRW Inc., and Chandra X-Ray Center, Smithsonian Astrophysical Observatory*)

observatories. The optics of the *Chandra X-Ray Observatory* consist of four nested, barrel-shaped telescopes called Wolter Type I telescopes (**Fig. 9**). These telescopes operate such that x-rays will be reflected if they strike a mirror at grazing angles, while the rays would simply pass through the mirror if they struck it at nearly right angles. The four nested telescopes provide a large cross section to collect adequate quantities of x-rays, and provide about 0.5-arcsecond resolution. The telescopes are made of Zerodur glass, coated with iridium. *See* CHANDRA X-RAY OBSERVATORY; X-RAY ASTRONOMY.

The *James Webb Space Telescope* (JWST) will have a deployable 6.5-m-diameter (252-in.) mirror, and will be placed in an orbit approximately 1.5 km (0.93 mi) from the Earth. It will be optimized to study infrared radiation. Current plans are for a launch in the 2013 time frame. *See* SATELLITE (ASTRONOMY); SCIENTIFIC SATELLITES.

Robert D. Chapman; William M. Sinton

Bibliography. Keck VLT interferometers break new ground, *Sky Telesc.*, 106(4):24, October 2003; P. R. Lawson, Optical interferometry comes of age, *Sky Telesc.*, 105(5):30–39, May 2003; L. Mertz, *Excursions in Astronomical Optics*, Springer-Verlag, New York, 1996; P. Moore, *Eyes on the Universe: The Story of the Telescope*, Springer-Verlag, London, 1997; H. S. Stockman, *Next Generation Space Telescope: Visiting a Time When Galaxies Were Young*, AURA, 1997; Subaru sees first light, *Sky Telesc.*, 97(5):18, May 1999.

Telestacea

An order of the cnidarian subclass Alcyonaria (Octocorallia). Telestacea are typified by *Telesto*, which forms an erect branching colony by lateral budding from the body wall of an elongated primary or axial polyp. The stolon is bandlike or membranous. Sclerites are scattered singly, partly fused, or entirely fused to form a rigid tube. *See* OCTOCORALLIA (ALCYONARIA).

Kenji Atoda

Teletypewriter

An electronic-mechanical device, also called a teleprinter, for preparing, storing, transmitting, and receiving messages over a telegraph or data communication circuit. The term teletypewriter was used first for teleprinters utilized in teletypewriter exchange (TWX) service in the early 1930s. It was later applied to terminals used in other services, and became interchangeable with the terms teleprinter and printing telegraph terminal. A generic term, it should be used instead of Teletype, which is a registered trademark of the AT&T Teletype Corporation. Teletypewriters were invented by several individuals in the early 1900s, initially being adaptations of the commercial typewriters of the day. Relays, magnets, and solenoids were added to provide telegraph signal generation from the operation of the typewriter keyboards, and caused the typebaskets

or typewheels to print automatically on sheets of paper.

When greater ruggedness was required, the inventors devised more sturdy mechanisms. Unreliable machinery and open-wire telegraph circuits caused many errors to occur, necessitating repetition of an errored message to get good printed copies. The use of punched tape helped somewhat, with operators fixing their perceived keyboarding errors before transmission, and expedited the repetition process by allowing reruns of the punched tape. The next generation of teletypewriters printed on narrow paper tape that was gummed on the back. Each line of a message was moistened from a sponge or wick, cut off, and pasted onto a telegram form, with errors being corrected by pasteovers (a repetition of the error portion of the message, in correct form). Although this was labor-intensive, with low wage scales the tape printers were commercially successful. Later, the increasing cost of labor led to the return to page printers exclusively. *See* TELEGRAPHY.

Commercially obsolete machines can be used by deaf persons who cannot use regular telephones. Several thousand teletypewriters have been equipped with modems or acoustic couplers for attachment to telephone handsets or lines. A visible alarm lamp is provided to signal the deaf person whenever a call is being received. Many of the sets are equipped with automatic answering facilities, including an answer-back message transmitter that identifies the terminal, so that messages can be received in the owner's absence. If a regular telephone happens to connect to one of these terminals, the person making the call will hear a whistle as the station answers. An improved instrument for deaf people is a portable teletypewriter with a built-in modem which can couple acoustically to any standard telephone handset, allowing use anywhere. These teletypewriters have compact keyboards, and some are equipped with a display that shows one line of characters, allowing conversational visible communication. Others may have a very compact printing mechanism that uses a narrow roll of paper, similar to that used in some calculators; other designs feature a small CRT display. Ransom D. Slayton

Bibliography. *Teletypewriter Circuits and Equipment: Fundamentals*, 1991.

Television

The electrical transmission and reception of transient visual images. Like motion pictures, television consists of a series of successive images which are registered in the brain as a continuous picture because of the persistence of vision. Each visual image impressed on the eye persists for a fraction of a second. At the television transmitter, minute portions of a scene are sampled individually for brightness and color, and the information for each portion is transmitted consecutively. At the receiver, each portion is synchronized and reproduced in its proper position and with correct brightness and color to reproduce the original scene.

Fundamental Principles

The technology of television is based on the conversion of light rays from still or moving scenes and pictures into electronic signals for transmission or storage, and subsequent reconversion into visual images on a screen. A similar function is provided in the production of motion picture film. However, whereas film records the brightness variations of a complete scene on a single frame in an exposure no longer than a fraction of a second, the elements of a television picture must be scanned one piece at a time. In the television system, a scene is dissected into a frame composed of a mosaic of picture elements or pixels. (A pixel is defined as the smallest area of a television image that can be transmitted within the parameters of the system.) This process is accomplished by (1) analyzing the image with a photoelectric device in a sequence of horizontal scans from the top to the bottom of the image to produce an electric signal in which the brightness and color values of the individual picture elements are represented as voltage levels of a video waveform; (2) transmitting the values of the picture elements in sequence as voltage levels of a video signal; and (3) reproducing the image of the original scene in a video signal display of parallel scanning lines on a viewing screen.

Scanning lines and fields. The image pattern of electrical charges on a camera pickup element (corresponding to the brightness levels of a scene) are converted to a video signal in a sequential order of picture elements in the scanning process. At the end of each horizontal line sweep, the video signal is blanked while the beam returns rapidly to the left side of the scene to start scanning the next line. This process continues until the image has been scanned from top to bottom to complete one field scan. *See* TELEVISION CAMERA.

After completion of this first field scan, at the midpoint of the last line, the beam again is blanked as it returns to the top center of the target where the process is repeated to provide a second field scan. The spot size of the beam as it impinges upon the target must be sufficiently fine to leave unscanned areas between lines for the second scan. The pattern of scanning lines covering the area of the target, or the screen of a picture display, is called a raster.

Because of the half-line offset for the start of the beam return to the top of the raster and for the start of the second field, the lines of the second field lie in between the lines of the first field. Thus, the lines of the two are interlaced. The two interlaced fields constitute a single television frame. **Figure 1** shows a frame scan with interlacing of the lines of two fields.

Reproduction of the camera image on a cathode-ray tube (CRT) is accomplished by an identical operation, with the scanning beam modulated in density by the video signal applied to an element of the electron gun. This control voltage to the CRT varies the brightness of each picture element on the phosphor

screen. Alternatively, the image may be displayed on a flat-panel display device. *See* CATHODE-RAY TUBE; FLAT-PANEL DISPLAY DEVICE; PICTURE TUBE.

Blanking of the scanning beam during the return trace (retrace) is provided for in the video signal by a "blacker-than-black" pulse waveform. In addition, in most receivers and monitors another blanking pulse is generated from the horizontal and vertical scanning circuits and applied to the CRT electron gun to ensure a black screen during scanning retrace.

The interlaced scanning format, standardized for monochrome and compatible color, was chosen primarily for two partially related and equally important reasons: (1) to eliminate viewer perception of the intermittent presentation of images, known as flicker, and (2) to reduce the video bandwidth requirements for an acceptable flicker threshold level.

The standard adopted by the Federal Communications Commission (FCC) for monochrome television in the United States specified a system of 525 lines per frame, transmitted at a frame rate of 30 Hz, with each frame composed of two interlaced fields of horizontal lines. Initially in the development of television transmission standards, the 60-Hz power line waveform was chosen as a convenient reference for vertical scan. Furthermore, in the event of coupling of power line hum into the video signal or scanning-deflection circuits, the visible effects would be stationary and less objectionable than moving hum bars or distortion of horizontal-scanning geometry. In the United Kingdom and much of Europe, a 50-Hz interlaced system was chosen for many of the same reasons. With improvements in television receivers, the power line reference was replaced with a stable crystal oscillator.

The initial 525-line monochrome standard was retained for color in the recommendations of the National Television System Committee (NTSC) for compatible color television in the early 1950s. The NTSC system, adopted in 1953 by the FCC, specified a scanning system of 525 horizontal lines per frame, with each frame consisting of two interlaced fields of 262.5 lines at a field rate of 59.94 Hz. Forty-two of the 525 lines in each frame are blanked as black picture signals and reserved for transmission of the vertical scanning synchronizing signal. This results in 483 visible lines of picture information. *See* TELEVISION SCANNING.

Synchronizing video signals. In monochrome television transmission, two basic synchronizing signals are provided to control the timing of picture-scanning deflection: (1) horizontal sync pulses at the line rate, and (2) vertical sync pulses at the field rate in the form of an interval of wide horizontal sync pulses at the field rate. Included in this interval are equalizing pulses at twice the line rate to preserve interlace in each frame between the even and odd fields (offset by one-half line).

In color transmissions, a third synchronizing signal is added during horizontal scan blanking to provide a frequency and phase reference for signal encoding circuits in cameras and decoding circuits in receivers. These synchronizing and reference signals are combined with the picture video signal to form a composite video waveform (**Fig. 2**).

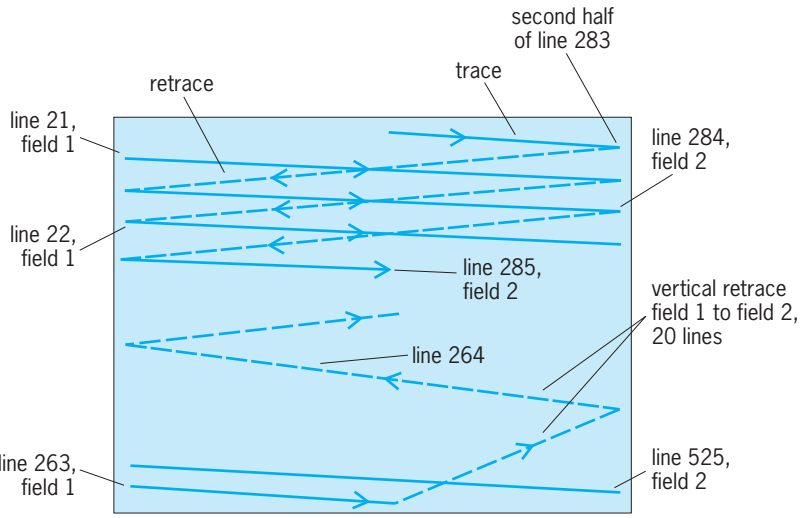


Fig. 1. Interlace scanning pattern (raster) of the television image.

The receiver scanning and color-decoding circuits must follow the frequency and phase of the synchronizing signals to produce a stable and geometrically accurate image of the proper color hue and saturation. Any change in timing of successive vertical scans can impair the interlace of even and odd fields in a frame. Small errors in horizontal scan timing of lines in a field can result in a loss of resolution in vertical line structures. Periodic errors over several lines that may be out of the range of the horizontal-scan automatic frequency control circuit in the receiver will be evident as jagged vertical lines.

Industry standards. Three primary color transmission standards are in use today:

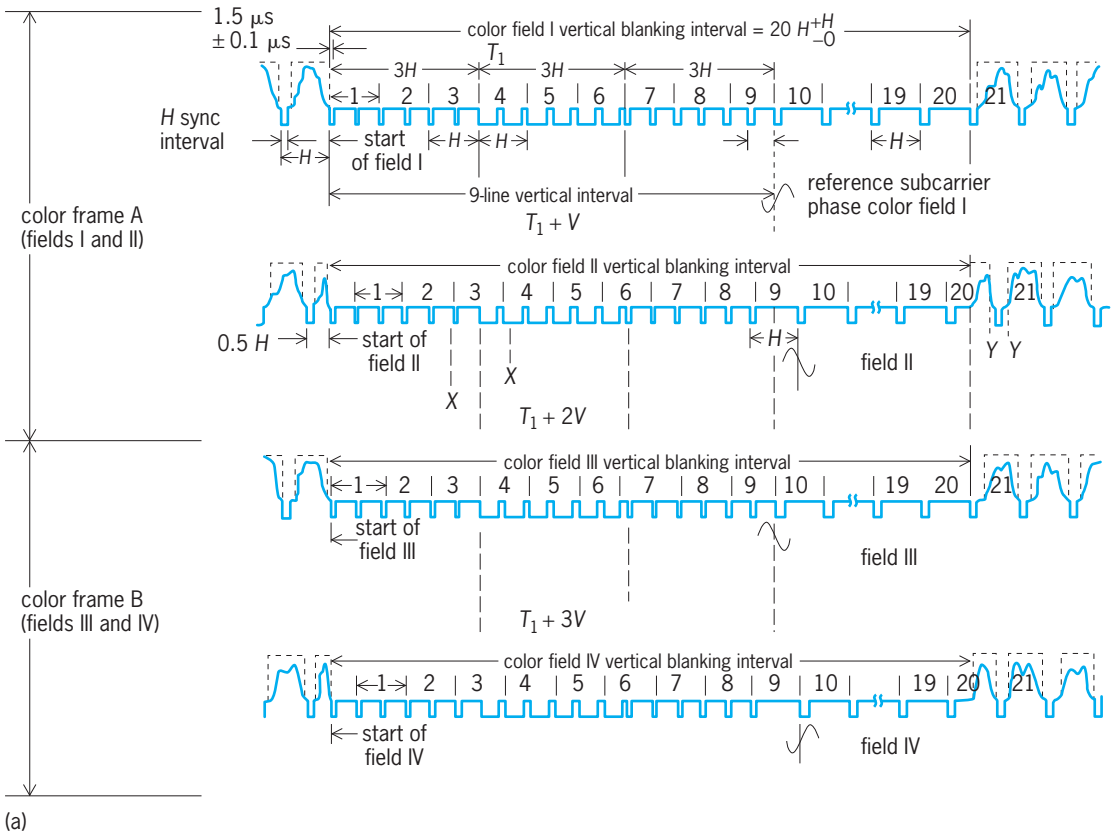
1. NTSC (National Television Systems Committee)—used in the United States, Canada, Mexico, Central America, some of South America, and Japan. In addition, NTSC is used in various countries or possessions heavily influenced by the United States.

2. PAL (Phase Alternate each Line)—used in the United Kingdom, most countries and possessions influenced by England, most European countries, and China. Variation exists in PAL systems.

3. SECAM (SEquential Color with [Avec] Memory)—used in France, countries and possessions influenced by France, the nations of the former Soviet Union, other former Soviet Bloc nations, including East Germany, and other areas influenced by Russia.

The three standards are incompatible for a variety of technical reasons, but standard converter devices are commercially available.

Transmission frequencies. The band of frequencies assigned to a television station for the transmission of synchronized picture and sound signals is called a television channel. In the United States a television channel is 6 MHz wide, with the visual carrier frequency 1.25 MHz above the lower edge of the band and the aural carrier 0.25 MHz below the upper edge of the band.



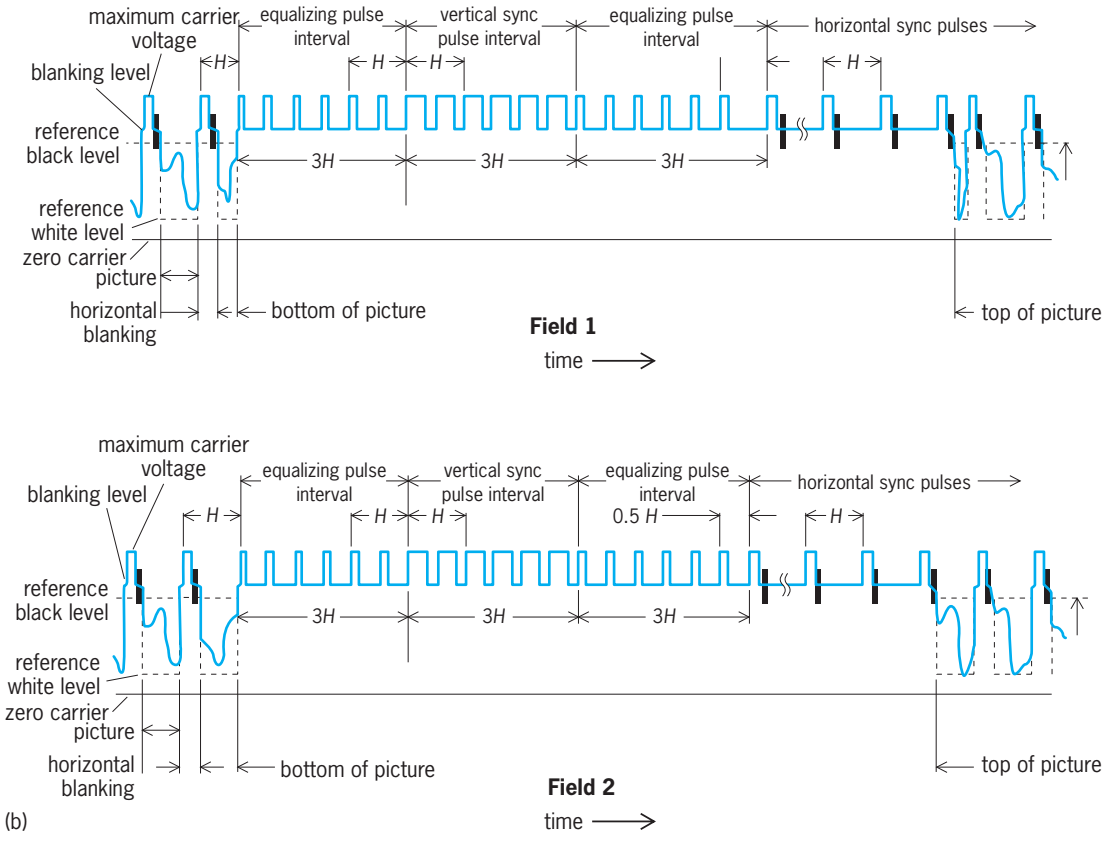


Fig. 2. NTSC color television waveform. (*a*) Principal components. (*b*) Detail of picture elements. (*Electronic Industries Association*)

TABLE 1. Television channels in the United States

Channel number	Frequency band, MHz	Channel number	Frequency band, MHz
2	54–60	36	602–608
3	60–66	37	608–614
4	66–72	38	614–620
5	76–82	39	620–626
6	82–88	40	626–632
7	174–180	41	632–638
8	180–186	42	638–644
9	186–192	43	644–650
10	192–198	44	650–656
11	198–204	45	656–662
12	204–210	46	662–668
13	210–216	47	668–674
14	470–476	48	674–680
15	476–482	49	680–686
16	482–488	50	686–692
17	488–494	51	692–698
18	494–500	52	698–704
19	500–506	53	704–710
20	506–512	54	710–716
21	512–518	55	716–722
22	518–524	56	722–728
23	524–530	57	728–734
24	530–536	58	734–740
25	536–542	59	740–746
26	542–548	60	746–752
27	548–554	61	752–758
28	554–560	62	758–764
29	560–566	63	764–770
30	566–572	64	770–776
31	572–578	65	776–782
32	578–584	66	782–788
33	584–590	67	788–794
34	590–596	68	794–800
35	596–602	69	800–806

Television channels in the United States are identified by numbers, starting with channel 2. The frequency originally assigned to channel 1 was later reassigned to other uses. These channels are in three frequency bands (**Table 1**). Channels 2–6 occupy 54 to 88 MHz, channels 7–13 are 174 to 216 MHz, and channels 14–69 are 470 to 806 MHz. The first two groups of channels fall in the very high frequency (VHF) band; the channels in the last group are in the ultrahigh-frequency (UHF) band. When digital television, whose stations operate at different radio frequencies, is widely implemented, the current radio frequencies used by the analog television stations will be used for other transmision services. *See* RADIO SPECTRUM ALLOCATIONS.

Grades of service. The FCC has established two grades of television service for the United States. Grade A service provides relatively high freedom from interference from other television stations and also good freedom from artificial and receiver noise. It specifies that picture quality acceptable to the average observer is expected to be available at least 90% of the time at the best 70% of all receiver locations at the outer geographical limits of the service area. Grade B service may be more vulnerable to interference and noise. It specifies that service equal to that of grade A is available to only 50% of all receiver locations at the service area limit.

Sound transmission. In the United States, the sound portion of the program is transmitted by frequency modulation (FM) at a carrier frequency 4.5 MHz above the picture carrier. Maximum frequency deviation (bandwidth) of the sound signals is 25 kHz. Additional sound channels can be provided using narrowband FM subcarriers located near the primary audio channel. *See* FREQUENCY MODULATION.

The normal frequency response is altered in the transmitter to emphasize the higher audio frequencies with respect to the lower frequencies. This preemphasis is accomplished by a circuit that causes the audio response to increase with frequency. A corresponding circuit is used in the receiver to produce an equal and opposite decrease of response to higher audio frequencies. By so doing, noise produced in the receiver is attenuated without the overall audio-frequency response being affected. *See* TELEVISION RECEIVER; TELEVISION STANDARDS; TELEVISION TRANSMITTER.

Digital Television

Digital television (DTV) has ushered in a new era in television broadcasting. The impact of DTV is more significant than simply moving from an analog system to a digital system. DTV permits a level of flexibility wholly unattainable with analog broadcasting. An important element of this flexibility is the ability to expand system functions by building upon the technical foundations specified in the basis standards.

With NTSC, and its PAL and SECAM counterparts, the video, audio, and some limited data information are conveyed by modulating a radio-frequency (RF) carrier in such a way that a receiver of relatively simple design can decode and reassemble the various elements of the signal to produce a program consisting of video and audio, and perhaps related data (for example, closed captioning). As such, a complete program is transmitted by the broadcaster that is essentially in finished form. In the DTV system, however, additional levels of processing are required after the receiver demodulates the RF signal. The receiver processes the digital bit stream extracted from the received signal to yield a collection of program elements (video, audio, or data) that match the services the consumer has selected. This selection is made using system and service information that is also transmitted. Audio and video are delivered in digitally compressed form and must be decoded for presentation. Audio may be monophonic, stereo, or multichannel. Data may supplement the main video-audio program (for example, closed captioning, descriptive text, or commentary), or it may be a standalone service (for example, a stock or news ticker). *See* CLOSED CAPTION TELEVISION.

The nature of the digital television system is such that it is possible to provide new features that build upon the infrastructure within the broadcast plant and the receiver. One of the major enabling developments of digital television, in fact, is the integration of significant processing power in the receiving device itself. Historically, in the design of any broadcast system—be it radio or television—the goal has always been to concentrate technical sophistication (when needed) at the transmission end and thereby

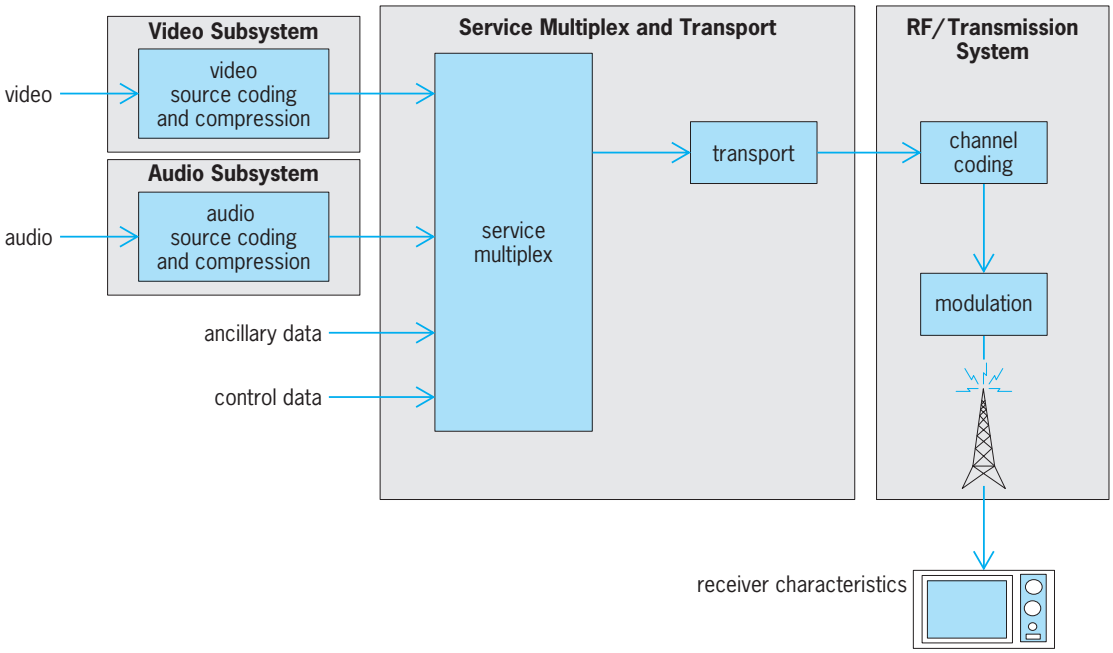


Fig. 3. ITU-R Digital terrestrial television broadcasting model of the International Telecommunications Union, Radiocommunication Sector (ITU-R). (*From Advanced Television Systems Committee, ATSC Digital Television Standard, Doc. A/53B, 2001*)

facilitate simpler receivers. Because there are far more receivers than transmitters, this approach has obvious business advantages. While this trend continues to be true, the complexity of the transmitted bit stream and compression of the audio and video components require a significant amount of processing power in the receiver, which is practical because of the enormous advancements made in computing technology. Once a receiver reaches a certain level of sophistication (and market success), additional processing power is essentially "free."

The development of digital television occurred over a period of many years on several continents. Three major systems emerged:

1. The Advanced Television Systems Committee (ATSC), a North American–based DTV standard organization, developed the ATSC terrestrial DTV series of standards.
2. The Digital Video Broadcasting Project (DVB), a European-based standard organization, developed the DVB series of DTV standards, which were standardized by the European Telecommunication Standard Institute (ETSI).
3. The Integrated Service Digital Broadcasting (ISDB) standards, a series of DTV standards, were developed and standardized by the Association of Radio Industries and Business (ARIB) in Japan.

The remainder of this article focuses on the ATSC DTV system.

ATSC DTV system overview. The Digital Television Standard describes a system designed to transmit high-quality video and audio and ancillary data over a single 6-MHz channel. The system can deliver about 19 megabits per second in a 6-MHz terrestrial broadcasting channel and about 38 Mbits/s in a 6-MHz cable television channel. This means that encoding high-definition video essence at raw data rates that are typically greater than 1 gigabit per second requires a bit rate reduction by about a factor of 50. To achieve this bit rate reduction, the system is designed to be efficient in utilizing available channel capacity by exploiting complex video and audio compression technology. The compression scheme optimizes the throughput of the transmission channel by representing the video, audio, and data sources with as few bits as possible while preserving the level of quality required for the given application. *See* DATA COMPRESSION.

A basic block diagram representation of the ATSC DTV system is shown in **Fig. 3**. According to this model, the digital television system can be seen to consist of three subsystems: source coding and compression, service multiplex and transport, and RF/transmission. The RF/transmission subsystems described in the Digital Television Standard are designed specifically for terrestrial and cable applications. The structure is such that the video, audio, and service multiplex and transport subsystems are useful in other applications.

Source coding and compression refers to the bit-rate reduction methods (data compression) appropriate for application to the video, audio, and ancillary digital data streams. (The term ancillary data encompasses control data; conditional-access (private) data; data associated with the program audio and video services, such as closed captioning; and independent program services.) The purpose of the coder is to minimize the number of bits needed to represent the audio and video information.

The modulation subsystem offers two modes: the terrestrial broadcast mode, 8-VSB (8-level vestigal sideband); and the high-data-rate mode, 16-VSB

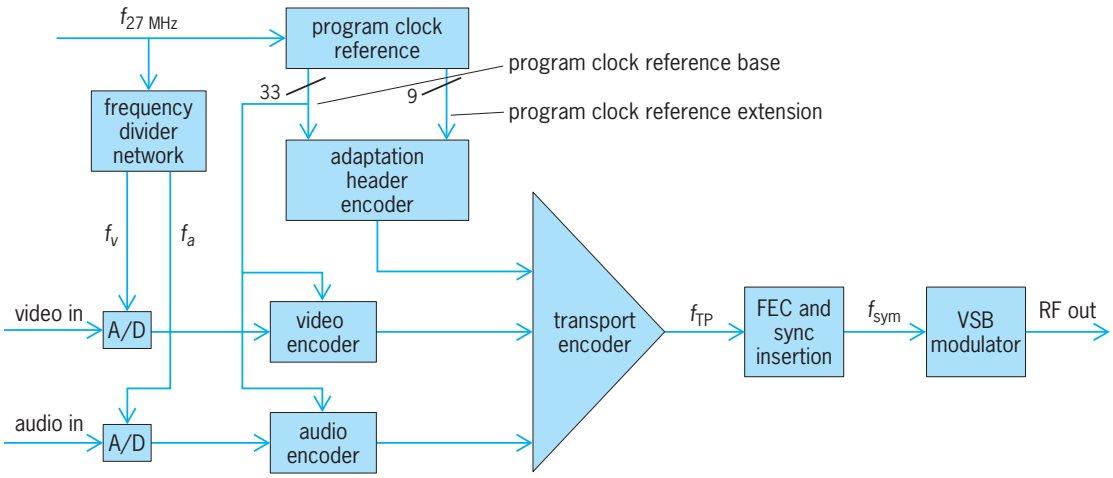


Fig. 4. High-level view of the DTV encoding system, illustrating the relationship and derivation of various clock frequencies within the encoder and the use of forward error correction (FEC). A/D = analog-to-digital. VSB = vestigial sideband. (*From Advanced Television Systems Committee, ATSC Digital Television Standard, Doc. A/53B, 2001*)

(16-level vestigial sideband). *See* AMPLITUDE MODULATION.

Figure 4 gives a high-level view of the encoding equipment. This view is not intended to be complete, but is used to illustrate the relationship and derivation of various clock frequencies within the encoder and the use of forward error correction (FEC) in the system.

The DTV video-compression algorithm conforms to the Main Profile syntax of the standard ISO/IEC 13818-2 (MPEG-2). The allowable parameters are bounded by the upper limits specified for the Main Profile/High Level in this standard. **Table 2** lists the allowed compression formats under the ATSC DTV Standard.

ATSC system block diagram. A basic block diagram representation of the ATSC DTV system is shown in **Fig. 5**. According to this model, the digital television system consists of four major elements, three within the broadcast plant plus the receiver.

Application encoders and decoders. The application encoders and application decoders, as used in Fig. 5, apply bit-rate reduction methods, also known as data compression, appropriate for application to the video, audio, and ancillary digital data streams. The purpose of compression is to minimize the number of bits needed to represent the audio and video information. The DTV system employs the MPEG-2 video stream syntax for the coding of video and the ATSC standard Digital Audio Compression (AC-3) for the coding of audio.

As noted above, ancillary data is a broad term that includes control data and data associated with the program audio and video services. As standards were developed to define how to transport and process data, it became clear that different forms of data served very different purposes, and different standards were needed for metadata and essence. (essence = content − metadata). In this context, (video) essence is the image itself without any ancillary data, and metadata is information about the data essence. Including system information as ancillary data is not strictly proper, as some such data are needed to reassemble the audio, video, and data services. Data delivered as a separate payload can provide independent services as well as data elements related to an audio- or video-based service.

Transport packetization and multiplexing. Transport packetization and multiplexing refers to the means of dividing each bit stream into "packets" of information, the means of uniquely identifying each packet or packet type, and the appropriate methods of

TABLE 2. ATSC DTV compression format constraints

Vertical size value	Horizontal size value	Aspect ratio	Frame-rate code*	Scanning sequence
1080†	1920	16:9, square pixels	1,2,4,5	Progressive
			4,5	Interlaced
720	1280	16:9, square pixels	1,2,4,5,7,8	Progressive
480	704	4:3, 16:9	1,2,4,5,7,8	Progressive
			4,5	Interlaced
	640	4:3, square pixels	1,2,4,5,7,8	Progressive
			4,5	Interlaced

*Frame-rate code: 1 = 23.976 Hz, 2 = 24 Hz, 4 = 29.97 Hz, 5 = 30 Hz, 7 = 59.94 Hz, 8 = 60 Hz.
†1088 lines actually are coded in order to satisfy the MPEG-2 requirement that the coded vertical size be a multiple of 16 (progressive scan) or 32 (interlaced scan).
source: After Advanced Television Systems Committee, *ATSC Digital Television Standard*, Doc. A/53B, 2001.

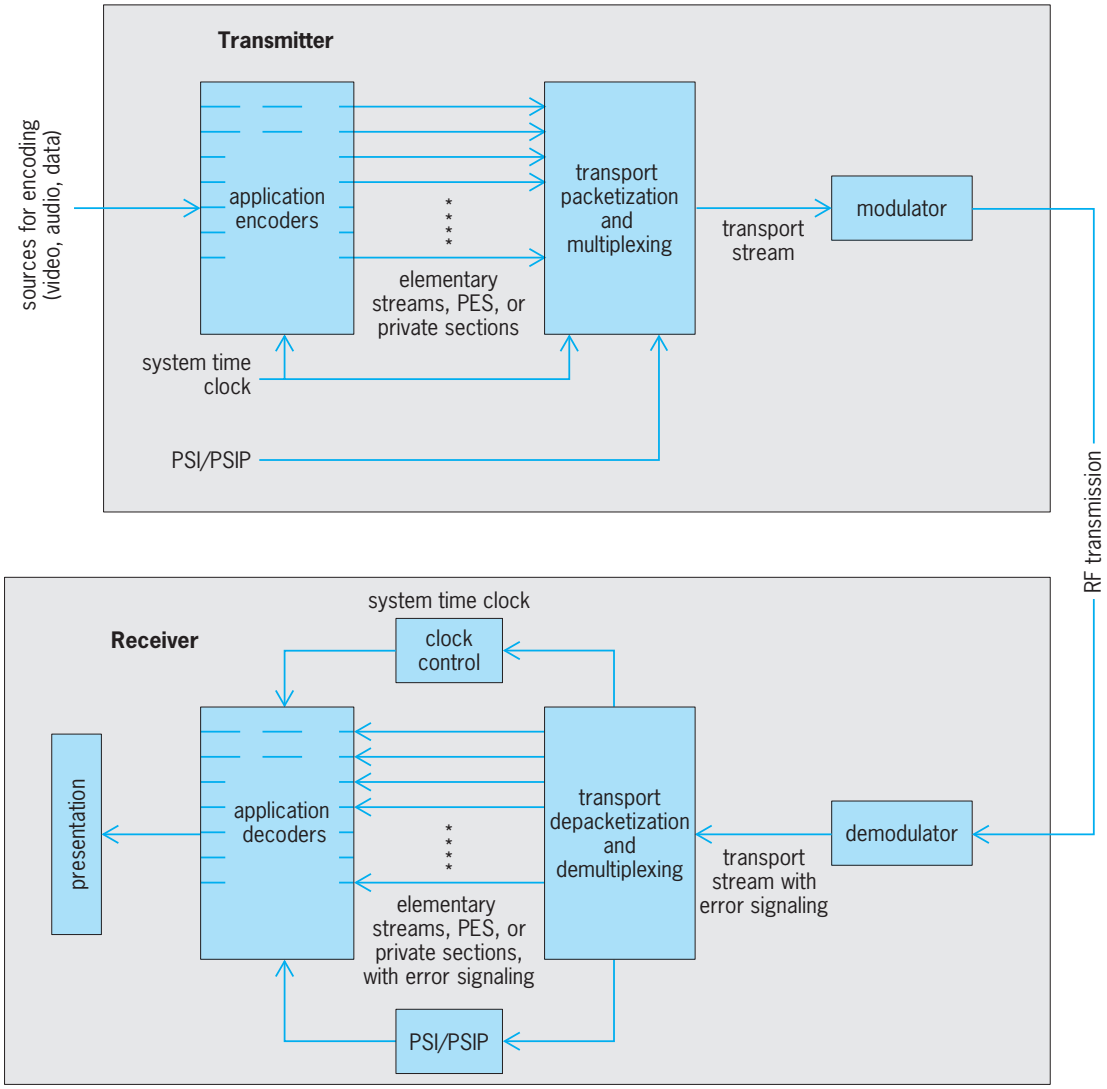


Fig. 5. Sample organization of functionality in a transmitter-receiver pair for a single DTV program. PSIP = Progam and System Information Protocol, PES = packetized elementary stream, PSI = program-specific information. (*From Advanced Television Systems Committee, Guide to the Use of the Digital Television Standard, Doc. A/54A, 2003*)

interleaving or multiplexing video bit stream packets, audio bit stream packets, and data bit stream packets into a single transport mechanism. (The inverse operations of transport depacketization and demultiplexing are carried out in the receiver.) The structure and relationships of these essence bit streams are carried in service information bit streams, also multiplexed into the single transport mechanism. In developing the transport mechanism, interoperability among digital media—such as terrestrial broadcasting, cable distribution, satellite distribution, recording media, and computer interfaces—was a prime consideration. The DTV system employs the MPEG-2 transport stream syntax for the packetization and multiplexing of video, audio, and data signals for digital broadcasting systems. The MPEG-2 transport stream syntax was developed for applications where channel bandwidth or recording media capacity is limited and the requirement for an efficient transport mechanism is paramount.

RF transmission. RF transmission refers to channel coding and modulation. The channel coder takes the digital bit stream and adds information that can be used by the receiver to reconstruct the data from the received signal which, due to transmission impairments, may not accurately represent the actual transmitted signal. The modulation (or physical layer) uses the digital bit stream information to modulate a carrier for the transmitted signal. As noted above, the modulation subsystem offers two modes: 8-VSB and 16-VSB.

Receiver. The ATSC receiver must recover the bits representing the original video, audio, and other data from the modulated signal. In particular, the receiver must tune to the selected 6-MHz channel; reject adjacent channels and other sources of interference; demodulate (equalize as necessary) the received signal, applying error correction to produce a transport bit stream; identify the elements of the bit stream using a transport layer processor; select each desired element and send it to its appropriate processor; decode and synchronize each element; and present the programming. *See* DEMODULATION.

Noise, interference, and multipath are elements of the terrestrial transmission path, and the receiver circuits are expected to deal with these impairments.

Innovations in equalization, automatic gain control, interference cancellation, and carrier and timing recovery create product performance differentiation and improve signal coverage. *See* AUTOMATIC GAIN CONTROL (AGC); EQUALIZER.

High-definition television. The resolution of the displayed picture is the most basic attribute of any video production system. Generally speaking, a high-definition television (HDTV) image has approximately twice as much luminance definition horizontally and vertically as the 525-line NTSC system or the 625-line PAL and SECAM systems. The total number of luminance picture elements (pixels) in the image, therefore, is four times as great. The wider aspect ratio of the HDTV system adds even more visual information. The HDTV image is 25% wider than the conventional video image for a given image height; the ratio of image width to height in HDTV systems is 16:9, or 1.78. The conventional video image has a 4:3 aspect ratio.

As a result of these attributes, the HDTV image may be viewed more closely than is customary in conventional television systems. Full visual resolution of the detail of conventional television is available when the image is viewed at a distance equal to about six or seven times the height of the display. The HDTV image may be viewed from a distance of about three times picture height for the full detail of the scene to be resolved.

Definition. Although there is no single, universal definition for HDTV, it is generally accepted to encompass several elements that are described by various consumer, broadcast, and regulatory groups. HDTV offers the potential for approximately twice the horizontal and twice the vertical resolution of current (NTSC) television. When combined with a wide-screen format (16:9 aspect ratio), this can result in considerably more visual information than conventional television. HDTV consumer television sets can have 720 or 1080 active vertical scanning lines, and are capable of decoding the transmitted 720 × 1280 and 1080 × 1920 ATSC formats and displaying them as a 16:9 aspect ratio image. These two high-definition formats can potentially provide over eight times as much picture information as delivered over broadcast NTSC. While these high-definition transmission formats will be supported by such sets, the actual delivered resolution may vary by broadcaster, by product, and by program. HDTV is normally accompanied by digital surround-sound capability.

Production versus transmission systems. Bandwidth is perhaps the most basic factor that separates production HDTV systems from transmission-oriented systems for broadcasting. A closed-circuit system does not suffer the same restraints imposed upon a video image that must be transported by radio-frequency means from an origination center to consumers. This distinction has led to the development of widely varied systems for production and transmission applications. *See* CLOSED-CIRCUIT TELEVISION.

Format development. In the staging of motion picture films intended for theatrical distribution, no provision generally is made for the limitations of conventional video displays. Instead, the full screen, in formats such as CinemaScope with wide aspect ratios, is used by directors for maximum dramatic and sensory impact. Consequently, cropping of essential information often may be encountered on the video screen. This problem is particularly acute in wide-screen features where cropping of the sides of the film frame is necessary to produce a print for video transmission. This objective is met in one of the following ways:

1. Letterbox transmission, with blank areas above and below the wide-screen frame. Audiences in North America and Japan have not generally accepted this presentation format, primarily because of the reduced size of the picture images and the distraction of the blank screen areas.
2. Printing the full frame height and cropping equal portions of the left and right sides to provide a 4:3 aspect ratio. This process frequently is not ideal because, depending upon the scene, important visual elements may be eliminated.
3. Programming the horizontal placement of a 4:3 aperture to follow the essential picture information. Called pan and scan, this process is used in producing a print or in making a film-to-tape transfer for video viewing. Editorial judgment is required for determining the scanning cues for horizontal positioning and, if panning is used, the rate of horizontal movement. This is an expensive and laborious procedure and, at best, it compromises the artistic judgments made by the director and the cinematographer in staging and shooting and by the film editor in postproduction.

One reason for moving to a 16:9 format is to take advantage of consumer acceptance of the 16:9 aspect ratio commonly found in motion picture films. Actually, motion pictures are produced in several formats, including 4:3 (1.33); 2.35, used for 35mm anamorphic CinemaScope film; and 2.2 in a 70mm format Still, the 16:9 aspect ratio generally is supported by the motion picture industry. **Figure 6** illustrates some common aspect ratios. *See* CINEMATOGRAPHY.

Jerry C. Whitaker

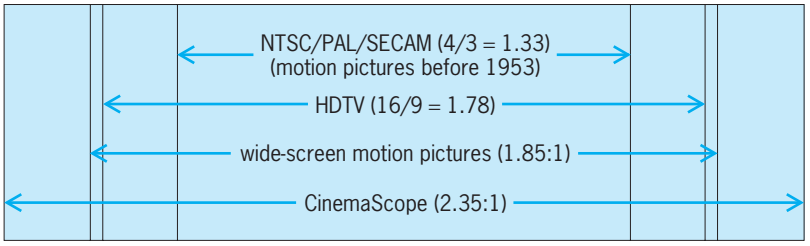


Fig. 6. Common image aspect ratio formats. (*After K. B. Benson and D. G. Fink: HDTV: Advanced Television for the 1990s, McGraw-Hill, New York, 1990*)

Bibliography. J. Boston, *DTV Survival Guide*, McGraw-Hill, New York, 2000; M. Robin and M. Poulin, *Digital Television Fundamentals*, 2d ed., McGraw-Hill, New York, 2000; P. Symes, *Video Compression Demystified*, McGraw-Hill, New York, 2001; J. C. Whitaker (ed.), *NAB Engineering Handbook*, 9th ed., National Association of Broadcasters, Washington, DC, 1997; J. C. Whitaker, *Standard Handbook of Broadcast Engineering*, McGraw-Hill, New York, 2005; J. C. Whitaker, *Standard Handbook of Video and Television Engineering*, 4th ed., McGraw-Hill, New York, 2003.

Television camera

An electrooptical system used to pick up and convert a visual image or scene into an electrical signal called video. The video may be transmitted by cable or wireless means to a suitable receiver or monitor some distance from the actual scene. It may also be recorded on a video tape recorder for playback at a later time.

A television camera may fall within one of several categories: studio (**Fig. 1**), portable, or telecine. It may also be one of several highly specialized cameras used for remote viewing of inaccessible places, such as the ocean bottom or the interior of nuclear power reactors. The camera may be capable of producing color or monochrome (black and white) pictures. Most modern cameras are entirely solid-state, including the light-sensitive element, which is composed of semiconductors called charge-coupled devices (CCDs). Inexpensive or special-purpose cameras, however, may use one or more vacuum tubes, called vidicons, with a light-sensitive surface in lieu of the charge-coupled devices. *See* CHARGE-COUPLED DEVICES.

Television cameras intended for industrial, consumer, or broadcast portable use are usually one piece, with all elements of the camera system contained in one assembly. Such cameras may be combined with a detachable or built-in videocassette recorder to form a camera recorder or camcorder (**Fig. 2**). A broadcast-quality studio camera, on the other hand, usually consists of a separate head and camera control unit (CCU) connected by a multiple-conductor cable.

In such a two-piece camera, the head unit consists of the optical system with lens, the picture pickup devices, and a minimum of electronics necessary to generate and amplify the signals from the pickup devices. It will also have a small, built-in television screen that functions as a viewfinder. The camera control unit may contain electronics and controls that allow a skilled operator to adjust brightness (luminance), color (balance, saturation, and hue), and, for cameras that employ pickup tubes, certain correction circuits (registration, gamma, and aperture) that improve the picture. A special-purpose oscilloscope, called a waveform monitor, is generally provided to the operator of the camera control unit so that accurate voltage levels may be set. Modern cameras provide for automatic as well as manual setup of some or all of the above adjustments. Some cameras have a built-in microprocessor that enables complete setup of all parameters by simply pushing a button. Triax cameras utilize a small-diameter triaxial cable (three concentric conductors) to connect the camera head to the camera control unit, instead of heavier, multiple-conductor cable with as many as 81 conductors. Triax can be used because all of the normal signals are multiplexed onto the three wires. The one-piece cameras generally incorporate similar functions with somewhat lower levels of performance and flexibility, owing to their smaller size. *See* MICROPROCESSOR.

Fig. 1. Television studio camera. (*Thomson Multimedia*)

Fig. 2. Televison camcorder. (*Thomson Multimedia*)

Essential elements. Every camera shares certain essential elements: an optical system, one or more picture pickup devices, preamplifiers, scanning circuits, blanking and synchronizing circuits, video processing circuits, and control circuits. Color cameras also include some kind of color-encoding circuit.

Optical system. The optical system consists minimally of a fixed-focal-length lens placed directly in front of a pickup device. Provision must be made to focus the image on the focal plane of the pickup device. Usually the pickup device, complete with its mounting and scanning components, is adjusted with respect to the lens, while the lens is set at the distant extreme of its focusing range. This establishes the correct back-focus so that a sharp image can be obtained by adjusting only the lens-focusing ring as camera-to-subject distances vary.

All but the most rudimentary cameras replace the fixed-focal-length lens with a zoom-type lens that allows a smooth, continuous transition from wide-angle to telephoto focal lengths while the image remains focused. Modern consumer cameras and camcorders almost always provide an autofocus feature, based on an infrared rangefinder circuit, so that subjects at various distances from the camera will

remain in focus without operator intervention. Some advanced professional cameras, and many consumer-grade cameras, offer an image stabilization feature that makes steady pictures attainable even if the camera cannot be held steady. Most cameras provide for the insertion of various filters into the optical path in order to correct for the color temperature of the lighting or to create special optical effects. *See* COLOR FILTER; GYROSCOPE; RANGEFINDER (OPTICS).

Some portable cameras, especially those intended to capture fast-moving subjects, are equipped with either mechanical or electronic shutters, which can be used to increase the effective frame rate. This significantly reduces the blurring of the image that would occur at the normal video vertical field rate [59.94 Hz for the NTSC (National Television Systems Committee) color system] when, for example, the camera is shooting a fast-moving object. Shutters also cause a loss of sensitivity, which increases with the shutter speed.

Color cameras must split the incoming light into suitable primary colors (usually red, green, and blue). On very inexpensive consumer cameras, this is generally done by means of a stripe filter (with thin vertical, alternating stripes of red, green, and blue) placed in front of a single pickup device. Somewhat better performance is achieved in industrial-grade cameras by using two or more pickup devices with a relay lens and dichroic mirror system placed between the objective lens and the pickup devices (**Fig. 3***a*). Dichroic mirrors reflect light of one color while passing all others; that is, a red dichroic mirror reflects red and passes all other colors. The highest-performance (and most expensive) broadcast-quality cameras utilize three pickup devices and a glass prism with dichroic materials coated onto the glass (Fig. 3*b*).

The advantages of the costlier prism system include ruggedness and, most importantly, the elimination of air-to-glass surfaces except at the entry and exit points. This reduces image deterioration due to misalignment of the mirrors and dust accumulation on the surfaces. The prism also makes a good mount for charge-coupled-device pickup devices, which can be aligned and permanently bonded to the prism by the manufacturer. *See* DICHROISM.

Pickup device. The picture pickup device used in most modern cameras is a type of highly integrated solid-state circuit called a charge-coupled device. Charge-coupled devices consist of a large number of photodiodes (light-sensitive semiconductor junctions) aligned in a matrix so that each diode's output voltage can be related to a particular point in the picture that is focused on the array (**Fig. 4**). As photons of light strike the junctions, small voltages are developed; these voltages are amplified and processed many times each second by circuitry attached to the charge-coupled device. This circuitry constructs an electrical signal that represents the total image striking the charge-coupled device. Charge-coupled device pickups are very resistant to shock, have precise image geometry, and should have nearly indefinite life. They also offer freedom from microphonic noise and image burn-in, which are commonly found in pickup tubes. *See* PHOTODIODE.

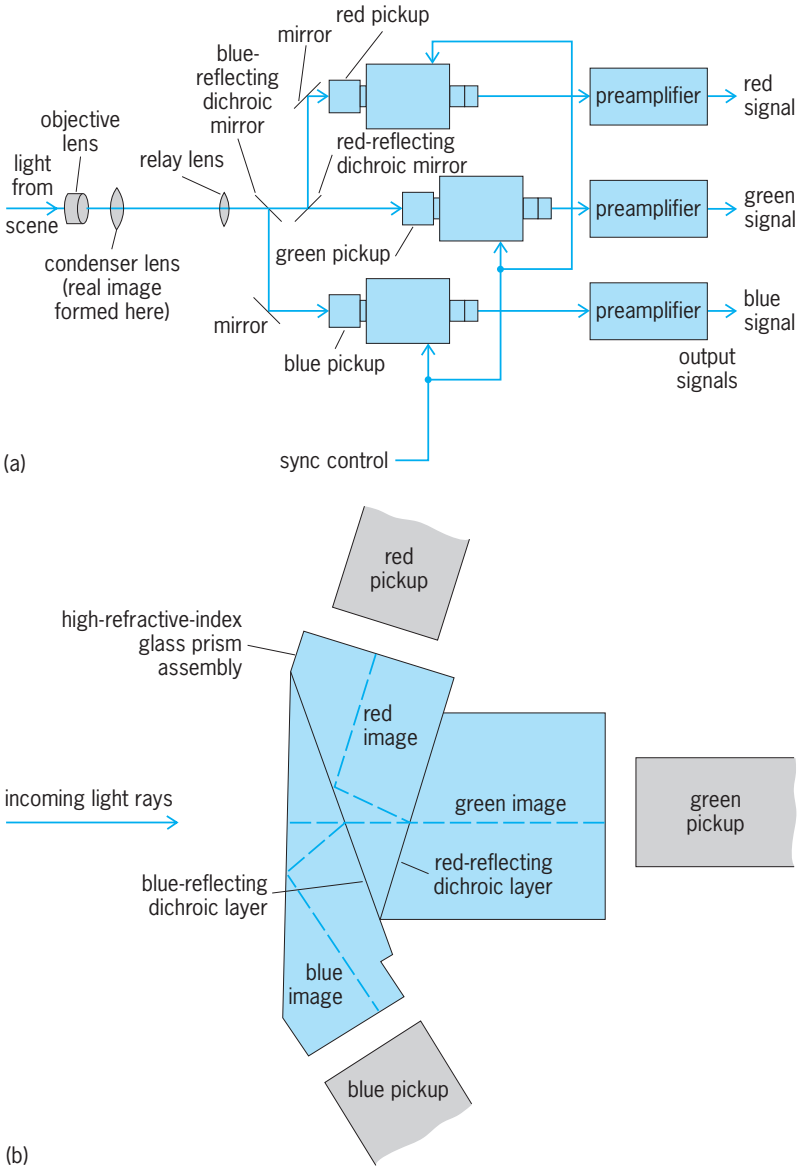


Fig. 3. Color television camera optical systems. (*a*) Dichroic mirror system (*after G. W. Bartlett, ed., NAB Engineering Handbook, 6th ed., National Association of Broadcasters, 1975*). (*b*) Dichroic prism system.

In those older cameras that still use vacuum pickup tubes, the tube's light-sensitive target is scanned in a geometric raster by an electron beam, resulting in an output voltage that varies in proportion to the amount of light striking each point on the target. Monochrome cameras and less sophisticated, consumer-type color cameras may utilize vidicon pickup tubes. High-quality broadcast cameras may use various types of tubes, such as lead oxide target tubes.

Monochrome cameras invariably have only one pickup device; color cameras may have one, two, or three pickup devices. Modern broadcast color cameras employ three pickup devices, one for each of three primary colors used to derive the full-color spectrum. Red, green, and blue are used in additive

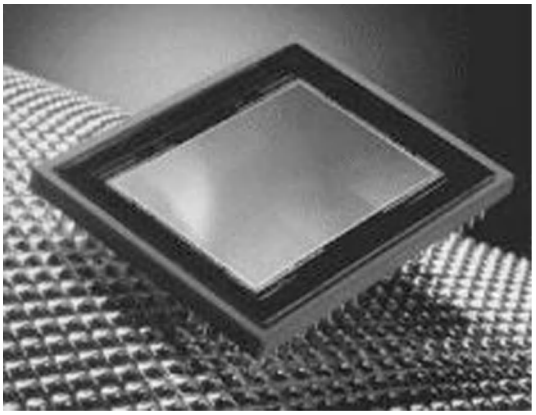


Fig. 4. Charge-coupled-device (CCD) image sensor array. (*Photo courtesy of Philips Semiconductors*)

color systems. While subtractive color systems have been devised that use white (full-spectrum luminance), red, and blue, most modern color cameras use the additive scheme. Each pickup device is scanned in synchronism so that separate red, green, and blue representations of the scene being viewed are always being generated within the camera. *See* COLOR; TELEVISION CAMERA TUBE.

Preamplifier. The pickup device's output is fed to a preamplifier, which helps to maintain a high signal-to-noise ratio and also provides a level sufficient to operate an electronic viewfinder at the camera. *See* PREAMPLIFIER.

Scanning, blanking, and synchronizing circuits. A major difference between charge-coupled devices and pickup tubes is the complexity of required scanning circuits.

The electron beam that scans the photosensitive target of the pickup tube is caused to sweep by means of the action of magnetic fields impressed on the tube by using deflection coils. One coil controls horizontal deflection while the other controls vertical deflection. The coils are often assembled into an integrated yoke assembly. Yokes must be made with great precision and are often computer-matched so that all pickup tubes in a multiple-tube camera scan with precisely the same geometry; otherwise misregistration and color fringing will result. In a mixed-field tube, one of the coils is replaced by electrostatic deflection plates. Horizontal and vertical drive circuits cause the deflection yokes to deflect the electron beam so that it scans the photosensitive target in the pickup tube according to a definite pattern, called a raster. These drive circuits are synchronized by pulses from a synchronizing generator, which also provides horizontal and vertical blanking pulses to suppress the pickup tube output during the retrace interval. If the pickup tube output was not suppressed, there would be objectionable lines through the camera's picture output. *See* TELEVISION SCANNING.

Charge-coupled-device cameras, on the other hand, do not require complicated deflection or retrace blanking circuits. By their very nature, they are scanned by using computer-type addressing techniques. Pulse counters, divider chains, and memory provide the correct sequential readout of the charge-coupled-device array so that a picture is recovered from the individual photodiode outputs.

Before the camera output can be viewed on a conventional video monitor, horizontal and vertical synchronizing (sync) and blanking pulses must be added. In two-piece cameras this is ordinarily done in the camera control unit.

Video processing and control circuits. Other functions that are necessary to obtain high-quality pictures include gamma correction, aperture correction, registration, and color balance. Gamma correction is required because the pickup devices do not respond linearly to increasing light levels. It allows the camera to capture detail in the dark areas of high-contrast scenes, essentially by "stretching" the video levels in those areas. Aperture correction provides several benefits mainly related to an even overall response to scenes with more or less detail. It also helps to improve the signal-to-noise ratio of the camera's output video. Registration must be adjusted on multiple-tube cameras to ensure that the separate red, blue, and green images are precisely aligned on one another; charge-coupled-device cameras are usually registered once, at the factory. Color balance must be properly set on color cameras and must be consistent from dark scenes to bright scenes, or there will be an objectionable tint to the camera output. This is referred to as black balance and white balance. The former is set up with all light blocked from the camera's lens, while the latter is set up by using a so-called white card under the actual lighting conditions of the scene. Color balance is frequently an automatic push-button feature on modern cameras.

All color cameras have a color encoder which combines the three primary colors into a composite signal that must conform to one of several standards used in various countries. In the additive color NTSC system used in the United States, the red (R), green (G), and blue (B) signals are matrixed by a combining network and amplified to provide three new signals, known as M, I, and Q. The M signal contains luminance information, while I and Q contain the color, or chrominance, information from the televised scene. Component M consists of 30% red plus 59% green plus 11% blue. The I signal consists of 60% red minus 28% green minus 32% blue. Component Q is made up of 21% red minus 52% green plus 31% blue. It is possible to subtract voltage levels by inverting the phase and summing, and this is exactly what is done in the color encoder (**Fig. 5**). I and Q then modulate a 3.58-MHz carrier in a two-phase (90° quadrature), balanced modulation system. The resulting amplitude-modulated sidebands are added to the M signal along with the synchronizing and blanking signals already mentioned. A short reference burst of unmodulated 3.58-MHz carrier is also added before the start of each television line. The resultant signal, which is known as composite NTSC video, contains all necessary information to recreate the original color scene on a color monitor or receiver. The M signal is used by black and white

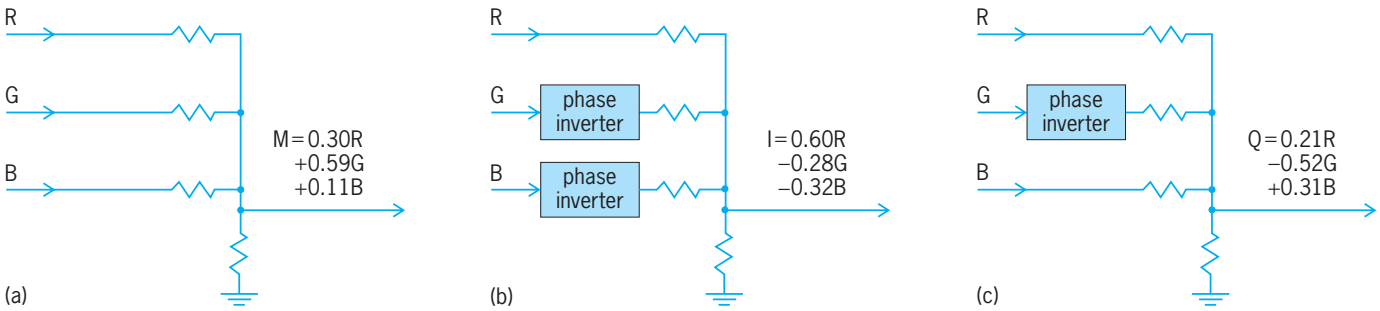


Fig. 5. Color matrix encoding diagrams showing derivation of (*a*) M, (*b*) I, and (*c*) Q signals. (*After G. W. Bartlett, ed., NAB Engineering Handbook, 6th ed., National Association of Broadcasters, 1975*)

monitors or receivers, which are insensitive to I and Q, to create an image in shades of gray. Color monitors and receivers have special circuits to process the I and Q signals back into the primary colors. In a black and white scene the red, green, and blue signals are equal, and therefore I and Q become zero and a black and white picture is produced on both monochrome and color television monitors. Most modern cameras rely heavily on digital integrated circuits and digital signal processing to achieve excellent performance in a compact form factor. *See* INTEGRATED CIRCUITS.

Typical configurations. Studio cameras (Fig. 1) are equipped with several ancillary systems to enhance their operation. An electronic viewfinder (actually a small television monitor) shows the camera operator what the camera is seeing, making it possible to frame and focus the picture. The tally system consists of one or more red lights that illuminate when the camera's picture is "on the line" so that production and on-camera personnel know which camera is active. Generally an intercom system is built into the camera so that the director can communicate with the camera operator. The camera itself may be mounted upon a tripod, but more often it is on a dolly and pedestal, which allows the camera to be moved around on the studio floor and raised or lowered as desired. A pan head permits the camera to be rotated to the left or right and furnishes the actual mounting plate for the camera. The lens zoom and focus controls are mounted on a panning handle convenient to the operator. A common accessory on studio cameras is the videoprompter. This is a television monitor on which the program script may be displayed and read by the on-camera personnel.

Some studio cameras are now being equipped with robotic control systems. These systems can pan, tilt, zoom, and focus the camera. More advanced systems can raise or lower the camera pedestal height, and some can move the camera around the studio floor under the control of a remote operator. The cameras can even be placed entirely under the control of a computer that makes all of the adjustments necessary to obtain the desired camera shots. *See* REMOTE-CONTROL SYSTEM; ROBOTICS.

Telecine cameras are used in conjunction with film or slide projectors to televise motion pictures and still images. Many of the usual controls are automatic so as to require less operator attention

The film projectors used in television may utilize either a constant-rate pulldown or 3-2 intermittent mechanism to translate the motion picture's 24 frames per second to the 30 frames-per-second television frame rate. The constant-rate pulldown has a shutter which is closed before the film is rapidly pulled. The shutter opens 120 times per second so that the camera sees each film frame in a 5:4 ratio. The 3-2 pulldown moves the film intermittently, so that one film frame is held in the film gate for two television fields ($^2/_{60}$ s) and the next film frame is held in the film gate for three television fields ($^3/_{60}$ s). Hence two film frames are displayed in $^5/_{60}$ s, and 24 film frames are shown in 1 s (30 television frames).

Generally, one telecine camera serves several projectors through the use of an optical multiplexer. The optical multiplexer may use stationary, half-silvered mirrors or movable, front-silvered mirrors to direct the image from each projector into the camera. Movable mirrors offer superior light-transmission efficiency and are preferable in color systems to minimize light losses. Telecine camera systems have largely been supplanted by flying-spot or linear-array charge-coupled-device scanners, which convert film images line-by-line directly into a video raster more precisely than the older telecine.

Portable cameras usually combine all of the basic elements into one package and may be used for a multitude of purposes. They have found their way into electronic news gathering for broadcast television, and into electronic field production, where they can be used for production of broadcast programs, commercials, and educational programs. The units often have built-in microphones, videocassette recorders, and batteries for completely self-contained operation (Fig. 2). One person can easily handle these compact and lightweight camcorders. The most compact units fit in the palm of the hand. They have become so popular and inexpensive that they have almost completely supplanted film-based home movie cameras.

HDTV cameras. Cameras used in high-definition television (HDTV) are fundamentally similar in appearance and operation to previous cameras. In fact, some modern cameras are switchable to produce either a conventional output or an HDTV output. The conventional output has a 4:3 aspect ratio raster

and the scan rates match the 525-horizontal-line, 59.94-Hz-vertical-field-rate NTSC standard in the United States. When switched to HDTV mode, the aspect ratio becomes 16:9 and the horizontal scan rate is usually increased to either 720 progressively scanned lines or 1080 interlace-scanned lines with a 60-Hz vertical field rate. Various manufacturers employ proprietary algorithms to derive 16:9 images from a 4:3 charge-coupled-device array and vice versa. *See* TELEVISION; TELEVISION RECEIVER; TELEVISION STANDARDS; TELEVISION STUDIO; TELEVISION TRANSMITTER.

Earl F. Arbuckle, III

Bibliography. T. D. Burrows et al., *Television Production*, 6th ed., 1995; P. Hodges, *Video Camera Operator's Handbook*, 1994; A. C. Luther, *Video Camera Technology*, 1998; National Association of Broadcasters, *NAB Engineering Handbook*, 8th ed., 1992; J. C. Whitaker and K. B. Benson (eds.), *Standard Handbook of Video and Television Engineering*, 3d ed., 2000; H. Zettl, *Video Basics*, 3d ed., 2000.

Television camera tube

An electron tube having a light-sensitive receptor that converts an optical image into an electrical television video signal. The tube is used in a television camera to generate a train of electrical pulses representing the light intensities present in an optical image focused on the tube. Each point of this image is interrogated in its proper turn by the tube, and an electrical impulse corresponding to the amount of light at that point of the optical image is generated by the tube. This signal represents the video or picture portion of a television signal. Television camera tubes are designed for broadcast television to pick up live programs, indoors or outdoors, as well as to reproduce motion pictures and other filmed material. *See* TELEVISION CAMERA.

The tubes are also used extensively in closed-circuit cameras for surveillance, and in training studios, schools, video tape recorder cameras, and military special-purpose cameras. Special versions are designed to work with intensifier tubes which increase the effective sensitivity so that the cameras can operate at very low light levels. These are used for nighttime surveillance work, television astronomy, and viewing low-intensity x-ray fluoroscope images in medical x-rays and in baggage inspection units in airports. In general, three tubes are used in color television cameras. A class of tubes has built-in stripe color filters which allow a single tube to develop a complete color picture, although with somewhat reduced fidelity compared to the multitube cameras. Although the television camera tube is sensitive primarily to visible light, special tubes are sensitive to radiant energy in the infrared and the ultraviolet.

Charge-coupled devices are a new generation of solid-state electronic image sensors which can produce a television signal from an optical image. These can operate independently or can be incorporated into an intensifier-type vacuum tube to achieve enhanced sensitivity. **Figure 1** shows contemporary television camera tubes. *See* CHARGE-COUPLED DEVICES.

Image orthicon. The image orthicon made broadcast television practical. It was used for more than 20 years as the primary studio and field camera tube for black and white and color television programming because of its high sensitivity and its ability to handle a wide range of scene contrast and to operate at very low light levels. It is one of the most complicated camera tubes. It is an outgrowth of the earlier multiplier orthicon and image iconoscope, and was made possible by the invention of the "two-sided" storage target. The image orthicon is divided into an image section, a scanning section, and a multiplier section (**Fig. 2**), within a single vacuum envelope.

Image section. A light image is focused on the photoemissive layer, which is a continuous film inside the tube faceplate. Electrons absorb the energy and leave the surface in numbers proportional to the intensity of the illumination at each point. These photoelectrons flow in essentially parallel streams through the image section. The magnetic field focuses each to a sharp focus at the target plane. *See* PHOTOEMISSION.

The two-sided target consists of a fine wire mesh screen placed several thousandths of an inch from a glass membrane less than 0.0002 in. (5.0 micrometers) thick. Most of the photoelectrons pass through the target mesh and hit the front surface of the target glass. Each photoelectron knocks several additional electrons from the target glass surface, producing a positive charge at the impact point. The secondary electrons are collected by the target mesh, which is held at a slightly more positive voltage. *See* SECONDARY EMISSION.

Scanning section. The positive charge pattern is stored on the front side of the target glass. A beam of low-velocity electrons generated by an electron gun is made to scan the rear surface of the glass by varying magnetic fields produced by the deflecting coils. As the beam moves across the glass, it deposits electrons wherever positive charges are built up on the image side. The glass resistance is controlled so that charges can move from one face to the other before the scanning beam returns to the same spot; yet the glass is of high enough resistance to inhibit lateral movement of the charges. When the scanning beam has deposited enough electrons at each point to neutralize the charge on the glass and reduce it to the voltage of the electron gun cathode, the remaining electrons return toward the electron gun. When the beam scans an uncharged (dark) area, the full beam is returned. When the beam scans a highly charged (bright) area, most of the beam is deposited and little returns. The variations in the return beam current constitute the television picture information, at low intensity. The return beam is amplified about 1000 times in the electron multiplier section of the tube and then is taken out at the anode of the multiplier as a video signal current. *See* CATHODE-RAY TUBE.

Fig. 1. Typical television tubes in modern use. From left to right: image orthicon, lead oxide vidicon, industrial-type vidicon, miniature vidicon, silicon intensifier vidicon, and charge-coupled device. (*RCA Corp.*)

Multiplier section. The electron multiplier is of unique construction, although it operates like the multiplier used in a multiplier phototube. It consists of a flat first-dynode structure and a series of pinwheel multipliers. When the return beam strikes the first dynode, a shower of secondary electrons cascades through the pinwheels, which are maintained at progressively higher voltages, where repeated secondary emission multiplies their number. The final group of electrons is collected by the anode and forms the video signal current. *See* PHOTOMULTIPLIER.

Image isocon. The image isocon is a further development of the image orthicon. The excess primary electrons in the scanning beam of the image isocon are returned from the target in two components: the scattered electrons and those that are reflected (**Fig. 3**). The image isocon works on the principle of separating out these two components of the return scanning beam and utilizing the scattered electron component, which has the highest signal compared to the random noise in the beam current. The separation section (**Fig. 4**) directs only the scattered beam into the electron multiplier. This improves the signal-to-noise ratio of the output signal and allows the camera utilizing the image isocon to operate at very low light levels in such fields as astronomy and intensification of x-ray fluoroscopic images.

Photoconductive tubes. These types have a photoconductor as the light-sensitive portion. A photoconductor is a material that absorbs light and transfers the energy of the photons of light to electrons in the material. This frees some of the electrons and allows them to move through the material, and thereby changes the electrical conductivity of the material where the light is absorbed. The electron tube is designed to detect this change in electrical conductivity and develop a television signal. *See* PHOTOCONDUCTIVITY.

The name vidicon was applied to the first photoconductive camera tube developed by RCA. It is

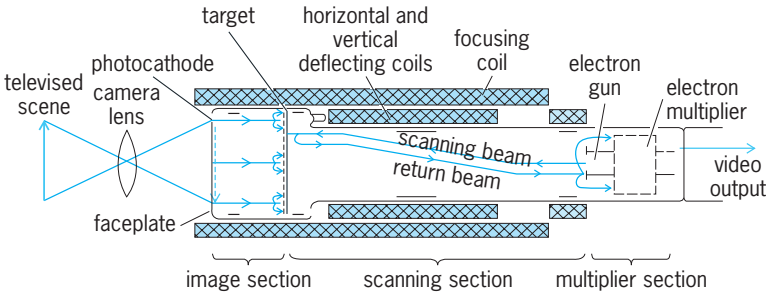


Fig. 2. Image orthicon and its associated deflecting and focusing coils. (*After D. G. Fink, ed., Television Engineering Handbook, McGraw-Hill, 1957*)

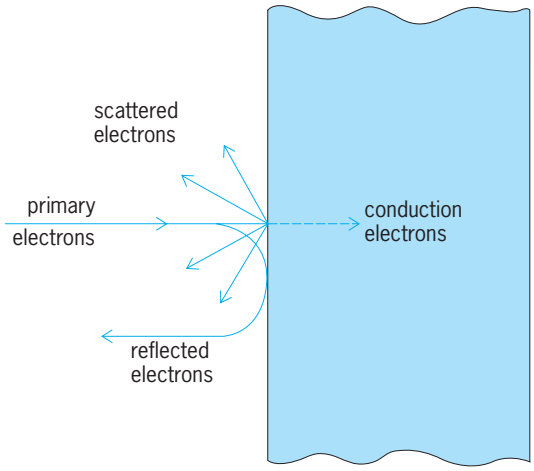


Fig. 3. Formation of the two components in the return beam of the image isocon: the scattered and the reflected electrons.

loosely applied to all photoconductive camera tubes, although some manufacturers adopt their own brand names to identify the manufacturer or the type of photoconductive material used.

The vidicon tube is a small tube that was first developed as a closed-circuit or industrial surveillance television camera tube. The development of new photoconductors has improved its performance to the point where it is now utilized in one form or another in most television cameras. Its small size and simplicity of operation make it well suited for use in systems to be operated by relatively unskilled people.

The vidicon is a simply constructed storage type of camera tube (**Fig. 5**). The signal output is developed directly from the target of the tube and is generated by a low-velocity scanning beam from an electron gun. The target generally consists of a transparent signal electrode deposited on the faceplate of the tube and a thin layer of photoconductive material, which is deposited over the electrode. The photoconductive layer serves two purposes. It is the light-sensitive element, and it forms the storage surface for the electrical charge pattern that corresponds to the light image falling on the signal electrode.

The photoconductor has a fairly high resistance when in the dark. Light falling on the material excites additional electrons into a conducting state, lowering the resistance of the photoconductive material at the point of illumination. A positive voltage is applied to one side of the photoconductive layer by means of the signal electrode. On the other side, the scanning beam deposits sufficient electrons at low velocity to establish a zero voltage. In the interval between successive scans of a particular spot, the light lowers the resistance in relation to its intensity. Current then flows through the surface at this point, and the back surface builds up a positive voltage until the beam returns to scan the point. The signal output current is generated when the beam deposits electrons on these positively charged areas. An equal number of electrons flow out of the signal electrode and through a load resistor, developing a signal voltage that is fed directly to a low-noise video signal amplifier.

A fine-mesh screen stretched across the tube near the target causes the electron scanning beam to decelerate uniformly at all points and approach the target in a perpendicular manner. The beam is brought to a sharp focus on the target by the longitudinal magnetic field of the focusing coil and the proper voltage for the focusing electrode. The beam scans the target under the influence of the varying magnetic fields of the deflecting coils.

Photoconductor properties determine to a large extent the performance of the different types of vidicon tubes. The first and still widely used photoconductor is porous antimony trisulfide. The

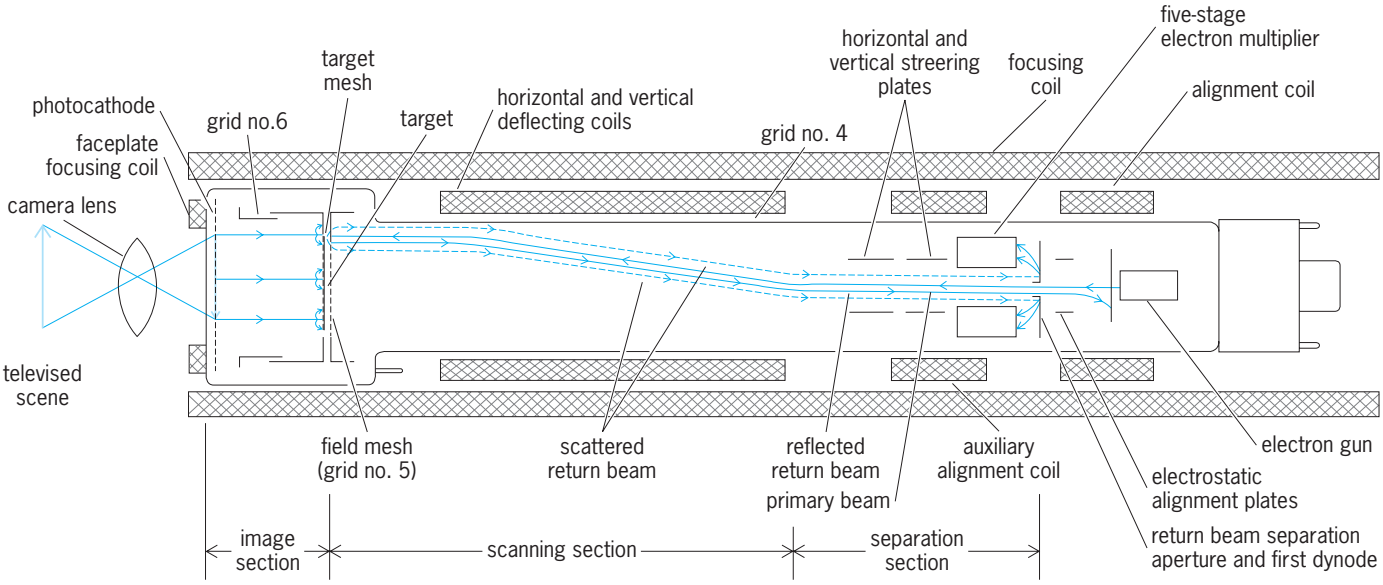


Fig. 4. Image isocon and 175 associated deflectron and focusing components. The separation section isolates the scattered and reflected return beam components. (*RCA Corp.*)

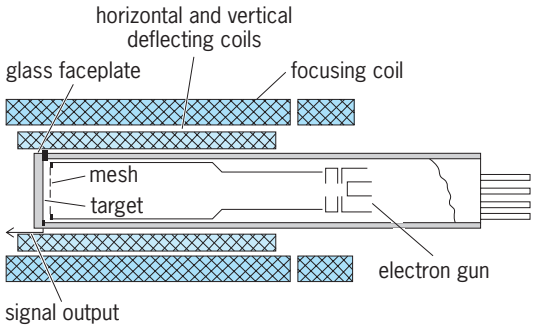


Fig. 5. Cross section of a vidicon tube and its associated deflection and focusing coils.

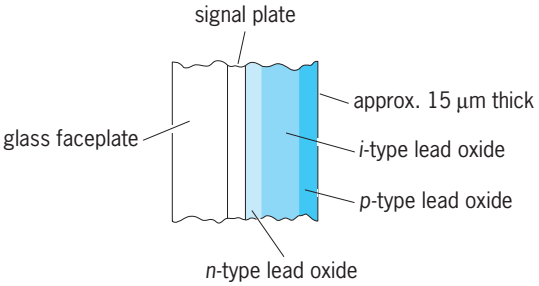


Fig. 6. Cross-sectional view of lead oxide barrier-layer type of photoconductor.

latest photoconductors are the lead oxide, selenium-arsenic-tellurium, cadmium selenide, zinc-cadmium telluride, and silicon diode arrays. All of these either improve the sensitivity or the speed of response (ability to capture motion without "smearing") or both. The first two are barrier-layer types that operate like large-area reversed-bias junctions. **Figure 6** illustrates the configuration of the barrier-layer or reversed-bias junction photoconductors; *n* refers to good electron conductivity, *i* refers to good electron and hole conductivity, and *p* refers to good hole conductivity. A positive voltage is applied to the *n* side through the signal plate, and a negative voltage is applied to the *p* side by the scanning beam electrons. This reverse-biases the junctions between the *i* and *p* and the *i* and *n* sections. This produces low dark current in the absence of light, a very desirable characteristic. Light is absorbed in the bulk of the material, where it produces charge carriers that provide a positive charge image on the side opposite the faceplate. The silicon diode tube is also a reversed-biased junction type, but consists of an array of hundreds of thousands of individual diodes on a wafer of silicon (**Fig. 7**). The photocarriers are generated in the silicon wafer, but are collected and stored on the diode cells (**Fig. 8**). *See* SEMICONDUCTOR.

Most color television cameras utilize the lead oxide or selenium-arsenic-tellurium tubes. Industrial and scientific-industrial cameras utilize the other types.

Silicon intensifier. The silicon intensifier camera tube utilizes a silicon diode target, but bombards it with a focused image of high-velocity electrons. These electrons are emitted by a photoemitter on the inside of the window on the front of the image section (**Fig. 9**). A fiber-optics window is utilized so that the emitting surface can be curved to produce good uniformity of focus of the high-energy electrons on the silicon diode target. Each high-energy electron can free thousands of electron carriers in the silicon wafer (compared to one carrier per photon of light on a silicon diode vidicon). This high amplification allows the camera to operate at light levels below that of the dark-adapted eye. With such a camera tube, it is possible to "see" the individual photons of light that compose a low-level optical image. The silicon intensifier tube is utilized for nighttime surveillance and other extremely low-light-level television uses in industrial, scientific, and military applications. It can be operated over a very wide range of light levels by varying the image section voltage. This changes the amplification over a range of more than 1000 to 1.

Solid-state imagers. These are solid-state devices in which the optical image is projected onto a large-scale integrated-circuit device which detects the light image and develops a television picture signal. Typical of these is the charge-coupled-device imager. The term charge-coupled device refers to the action of the device which detects, stores, and then reads out an accumulated electrical charge representing the light on each portion of the image. The charge-coupled device transfers the individual charges to the

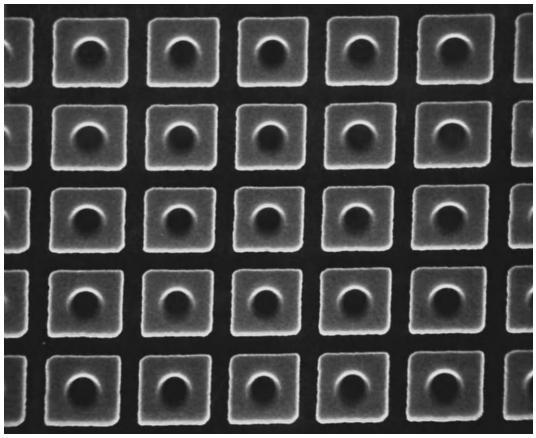


Fig. 7. Small section of the scanned side of a silicon diode target. Diodes are circles, and beam landing pads are squares. (*RCA Corp.*)

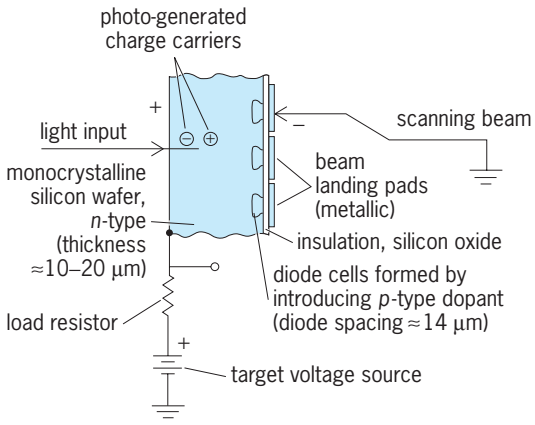


Fig. 8. Cross section of silicon diode target.

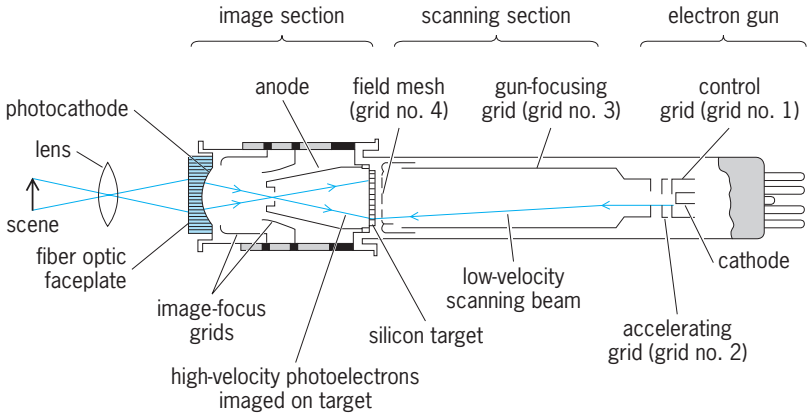
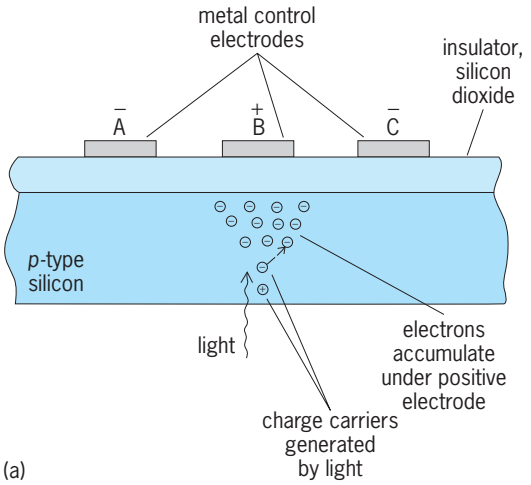


Fig. 9. Silicon intensifier vidicon. (*RCA Corp.*)

output in the proper television scanning sequence to constitute a television video signal. The device detects light by absorbing it in a photoconductive substrate, such as silicon. The charge carriers generated by the light are accumulated in isolated wells on the surface of the silicon that are formed by voltages applied to an array of electrodes on top of an oxide insulator formed on the surface of the silicon. These wells are actually small MOS (metal-oxide-semiconductor) capacitors (**Fig. 10***a*). Charges are transferred through the structure by varying the voltages on the metal electrodes. For example, if electrodes A and B are made more negative and C is made positive, the charges will move laterally from point 1 to point 2 (Fig. 10*b*). *See* INTEGRATED CIRCUITS.



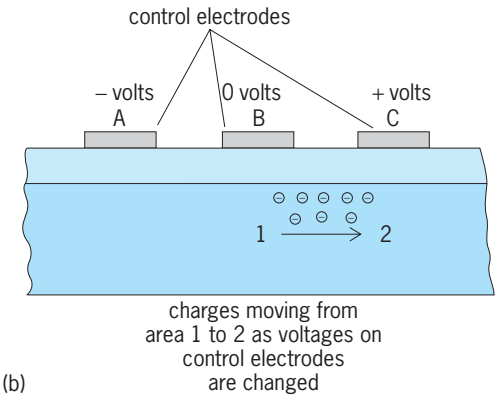


Fig. 10. Charge-coupled device. (*a*) Accumulation of an electron charge in a pixel element. (*b*) Movement of accumulated charge through the silicon by changing the voltages on the electrodes A, B, and C.

A practical charge-coupled-device imager employing these principles consists of a structure that forms several hundred thousand individual wells or pixels, and transfers the charges accumulated out to an output amplifier in the proper sequence.

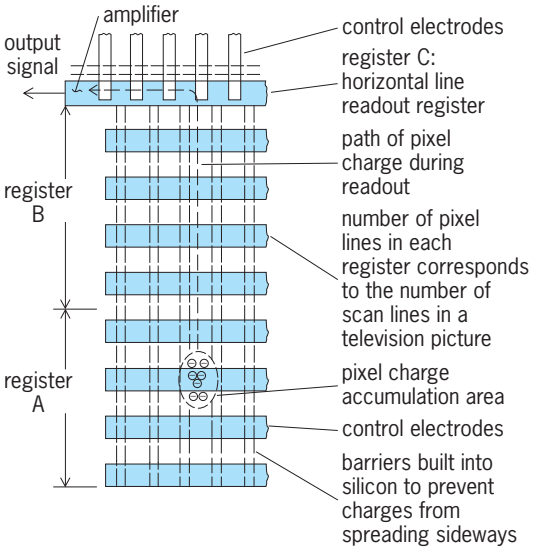


Fig. 11. One type of charge-coupled-device imager. Register A accumulates the pixel charges produced by photoconductivity generated by the light image. The B register stores the lines of pixel charges and transfers each line in turn into register C. Register C reads out the charges laterally as shown into the amplifier.

An example is shown in **Fig. 11**. Here the charges are accumulated by light exposure for the time it takes to complete a single television picture, or approximately $^1/_{60}$ s. Then all of the charges are rapidly transferred line by line upward into the storage register. In the storage area all charges are then moved upward one scan line at the end of each television line interval. The upper line of charges is moved into the horizontal readout register. Then the pixel charges are moved to the left through the horizontal readout register to the output amplifier. During this readout period of the lines in the storage area, a new group of image charges are being accumulated in the imaging area, and the complete sequence is repeated every $^1/_{60}$ s.

Investigation has been undertaken of many versions of solid-state imagers using different light-sensitive materials and charge storage and readout methods.

A charge-coupled-device imager has been incorporated in an image intensifier tube. In this device the charge carriers are generated by a focused image of high-energy electrons in the charge-coupled-device imager in the same manner as in the silicon intensifier tube. This greatly enhances the sensitivity of

the device and allows it to operate at very low light levels. *See* LIGHT AMPLIFIER. Robert G. Neuhauser

Bibliography. K. B. Benson and J. Whitaker, *Television Engineering Handbook: Featuring HDTV Systems*, rev. ed., 1992; K. Jackson and B. Townsend, *Television and Video Engineer's Reference Book*, 1991; B. Kazan (ed.), *Advances in Image Pickup and Display*, 6 vols., 1974–1983.

Television networks

Arrangements of communications channels, suitable for transmission of video and accompanying audio signals, which link together groups of television broadcasting stations or closed-circuit television users in different cities so that programs originating at one point can be fed simultaneously to all others.

In the United States, television network service is furnished by the long-distance or local-exchange carriers (hereafter identified as telephone companies) and satellite carriers, as well as by broadcaster-owned networks. The facilities, when provided by the telephone companies, consist of intercity channels, which interconnect the principal long-distance telephone offices in various cities, and local channels, which connect the telephone offices with the broadcasters' studios or other user locations in each city. In the terminating offices, the intercity and local channels are brought together in a television operating center (TOC) where means are provided for testing, monitoring, and connecting the channels in various patterns as required for service. *See* TELEPHONE SERVICE.

Use. The principal users of television network facilities are broadcast network organizations (broadcast networks). These may be national, regional, or local in scope, and may be regular or occasional in nature and may be commercial or noncommercial. In the United States, special regulations apply to the practices of commercial broadcast networks which offer an interconnected program service on a regular basis for 15 or more hours per week to at least 25 affiliated television licensees in 10 or more states. Four major broadcast networks currently meet this definition: ABC, CBS, Fox, and NBC. In addition, there are at least two other emerging national broadcast networks, and two national broadcast networks in the Spanish language. These regular national broadcast networks provide programming to affiliates on a daily basis at prearranged times. There are also occasional and specialized broadcast networks that provide coverage of local and regional news and of sports events, and meet other programming needs. Finally, there is PBS, a national noncommercial broadcast network that supplies programming to stations specifically allocated for educational purposes.

Broadcast networks typically consist of a programming entity which simultaneously feeds content over the network interconnection facilities to commonly owned and operated (O&O) stations, as well to other stations (affiliates) that have a contractual relationship (affiliation agreement) with the broadcast network. Commercial broadcast networks typically pay compensation to affiliates for carriage of the broadcast network's programming, which includes commercial advertisements (spots) by the network. This benefits the broadcast networks by spreading the cost of programming for their O&O stations over a much larger base, as well as generating revenue through the sale of advertising on a national scale; it benefits the affiliates by reducing program acquisition costs, providing a source of direct revenue, and increasing the value of spots retained by the affiliate within and adjacent to the broadcast network's programming.

In the United States, the broadcast networks have long-term arrangements to provide their primary feed to affiliates. They typically transmit their main programming to O&O stations and affiliates over two separate sets of communications channels. One feed is provided to those stations in the Eastern and Central Time Zones; a separate feed is provided to stations in the Pacific Time Zone to eliminate the need to tape-delay programming, and to permit programs to be seen at the same clock time by populations in the east and west. Some Mountain Time Zone stations are fed with the Eastern feed, others with the Pacific feed. In addition to these primary feeds, the broadcast networks provide a separate feed of programming in a digital format to be used by O&O stations and affiliates in the operation of their digital television (DTV) stations. These can be in either the high definition (HDTV) or standard definition (SDTV) formats.

In addition to their regular interconnection channels, the broadcast networks frequently order "occasional" channels to pick up programming from other than the usual originating points. An example is the origination of sports and special events programming from remote points, as well as the reception of material from diverse points to be included in news programming.

History. The first intercity television transmission in the United States occurred in 1927. Coaxial cable was installed in 1936 to permit the transmission of television images and sound between New York and Philadelphia. By 1940, coaxial cable was being used to provide programming for commercial broadcast.

During this period, United States broadcasters supported a private radio relay system linking television stations, while the telephone companies proposed use of coaxial cables and microwave links. This was resolved when federal regulators supported the common carrier approach favored by the telephone companies.

Following World War II, the telephone companies in the United States began to provide facilities in many major cities for television interconnection. Most major eastern cities were linked by the end of 1947. In 1949, a circuit linked the east with Chicago, first shared by all networks, and subsequently with each obtaining a discrete circuit. By August 1951, the transcontinental line was completed, enabling the broadcast networks to broadcast simultaneously from coast to coast.

The national grid grew explosively during the 1950s, so that by the end of the decade every significant community in the continental United States was capable of receiving a network feed. Stations in Hawaii and Alaska, however, continued to rely on "bicycling" of film, kineoscope transcriptions, and video tape from west coast stations until the late 1970s.

As time passed, the use of coaxial cable for intercity transmissions greatly diminished, replaced by a grid of microwave facilities that relay television and other communications transmissions. By the mid-1970s, communications satellites began to be used more and more for occasional channels to "backfeed" programming to the television networks, and the broadcast networks subsequently began to use communications satellites for the primary program feed to their O&O stations and affiliates. Now fiber-optic rings are being used for local television interconnection, and fiber-optic is being used for some long-haul video traffic.

Transmission requirements. United States television networks transmit color signals employing the National Television Systems Committee (NTSC) standards of monochrome information similar to that of an ordinary black-and-white picture in the frequency band from 30 Hz to about 4 MHz, plus color information transmitted by a color carrier at approximately 3.6 MHz which is both amplitude- and phase-modulated and whose upper sideband extends to 4.2 MHz. Two additional frequency-modulated (FM) subcarriers at 5.8 and 6.4 MHz are used to provide the simultaneous transmission of the audio portion of the signal over terrestrial systems. When transmission is over satellite, the subcarriers usually are at 6.2 and 6.8 MHz. Networks in other countries employ the PAL (phase alternation by line) or SECAM (sequential color with memory) standards. Some network have begun to transmit signals intended for broadcast by digital stations under the Advanced Television Standards Committee (ATSC) standards. *See* FREQUENCY MODULATION; TELEVISION STANDARDS.

Satisfactory transmission of these signals requires that the video channels pass the broad range of frequencies from 30 Hz to 6.5 MHz with negligible distortion. Performance standards (NTC7) were set by the National Transmission Committee, a joint effort of the major United States broadcast networks and the telephone companies. The channels used are designed to have attenuation and delay characteristics which are essentially flat over this range, and to produce a signal-to-noise ratio of 50 dB or better (peak power signal to root-mean-square noise) from pickup to ultimate user over a distance of up to 4000 mi (6500 km) or over a satellite transmission link. Differential distortion—that is, nonlinearity of output compared with input over the whole range of levels to be transmitted—must also be kept to a minimum. Otherwise, the color carrier could vary in amplitude or phase as the monochrome signal changes from black to white. *See* DISTORTION (ELECTRONIC CIRCUITS).

Interexchange carrier channels. The facilities for intercity video channels include terrestrial microwave networks, synchronous satellite systems, and fiber-optic systems. The terrestrial microwave networks are FM radio systems, usually in the frequency band of 3.7–4.2 GHz, with some links in the 6-GHz and 11-GHz bands and higher. Depending on service requirements and design, these microwave systems provide up to 12 wide-band channels in each direction. Relay stations are spaced an average of 25–30 mi (40–50 km) apart. Automatic switching is provided at section terminals that transfer service to a protection channel in case of equipment failure or severe fading on a regular channel. Satellite systems are used to provide both domestic intercity transmission and transmission to and from points located overseas. Fiber-optic links are used for local loops and for point-to-point transmissions over long distances. *See* COMMUNICATIONS SATELLITE; MICROWAVE.

Local channels. The local channels, which connect the telephone company office with the broadcasters' studios and transmitters and with closed-circuit users' locations, are largely composed of cable and fiber-optic facilities. To minimize low-frequency interference, cable channels make use of special balanced video cable pairs, generally 16-gage wire insulated with a low-loss dielectric. Each video pair is shielded by copper tapes or woven copper mesh. The attenuation of a 16-gage video pair is about 19 dB/mi (11 dB/km) at 4.5 MHz. Video amplifiers are spaced at intervals up to 4.5 mi (7.2 km) to compensate for attenuation. *See* COMMUNICATIONS CABLE.

Use of fiber-optic systems for local video distribution is growing in popularity, as fiber is a very effective means of point-to-point communications, but it is not yet economically viable for point-to-multipoint communications. The fiber-optic transmissions can be either FM or digitally modulated. The use of FM overcomes signal-to-noise-ratio and intermodulation distortion, removing the problems of signal quality from the optical to the electrical domain. The video baseband, with multiplexed FM subcarriers containing audio and data signals, is frequency-modulated on a carrier to produce a signal with a fixed amplitude excursion, the value of which modulates the fiber's light source over a linear portion of its range. Currently, fiber-optic systems using digital modulation often use dense wavelength division multiplexing (DWDM), which permits multiple independent signals to be transmitted over a single fiber by using different optical wavelengths. *See* MODULATION; OPTICAL COMMUNICATIONS; OPTICAL FIBERS; PULSE MODULATION.

Local channels may also be provided by fixed or portable microwave radio facilities. These can use either FM or digital modulation. Economic considerations usually determine whether radio, fiber, or cable should be used, but the decision may be influenced by the need for a speedy installation, such as for one-time news or sports events.

Television network control centers. In network operations, frequent rearrangement of the channel connections is required to change the point of origin

or to feed successive programs to different groups of broadcasting stations. These rearrangements are accomplished by means of the switching systems in network operating centers. Switches may be made either on a time schedule or on voice cue. Scheduled switching, when ordered, is done between the various 30-min segments into which the broadcasters divide their daily programming. Switching on cue is required for programs, such as athletic events, with indefinite completion times, or to rearrange a network in midprogram so that different commercials can be delivered to various geographical regions.

Although most television programs originate in the studios of the broadcasting companies, many originate at other locations, such as sports stadiums. Currently, most "backhaul" feeds from fixed locations hosting recurrent televised events, such as sports stadiums, are transmitted by fiber-optic facilities. Local feeds from other remote locations use microwave relay, and feeds to national program services are typically relayed directly by satellite to the network's main control center. From the control center the program is fed to the permanent network.

Fiber-optic distribution. There is also a growing use of fiber-optic distribution of television signals, particularly in backhaul and other point-to-point transmissions. Signal deterioration due to transmission-medium attenuation and noise addition is completely removed by regeneration, the process of detecting the digital signal, performing error correction, and retransmitting the signal to the next regeneration station. In this way, error-free transmission to any distant location is possible.

Fixed satellite networks. Transmission by satellite is widely used by broadcast networks, by cable television distributors, and in closed-circuit applications. The broad-frequency band-pass of the satellite transponder and the fact that the satellite transmitting antenna can be designed to provide transmission to a broad geographic area make satellites an attractive tool for providing point-to-multipoint video services. Satellite television service is configured so that one of several Earth stations sends the program (uplinks the program) to a satellite transponder. The signal is amplified, translated (converted from the uplink to the downlink frequency), and transmitted back to Earth. The satellite transmitting antenna pattern can be constructed so as to cover a large geographical area such as the contiguous United States, and on some satellites certain transponders serve offshore points such as Alaska, Hawaii, and Puerto Rico. It is this point-to-multipoint mode (that is, one station transmitting to many receiving stations) of satellite transmission that makes satellite distribution so attractive to program distributors. In a typical television network application, a master network transmitting Earth station can control many receiving Earth stations. Control may be extended to the receiving locations by a terrestrial data line or over the satellite on a low-power subcarrier along with the television signal. Regionalization of a satellite network, such as for football or baseball telecasting, where different games are to be seen in different cities, requires separate transponders for each event and close control of each receiving location.

Use of satellites for news backhaul. The ability to uplink to satellites from virtually anywhere in the world with portable equipment has made news reporting almost instantaneous. News backhaul usually uses Ku-band satellites, with uplink transmissions to the satellite at 14–14.5 GHz, and downlinks to receive stations at 11.7–12.2 GHz. The major portion of news backhaul signals are delivered to the editing point by frequency-modulation of the video, multiplexed with FM subcarriers of the related audio. This occupies about 27 MHz of bandwidth per program. Digital video compression and transmission enables more news backhaul to be completed with portable equipment, using lower satellite power and reduced transponder bandwidth.

Direct broadcast satellites. Geostationary satellite antenna footprints can be made to cover large areas of the Earth, thus making them ideal for point-to-multipoint transmissions. Direct broadcast satellite systems have been placed into operation in many developed nations. Their implementation in the United States is typical of other systems.

After the cable television industry in the United States began in the late 1970s to use the relatively low-powered (10–17 watts per channel) C-band fixed service satellites to expand the availability of program sources, individuals without access to cable saw satellite reception as a means of expanding their own programming options. The television receive-only (TVRO) industry began to provide consumers with receive antennas, typically using parabolic dishes of 10–12-ft (3–4-m) diameter, and associated electronic equipment. Program suppliers responded to this by using encoders to scramble their signals, permitting only authorized users access to decoding devices. The U.S. Congress subsequently passed legislation that required program suppliers to give consumers access to scrambled satellite programming, and integrated receiver decoders (IRDs) were marketed that permitted consumers to subscribe to and receive satellite programming services.

The Federal Communications Commission (FCC) allocated a portion of the Ku band, from 17.3 to 17.8 GHz, and downlink frequencies from 12.2 to 12.7 GHz to a new Broadcast Satellite Service in 1982. Space stations in that service are known as direct broadcast satellites (DBS). These satellites typically have 32 transponders of 120 W output each, although they can be reconfigured to operate with a lesser number of higher-power, 240-W transponders. These satellites receive and retransmit digital signals, and the amount of transponder power output has a direct bearing on the bit rate that can be handled by the satellite. *See* DIRECT BROADCASTING SATELLITE SYSTEMS.

In order to be competitive with terrestrial cable systems, the DBS system must provide consumers with far more than 32 program services. Therefore, the DBS providers have chosen to operate digital

systems using significant amounts of compression. This permits the 32 transponders to carry approximately 200 separate program services, often with a picture quality surpassing terrestrial television. To do this, the systems make use of the MPEG2 compression standard for video transmission, and Layer II of the MPEG1 compression standard for audio. The compression technology permits adaptive use of the available spectrum. Thus, programming containing little motion or change from scene to scene is more highly compressed than material requiring greater detail. This permits spectrum and transponder power conservation while preserving picture quality. *See* DATA COMPRESSION.

There are two principal DBS licensees operating in the United States: DirecTV and EchoStar, which does business as the Dish Network. The DirecTV operations center is in Castle Rock, Colorado, and the EchoStar operating center is in Cheyenne, Wyoming.

Earth stations at the operations centers receive programming materials from multiple sources. In addition, the centers contain banks of program origination equipment so that pay-per-view movies and other programming can be originated directly from the center. The source materials are input to video and audio compressors. The outputs, in the form of packetized elementary streams (PES), are input to a multiplexer. Several separate services are transformed into a single bitstream for each transponder. Forward error correction (FEC) is then applied to the bitstream from the multiplexer, and then input into a quadrature phase-shift keying (QPSK) modulator. The bitstream from the modulator is then frequency-upconverted to the appropriate transponder frequency and transmitted to the satellite. *See* INFORMATION THEORY.

The satellite transponds the signal and retransmits it to the Earth, where it is received by a small receive dish, typically an 18-in. (50-cm) reflector antenna. The antenna utilizes an offset low-noise block (LNB) amplifier. The low-noise block downconverts the signal to the L band (950–1450 MHz) and sends it via coaxial cable to the settop integrated receiver decoder. The integrated receiver decoder selects the appropriate transponder, demodulates the QPSK waveform, removes the forward error correction, demodulates it to extract the selected service from the bitstream, stores the signal in a buffer stage for further processing, and then decompresses the audio and video for reception by the television receiver.

During its initial years of operation, the DBS operators provided primarily nonbroadcast programming services to their subscribers. Subscribers wishing to receive local broadcast signals either had to install off-air antennas or subscribe to cable service in addition to their satellite subscriptions. Since the passage of the Satellite Home Viewer Improvement Act (SHVIA) in 1999, satellite television companies have begun to offer local programming to many areas throughout the United States, and as new satellites with spot-beam technology replace those currently in orbit, greater portions of the United States will be able to receive local signals by DBS.

Cable distribution by satellite. The C band (5925–6425 GHz up, and 3.7–4.2 GHz down) is used for cable television–satellite television networking. Cable television systems receive programming by satellite, and this programming is then sold to subscribers. Because of the premium value of these signals, they are frequently scrambled so they can be viewed only through appropriate equipment.

While all satellite television networks once used FM transmission, many operators now use digital video compression to transmit 4–10 signals over a single satellite transponder using quadrature phase-shift keying (QPSK) modulation. This changes the instantaneous amplitude of the two orthogonal carriers (I and Q) at a rate of about 38 Mbps. The received data require good forward error correction for perfect decompression to video and audio. Picture quality at four to six compressed video programs per satellite transponder is as good, or better, than FM-transmitted signals.

Cable delivery methods. Over 96% of television households in the United States have access to cable television, and most of them receive all their television programming in analog format via cable. Programming usually includes the local broadcast stations and satellite services.

Because of the closed nature of cable television systems, all the frequencies between 50 and 750 MHz and higher can be used for distribution. The analog signals received at cable television headends are combined side by side on a frequency-division multiplex (FDM) basis for carriage in the cable systems. Digital satellite signals are decompressed in integrated receiver decoders, and the programs are passed through to the subscribers as analog signals (**Fig. 1**). *See* MULTIPLEXING AND MULTIPLE ACCESS.

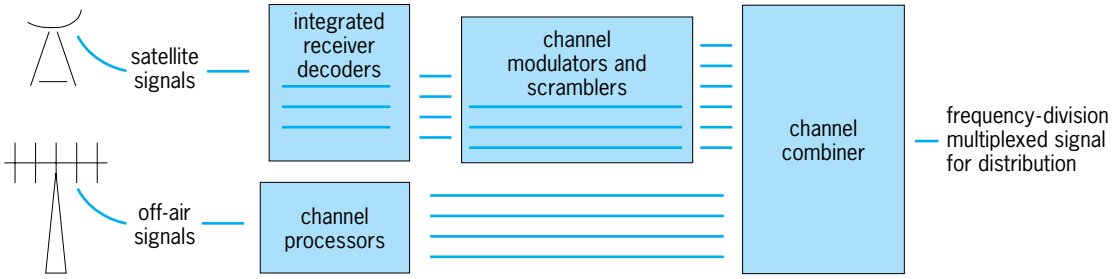


Fig. 1. Typical cable television headend. The processing used in creating the frequency-division-multiplex signal is shown. (*General Instrument Corp.*)

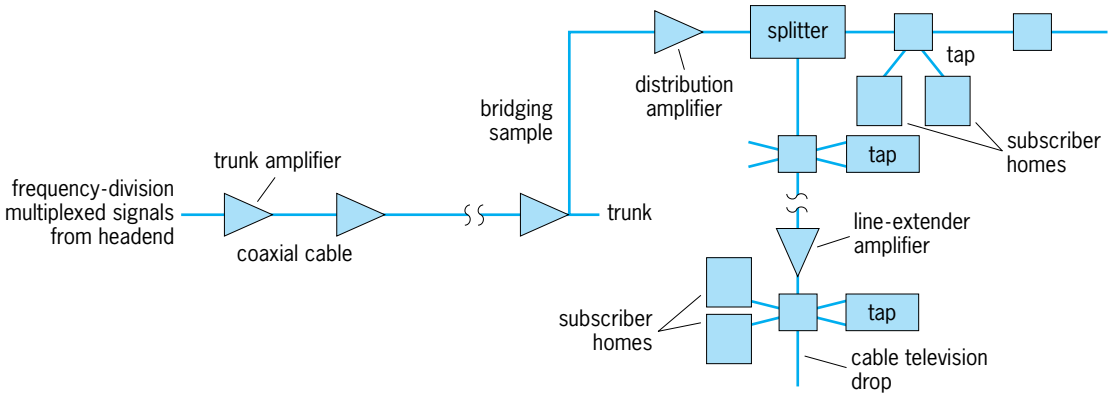


Fig. 2. Coaxial cable television system. (*General Instrument Corp.*)

Since many satellite-delivered programs are premium services, scrambling is applied to the cable signals and descramblers are leased to subscribers. Those descramblers often are necessary for distoration-free viewing of the available channels on the cable system.

Coaxial cable systems. Cable television systems used to have very long cascades of amplifiers, interconnected by coaxial cable. The loss of a coaxial cable increases in proportion to the square root of the frequency. Thus, the loss doubles when the frequency quadruples. For example, the loss at 200 MHz is twice that at 50 MHz, and the loss at 800 MHz is twice that at 200 MHz. *See* COAXIAL CABLE.

Amplification at higher frequencies is also more difficult to accomplish. When large numbers of amplifiers are cascaded, the flatness of the frequency response is critical because individual amplifier flatness deviations add to each other. Older cable television systems often had amplifier cascades of 40 or more. When an amplifier operates in a cable television system, the amplitude at which the signals can be carried is limited by the amplifier output capability, the noise contributed by the amplifier, and the number of television channels being carried. *See* AMPLIFIER.

A number of methods of improving the reach and quality of cable television signals are employed. One technique is to build the cable television facility in two parts. The trunk, with very large diameter cable to minimize loss, carries the signals most of the distance. The distribution system, built in parallel, bridges some of the trunk signal and, using less costly smaller-diameter cable, feeds the signals along the streets to the homes.

The amplifiers used in the trunk are very linear and often share the same housing as the bridger amplifier. Line-extender amplifiers are used when the taps feeding individual homes reduce the signal level to the point of requiring amplification. Distribution amplifiers generally have high power, lower flatness performance, and lower cost (**Fig. 2**).

In addition to larger coaxial cable, supertrunks are sometimes used in which FM signals are carried in a frequency-division-multiplex frequency plan and demodulated in a subheadend. This technique is expensive, but less than creating a second headend.

These techniques are sometimes supplemented by the use of amplitude-modulated links. Here the frequency-division multiplex cable television signals are sent by microwave radio to remote receivers acting as remote headends for normal cable television distribution. These systems are also beneficial when rivers and lakes block the routing of coaxial cable. They have problems of cost, licensing procedures, and rain fade, since they operate at frequencies near the Ku band.

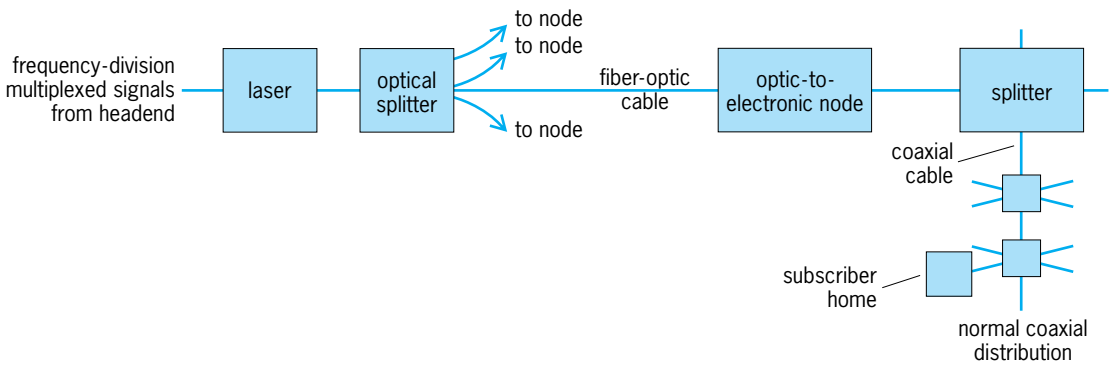


Fig. 3. Hybrid fiber-coaxial television system. (*General Instrument Corp.*)

Fiber optics in cable television. Most modern cable systems use fiber optics to replace the coaxial trunk cable completely, while imposing only the noise and distortion equal to one trunk amplifier. The wide use of fiber optics by cable systems contrasts with broadcast networks. The small optical-fiber light path requires lasers generating coherent light to launch significant light into the fiber. The modulation of these lasers is linear in that television signals are modulated by using frequency-division multiplexing onto a laser and then into the fiber-optic cable. *See* LASER.

The fiber-optic cables use two of the lowest optical frequencies capable of supporting low-loss single-mode transmission. Losses increase as the optical frequency is increased, just as happens in coaxial cable. The two widely used wavelengths are around 1310 and 1550 nanometers. Between them is a quite lossy water absorption line.

The optical-fiber cable used in cable television systems is either step-index or graded-index, both with almost equal performance. Both fibers have a loss of about 0.56 dB per mile (0.35 dB per kilometer) at 1310 nm, and 0.35 dB per mile (0.22 dB per kilometer) at 1550 nm. This is very low compared to 1 dB per 100 ft (33 dB per kilometer) for 1-in.-diameter coaxial cable at 750 MHz.

Distributed-feedback lasers operating at 1310 nm are generally used, with power outputs up to about 12 milliwatts. This power is seldom used for a single fiber-optic run, but is split to feed a number of cable television nodes. Each node is a local distribution system delivering cable television to 500–2000 subscribers. Usually, the area is limited by the number of customers or a maximum of four coaxial cable amplifiers in cascade. The desire is to design cable television distribution systems that will permit a reasonable number of customers to be served with all the planned services, including telephone, television, interactive services, multimedia, and two-way data services.

Hybrid fiber-coaxial (HFC) cable television systems require about 1 mW of received optical power in order to maintain a good carrier-to-noise ratio at the node where the optical power is converted to electronic power. For 1 mW out, a higher power input is required. For a 6.25-mi (10-km) span from the headend to the distribution node, about 4 dB of optical loss is incurred. This means that the power at the input must be 4 dB above 1 mW or 2.5 mW, permitting dividing the original 12 mW to supply four such paths (**Fig. 3**). *See* CABLE TELEVISION SYSTEM; TELEVISION.

Joseph B. Glaab; Clifford M. Harrington

Bibliography. W. Cicoria et al., *Modern Cable Television Technology*: *Video, Voice, and Data Communications*, Morgan Kaufman Publishers, New York, 2d ed., 2004; D. Roddy, *Satellite Communications*, 4th ed., McGraw-Hill Telecom Engineering, New York, 2006; J. Whitaker and B. Benson, *Standard Handbook of Video and Television Engineering*, 4th ed., McGraw-Hill, New York, 2003; E. Williams et al., *National Association of Broadcasters Engineering Handbook*, 10th ed., Elsevier, 2007.

Television receiver

The equipment used to receive the transmitted modulated radio-frequency signals and produce synchronized visual images and sound for entertainment or educational purposes. The radio-frequency portion operates on the superheterodyne principle. *See* MODULATION; RADIO RECEIVER.

The first television receivers to be mass-produced were monochrome; that is, they provided pictures in black and white only. Later, color receivers, which produce pictures in full color as well as black and white, became available. Some television receivers now can receive stereophonic sound or alternate language in accordance with multichannel television sound standards. For basic discussion of a television system. *See* TELEVISION; TELEVISION STANDARDS.

Early television receivers used vacuum-tube technology. Present-day receivers use solid-state technology with many functions integrated on a few chips. The only function still primarily implemented by using vacuum-tube technology is the display by the cathode-ray tube (CRT). *See* CATHODE-RAY TUBE; INTEGRATED CIRCUITS; VACUUM TUBE.

Monochrome receivers. Figure 1 shows a block diagram of a conventional monochrome television receiver.

Antenna and transmission line. Since most broadcast television transmissions in the United States are horizontally polarized (some are circularly polarized), the most basic type of television-receiving antenna is the horizontally mounted half-wave dipole. Because the stations serving a given area may operate on widely different frequencies, however, the dipole dimensions must be a compromise that permits reasonable performance on all the desired channels. More complex antennas combine several dipole elements of various lengths, and passive reflectors may be used to achieve some degree of horizontal directivity, which increases the amplitude of the receiver signal and reduces interference from other stations. Highly directive antennas are frequently mounted on remotely controlled rotators so that they can be pointed in the direction providing the best reception of the desired signal. The most common types of transmission line between the antenna and receiver are 300-ohm "twin-lead," employing polyethylene as a dielectric spacer between two uniformly spaced, unshielded wires. Also 75-ohm coaxial cable is used. *See* ANTENNA (ELECTROMAGNETISM); COAXIAL CABLE; POLARIZATION OF WAVES; TRANSMISSION LINES; YAGI-UDA ANTENNA.

Tuner. The tuner of a television receiver selects the desired channel and converts the frequencies received to lower frequencies within the passband of the intermediate-frequency amplifier. For very high-frequency (VHF) reception older tuners have 12 discrete switch positions, corresponding to channels 2–13. For ultrahigh-frequency (UHF) reception, continuous tuning is employed in older tuners. Most receivers now use a frequency synthesizer circuit for generating the local oscillator frequency, and may be stabilized with an automatic frequency control.

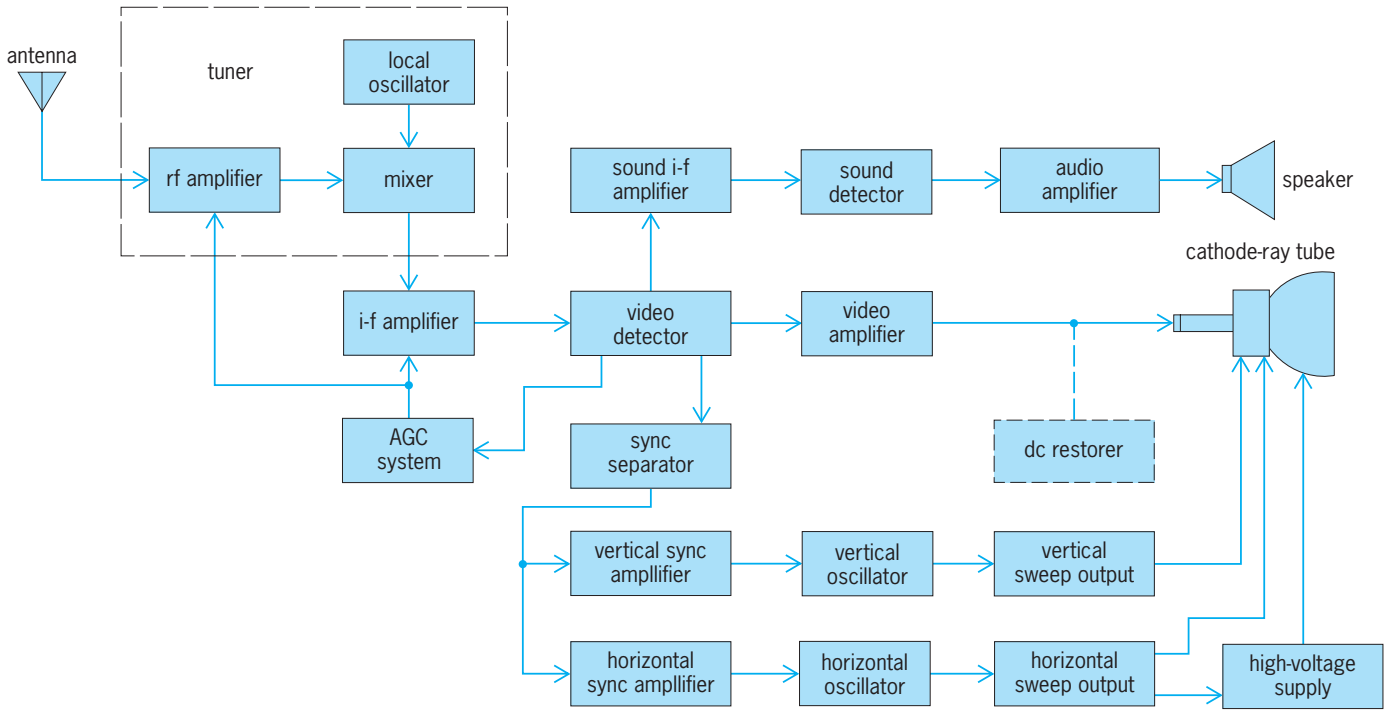


Fig. 1. Block diagram of a typical monochrome television receiver.

Nearly all VHF tuners employ a radio-frequency (rf) amplifier, a mixer, and local-oscillator circuits arranged as shown in Fig. 1. Some receivers are provided with additional channel-tuning capability for use with a community antenna television (CATV) system. *See* AMPLIFIER; AUTOMATIC FREQUENCY CONTROL (AFC); CLOSED-CIRCUIT TELEVISION; OSCILLATOR; RADIO-FREQUENCY AMPLIFIER.

The received signal and the local oscillator signal are applied to the mixer. Difference frequencies, representing the picture and sound carriers, are produced and remain essentially constant as the radio-frequency amplifier, mixer, and oscillator circuits are tuned to the different channels. Known as intermediate frequencies (41.25 MHz for sound and 45.75 MHz for picture), they are available for further amplification.

Such performance characteristics as noise factor, gain, bandwidth, and oscillator radiation must be optimized in the design of the tuner.

Intermediate-frequency amplifier. The output from the tuner is applied to the intermediate-frequency (i-f) amplifier. The gain of this amplifier is essentially constant from 43 to 45 MHz. Above 45 MHz the response decreases such that at 45.75 MHz, the picture carrier frequency, it is 50%. This slope is required to compensate for the vestigial sideband transmitted signal.

Below 43 MHz the response decreases until at 41.25 MHz, the sound carrier frequency, it is 5–10% of the flat response. This minimizes cross modulation between picture and sound carriers. Fixed tuned trap circuits are used to produce sharp cutoffs at the lower and upper limits of the intermediate-frequency passband. Sufficient selectivity is provided to minimize interference from signals originating in adjacent television channels.

Separation of video and audio. The output of the intermediate-frequency amplifier consists of two modulated radio-frequency signals. One of these, which is amplitude-modulated, provides a varying signal corresponding to the black and white portions of the picture, a blanking signal to render the return trace invisible on the picture tube, horizontal sync pulses to initiate the retrace of the beam at the end of each line, and vertical sync pulses to initiate the retrace of the beam at the end of each picture field. The other signal is frequency-modulated and contains the transmitted sound information. *See* AMPLITUDE MODULATION; FREQUENCY MODULATION.

There are two types of detector which may be used for video modulation detection: the envelope detector and the synchronous detector. In the envelope detector the two radio-frequency signals are applied to a diode which produces a rectified output that follows the instantaneous peak value of the amplitude-modulated picture carrier. *See* AMPLITUDE-MODULATION DETECTOR; DIODE.

The synchronous detector is a highly linear device and is used in the better television receivers where high-quality results are desired for color, teletext, alphanumeric and graphical data digital signals, and multichannel sound. The band-pass filter characteristic has much less attenuation for the sound carrier and upper video-frequency components. The synchronous detector is basically an analog multiplier.

The polarity of the video detector output depends upon the design of the video amplifier and method of picture-tube drive. Usually, maximum picture carrier (sync-pulse modulation) produces a negative output voltage.

Coincidentally a 4.5-MHz signal results from the

heterodyne beat of the picture and sound carriers. This signal contains the frequency-modulated sound information, which can be further amplified and detected in the sound channel. This is known as the intercarrier sound (ICS) system.

Following the detector is a video amplifier. An output level of about 100 volts is ordinarily sufficient to assure full drive of the picture tube over its modulation range. A 4.5-MHz trap is included in the video amplifier circuit to prevent the appearance of the intercarrier sound signal on the picture tube. The sound reproduction system is discussed below. *See* VIDEO AMPLIFIER.

Automatic gain control. Since television receivers, like radio receivers, may be subjected to widely varying incoming signal strengths, some form of automatic gain control (AGC) is necessary. Circuits for this function provide a nearly constant carrier signal level to the video detector. *See* AUTOMATIC GAIN CONTROL (AGC).

Sync separator circuits. Picture synchronizing information is obtained from the video signal by means of sync separation circuits. In addition, these circuits must separate this information from noise and interference during the reception of weak signals, particularly if impulse noise is present. In general, sync separation circuits perform the following functions: (1) separation, by means of amplitude clipping, of the sync information from the picture information; (2) separation of the desired horizontal and vertical timing information by means of frequency selection; and (3) rejection of noise signals that are higher in amplitude than sync pulses by amplitude limiting or gating (noise suicide) circuits. *See* ELECTRICAL NOISE.

Sweep systems. Two independent systems are employed in the vertical and horizontal sweep circuits. Each employs a timing generator, generally of the oscillatory type, controlled by the synchronizing information obtained from the sync separators. The oscillators are followed by drive and waveform-shaping circuits. These are followed by power amplifier stages capable of providing the currents required by the deflection coils of the yoke for picture-tube beam deflection. Substantially different techniques are required for vertical and horizontal scanning.

Vertical deflection. Generally the vertical oscillator is of the relaxation type operating at approximately 60 Hz. Its frequency is accurately controlled by a signal obtained from the sync separator. The output waveform of the sync separator consists of a train of pulses representing the horizontal and vertical synchronizing pulses. When these are passed through a low-pass filter or integrating circuit, a sawtooth-shaped voltage wave representing vertical sync is obtained. This is used to synchronize the vertical oscillator. A frequency control in the vertical oscillator circuit is so adjusted that its free-running frequency is slightly lower than the synchronizing signal frequency. For good interlace it is necessary that no horizontal frequency components be included in the vertical synchronizing voltage. *See* ELECTRIC FILTER; TELEVISION SCANNING.

The vertical output stage is generally operated as a class A amplifier. The yoke is transformer-coupled to the plate of the output tube to match the yoke impedance to the output tube impedance. Since the yoke impedance is partly resistive and partly inductive, the voltage waveform across it is the sum of a sawtooth and a rectangular pulse. The current through the yoke has essentially a sawtooth waveform, but each sawtooth has a symmetrical S shape to take care of picture-tube faceplate geometry and result in a linear scan. *See* POWER AMPLIFIER; TRANSFORMER.

Horizontal deflection. A more complex system is required for horizontal scanning. There are several basic reasons for this: (1) Horizontal sync pulses are of much shorter duration than are vertical sync pulses; (2) some form of automatic frequency control (AFC) of the horizontal oscillator is required to average the incoming horizontal sync information and retain accurate phase; and (3) considerably greater power output is required to generate the deflecting yoke fields as well as the high voltage, of 10–20 kV, for the picture tube. *See* AUTOMATIC FREQUENCY CONTROL (AFC).

The horizontal oscillator is generally of the Hartley or Colpitts type. The frequency of oscillation is determined both by a time-constant control and by a bias voltage derived from an AFC circuit. The AFC circuit may be a phase comparator, in which the pulses from the sync separator are compared to the oscillator output signal. The output of the comparator is a voltage proportional to the phase departure of the two signals. Many receivers lock the horizontal and vertical scanning waveforms together with a countdown circuit, thus eliminating the vertical oscillator. In this case the countdown begins at a harmonic of the horizontal scan rate.

The desired current waveform in the horizontal windings of the deflection yoke is a line-frequency sawtooth, possibly modified by the addition of a small amount of S curvature to compensate for picture-tube face geometry. Energy from the horizontal drive, or output tube, is normally supplied to the yoke (through the horizontal output transformer) only during approximately the last half of each sawtooth period. At the conclusion of the sawtooth period, the horizontal driver is cut off, and the energy stored in the form of current through the yoke causes an oscillation in the self-resonant circuit consisting of the yoke, horizontal output transformer, and the associated capacitances. This oscillation is permitted to continue for only one half-cycle, during which time the current through the yoke reverses in polarity and attains a negative value almost equal to the original positive value. The self-resonant frequency of the horizontal output circuit must be high enough to permit the full current reversal to be accomplished within the horizontal blanking interval. The oscillation is stopped after the first half-cycle by the action of a damper (normally a diode), which

controls the release of the energy stored in the yoke in such a way that the current follows the desired sawtooth waveform. In approximately the middle of the sawtooth period, the damper becomes nonconductive, and the horizontal driver takes over the task of supplying the energy required for the next cycle.

High-voltage supply. Since the impedance of the yoke at horizontal scan frequency is primarily inductive, the voltage across the horizontal deflection windings is essentially constant during active scan. During the retrace period, however, the high rate of current change causes the generation of a high-voltage pulse having a shape similar to that of a half sine wave and a duration equal to the retrace period. It is common practice to employ a stepup winding on the horizontal output transformer to raise this so-called kickback pulse to a still higher voltage level, which is commonly about 18 kV for black-and-white displays, and to pass it through a simple rectifier and filter to serve as the high-voltage supply for the cathode-ray tube.

Picture tubes. The display device for a monochrome television receiver is a cathode-ray tube, consisting of an evacuated bulb containing an electron gun and a phosphor screen, which emits light when excited by an electron beam. The intensity of the electron beam is controlled by the video signal, which is applied either to the grid or to the cathode of the electron gun. The position of the electron beam is controlled by electromagnetic fields produced by the deflection yoke placed around the neck of the tube. *See* PICTURE TUBE.

Controls. Certain controls are available to the user for adjustment of the receiver. These are the audio volume, channel selector, and brightness and contrast controls. In some receivers fine-tuning and horizontal hold and vertical hold controls are also available. Other controls, normally mounted on the rear of the chassis or under a removable panel, may include height, width, and linearity controls.

The on-off switch for the receiver is frequently mounted on the same shaft as the audio volume control, which controls the gain of the audio channel. The channel selector adjusts the tuner's selective circuits for optimum performance at the desired channel, and fine tuning is a vernier control for the frequency of the local oscillator. Brightness is usually a manual adjustment of the bias on the electron gun in the picture tube. The contrast control adjusts the level of the video signal by some means such as a potentiometer in one of the video amplifier stages.

The horizontal and vertical hold controls adjust the free-running frequencies of the horizontal and vertical oscillators to achieve the most reliable synchronization with the incoming signal. In some cases, the controls may actually consist of variable resistors in the circuits of the respective oscillators.

Vertical linearity is generally controlled by a variable resistance in the circuit of the vertical output stage, and picture height may be controlled by a variable resistor in the circuit of the vertical oscillator. Horizontal linearity may be controlled by a variable inductor placed between the damper and the source of voltage.

Color receivers. Television receivers designed to produce images in full color are necessarily more complex than those designed to produce monochrome images only, because additional information must be handled to produce color. In monochrome systems, the video signal controls only the luminance of the various areas of the image. In color systems, it is necessary to control both the luminance and the chrominance of the picture elements.

The chrominance of a color refers to those attributes which cause it to differ from a neutral (white or gray) color of the same luminance. While chrominance can be expressed in a great variety of ways, it is always necessary to employ at least two variables to express the full range of chrominance that can be perceived by the human eye. In qualitative terms, chrominance may be regarded as those properties of a color that control the psychological sensations of hue and saturation. For color television purposes, chrominance is most frequently expressed quantitatively in terms of the amounts of two hypothetical, zero-luminance primary colors (usually designated I and Q), which must be added to or subtracted from a neutral color of a given luminance to produce the color in question. *See* COLOR.

As a practical matter, color television receivers produce full-color images as additive combinations of red, green, and blue primary-color images, and it is necessary to process the luminance and chrominance information in a color signal in such a way as to make it usable by a practical reproducing device.

Color signal. Color television broadcasts in the United States employ signal specifications that are fully compatible with those used for monochrome, making it possible for color programs to be received on monochrome receivers and monochrome programs to be received on color receivers. (Color pictures are produced, of course, only when color programs are viewed through color receivers—in all other cases, the images are in black and white only.) Compatibility is achieved by encoding the color information at the transmitting end of a color television system in such a way that the transmitted signal consists essentially of a normal monochrome signal (conveying luminance information) supplemented by an additional modulated wave conveying chrominance information. **Figure 2** shows the major components of a color television signal. Although it is added directly to the monochrome signal component before transmission, the color subcarrier signal does not cause objectionable interference, because of the use of the frequency interlace technique. Because the chrominance information involves two variables, the modulated subcarrier signal varies in both amplitude and phase, and it is necessary to employ synchronous detectors to recover the two variables. A phase reference for the special local oscillator, which provides the synchronized carriers in each

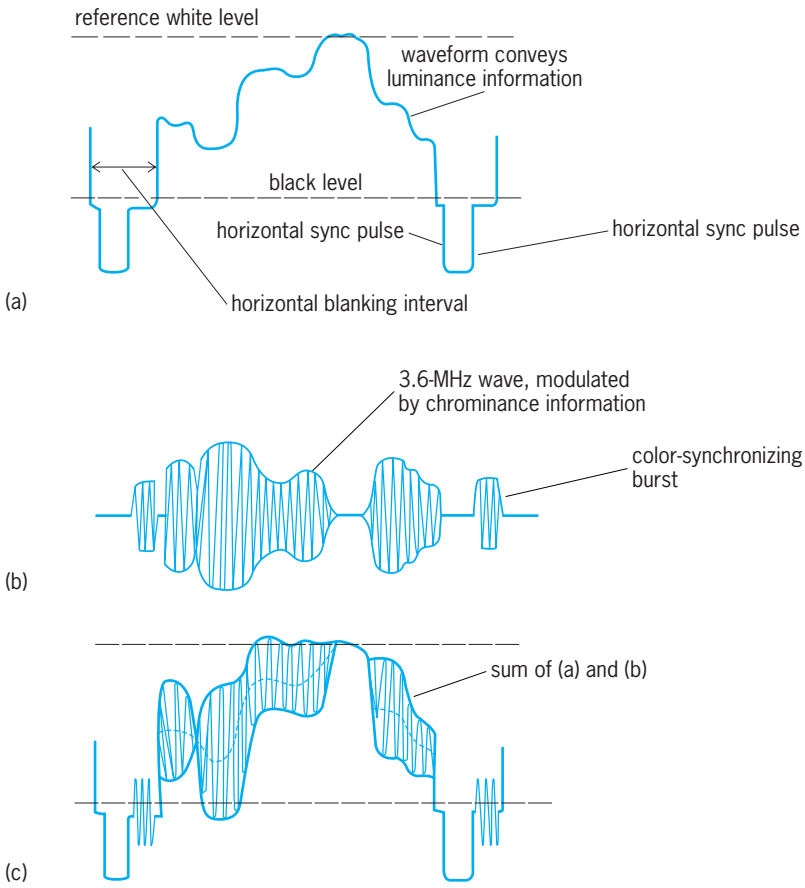


Fig. 2. Waveform sketches of major components of color television signal. (*a*) Normal monochrome signal. (*b*) Color subcarrier signal. (*c*) Complete color signal.

color receiver, is transmitted in the form of so-called color synchronizing bursts. These are short samples of unmodulated subcarrier transmitted during the horizontal blanking periods after the horizontal sync pulses.

Performance standards and requirements. A simplified block diagram for a color television receiver is shown in **Fig. 3**. Many of the circuits in a color receiver are the same in principle as the corresponding circuits in a monochrome receiver, but all circuits handling the complete color signal must be designed for high performance standards. Because the chrominance information is received in the form of sidebands occupying the upper portion of the video spectrum (centered on approximately 3.6 MHz), it is necessary that the antenna, tuner, intermediate-frequency amplifier, and video detector be designed to handle the full 4-MHz bandwidth provided in the broadcast transmission standards if degradation of the color information is to be avoided. Because the color subcarrier signal is simply added to the normal monochrome signal before transmission, it is necessary that all stages handling the complete signal be linear, so as to avoid intermodulation or distortion of the various signal components. The deflection circuits for a color receiver are similar in principle to those used in monochrome receivers, although the output stages are normally designed for a higher power level because of the greater deflection requirements for color cathode-ray tubes. Color displays use 22 to 30 kV for the high-voltage supply.

Color decoding circuits. Special decoding circuits are necessary in a color receiver to process the luminance and chrominance information in a color signal so that it can be used for the control of a practical color cathode-ray tube utilizing red, green, and blue primary colors. The major features of the most common approach to the color decoding circuits are shown within the broken lines in Fig. 3.

The video amplifier shown at the bottom handles the monochrome portion of the signal and is designed to provide attenuation in the vicinity of 3.6 MHz to block the passage of chrominance information. The chrominance information is recovered from the modulated subcarrier signal through a band-pass filter (centered at 3.6 MHz) and a pair of synchronous demodulators, in which the modulated wave is heterodyned against fixed carriers of two different phases but of the same frequency. In the most rigorous type of color decoding circuit, the chrominance components recovered from the modulated subcarrier signal are the same I and Q originally used to produce the modulated wave, but it is possible to use almost any two phase positions (not necessarily 90° apart) to recover any two independent combinations of the original I and Q signals. The bandwidths of the signals produced by the demodulators are normally adjusted somewhere between 0.5 and 1.5 MHz, and delay compensation may be required to keep all three signal components in time coincidence. The matrix circuit is essentially a linear cross-mixing network for combining the M, I, and Q signals in the proper proportions to produce red, green, and blue signals. If signals other than I and Q are produced by the chrominance demodulators, it is necessary only to design the matrix circuit so that it has slightly different mixing constants.

The synchronous carriers required for the demodulation of the chrominance information are provided by a subcarrier regenerator, which is usually a burst-controlled oscillator operating at the subcarrier frequency. Control information for the subcarrier regenerator is obtained from a burst separator, which is a gate circuit turned on only during the horizontal blanking periods by pulses derived from the horizontal deflection system. The separated bursts are compared with the output of the local subcarrier oscillator in a phase detector. If an error exists, a correction voltage is developed, which may be applied to restore the subcarrier oscillator to the proper frequency and phase. For good noise immunity, a time constant is normally provided so that control information is averaged over at least several line periods.

Color cathode-ray tube and convergence circuits. The great majority of color television receivers employ the shadow-mask color cathode-ray tube in which color images are produced in the form of closely intermingled red, green, and blue dots. The primary-color phosphor dots are excited by three separate electron beams, which are prevented from striking dots of the wrong color by the shadowing effect of an

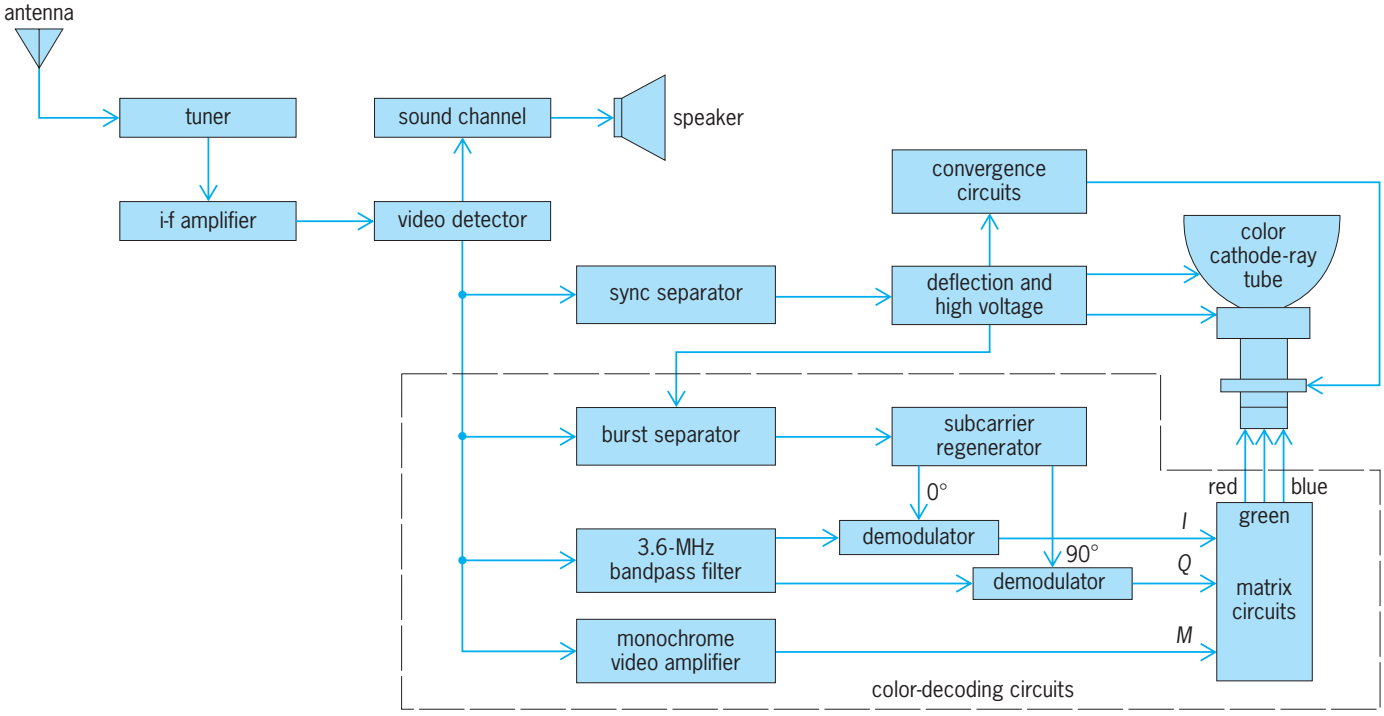


Fig. 3. Simplified block diagram of a color television receiver.

aperture mask located about $^1/_2$ in. (1.25 cm) behind the special phosphor screen. The beams in such a cathode-ray tube are deflected simultaneously by the fields produced by a single deflection yoke placed conventionally around the neck of the tube. New cathode-ray-tube designs and deflection yokes are self-converging and do not require auxiliary convergence deflection.

Color controls. In addition to the same controls required for monochrome receivers, color receivers normally have controls for convergence, hue, and saturation. The convergence controls, considered servicing adjustments only, adjust the relative amplitudes and phases of the signal components that are added together to form the proper waveforms for the convergence yoke. The hue control usually adjusts the phase of the burst-controlled oscillator and alters all the colors in the image in a systematic manner comparable to the effect achieved when a color circle diagram is rotated in one direction or the other. The proper setting for the hue control is normally determined by observing skin tones of persons on the television screen. The saturation control, frequently labeled chroma or simply color, adjusts the gain of the chrominance circuits relative to the monochrome channel and controls the saturation or vividness of the reproduced colors. When this control is set too low, the colors are all pale or pastel, and when it is reduced to zero, the picture is seen in black and white only.

Sound reproduction. The typical circuit for recovery of the sound from the aural transmitter consists of a band-pass filter to extract the frequency-modulated 4.5-MHz carrier from the video detector, amplification and amplitude limiting, frequency-modulation detection with a frequency discriminator, deemphasis with a resistance-capacitance (*RC*) low-pass filter having a time constant of 75 microseconds, followed by volume control, a power amplifier, and a loudspeaker. The result is single-channel monophonic sound. *See* LOUDSPEAKER.

Multichannel television sound. Standards have been adopted for the broadcasting of multichannel television sound (MTS) consisting of a monophonic service, a stereophonic service, and a simultaneous second-audio-program (SAP) service. A frequency spectrum plot of the multichannel sound modulation impressed upon the aural frequency-modulation transmitter is shown in **Fig. 4**. The main channel signal consists of the normal audio-frequency components up to 15 kHz, and in the case of stereophonic transmissions is made up of the arithmetic sum of the left and right audio signals ($L + R$), thus making for a compatible result for the monophonic listener. Stereophonic audio information in the form of the arithmetic difference between the left and right audio signals ($L - R$) is amplitude-compressed in

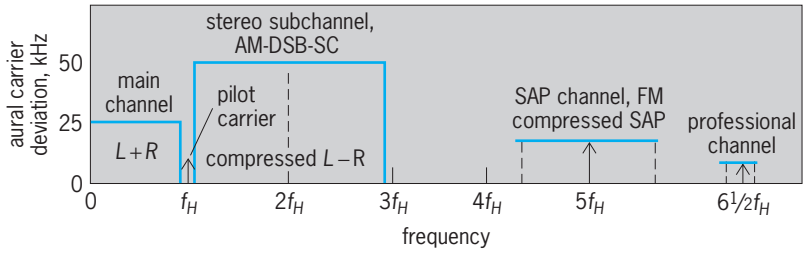


Fig. 4. Composite multichannel television sound signal baseband frequency spectrum. f_H is the horizontal scanning frequency, 15.734 kHz.

accordance with a specific algorithm and impressed as double-sideband, suppressed-carrier, amplitude modulation (AM-DSB-SC) on a carrier at twice the picture horizontal scanning frequency (31.47 kHz). A pilot carrier is sent at the picture horizontal scanning frequency (15.73 kHz) to synchronize the receiver stereo demodulator. Additionally, a second audio program (SAP; a second language, for example) is frequency-modulated on a carrier centered at five times the picture horizontal scanning frequency (78.67 kHz). The SAP audio is monophonic and is compressed with the same algorithm as the stereophonic difference ($L - R$) audio. A professional channel centered at a frequency of about 6.5 times the horizontal scanning frequency may be present and is reserved for use by the broadcaster. The sum of all the components using defined ratios is the composite multichannel baseband signal which frequency-modulates the aural carrier. *See* STEREOPHONIC RADIO TRANSMISSION; STEREOPHONIC SOUND.

Multichannel sound receiver. The MTS receiver design is still evolving, but many receivers make direct use of the 4.5-MHz frequency-modulation intercarrier available at the video detector to feed a tuned amplifier which has a wider bandwidth than that used in the monophonic receiver, a limiter, and a wideband discriminator (much wider than the monophonic equivalent) to reproduce the MTS composite baseband signal as described above. The main channel is deemphasized as in the monophonic case to recover $L + R$ audio. The stereophonic subcarrier at 31.47 kHz is synchronously demodulated with a reference carrier derived from the second harmonic of the pilot at 15.73 kHz to recover the compressed $L - R$ audio. A complementary expander using the inverse of the compression algorithm provides the correct $L - R$ audio signal which is matrixed with the $L + R$ audio to make the original left (L) and right (R) audio signals. A stereophonic amplifier and tone, balance, and volume controls complete the system together with appropriate loudspeakers. The SAP channel requires a separate FM subcarrier demodulator centered at 78.67 kHz to recover the compressed SAP audio. When the SAP audio is chosen for listening, a switch routes the compressed SAP audio to the $L - R$ expander, and disconnects both $L + R$ and $L - R$ audio signals. The compression-expansion (companding) scheme provides an effective way to reduce noise and other interferences encountered in the transmission path between transmitter and receiver, such as when the received signal is weak. The main monophonic channel noise performance is adequate without a companding scheme. *See* SOUND-REPRODUCING SYSTEMS; VOLUME CONTROL SYSTEMS.

Carl G. Eilers

Bibliography. C. G. Eilers, TV multichannel sound: The BTSC system, *IEEE Trans. Consumer Electr.*, CE-31:1-7, 1985; B. Grob and C. Herndon, *Basic Television and Video Systems*, 6th ed., 1998; M. S. Kiver and M. Kaufman, *Television Electronics: Theory and Servicing*, 8th ed., 1983; S. R. Prentiss, *Modern Television: Service and Repair*, 1989; L. B. Tyler, M. F. Davis, and W. A. Allen, A companding system for multichannel TV sound, *IEEE Trans. Consumer Electr.*, CE-30:633–640, 1984.

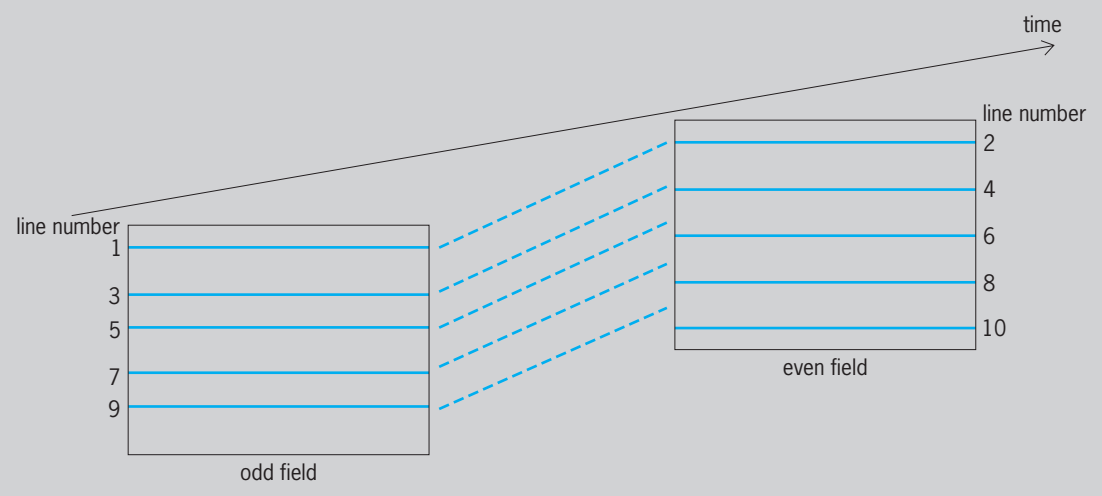
Television scanning

The process used to convert a three-dimensional image intensity into a one-dimensional television signal waveform. The image information captured by a television camera conveys color intensity (in terms of red, green, and blue primary colors) at each spatial location, with horizontal and vertical coordinates, and at each time instance. Thus, the image intensity is multidimensional, since it involves two spatial dimensions and time. It needs to be converted to a unidimensional signal so that processing, storage, communications, and display can take place.

First, the television scene is sampled many times per second in order to create a sequence of images (called frames). Then, within each frame, sampling is done vertically to create scan lines. Scanning proceeds sequentially, left to right and top to bottom. In a television camera, an electron beam scans across an electrically photosensitive target upon which the image is focused. At the other end of the television chain, with raster scanned displays, an electronic beam scans and lights up the picture elements in proportion to the light intensity. While it is convenient to think of all the samples of a single frame occurring at a single time (similar to the simultaneous exposure of a single frame for film), the scanning in a camera and in a display results in every sample corresponding to a different instance in time, and successive lines occur later in time. *See* TELEVISION CAMERA TUBE; TELEVISION RECEIVER.

Progressive and interlace scanning. There are two types of scanning approaches: progressive (also called sequential) and interlaced. In progressive scanning, the television scene is first sampled in time to create frames, and within each frame all the raster lines are scanned from top to bottom. Therefore, all the vertically adjacent scan lines are also temporally adjacent and are highly correlated even in the presence of rapid motion in the scene. Film can be thought of as naturally progressively scanned, since all the lines were originally exposed simultaneously, so the correlation between adjacent lines is guaranteed. Almost all computer displays (except some low-end computers) are sequentially scanned. *See* ELECTRONIC DISPLAY.

In interlaced scanning, all the odd-numbered lines in the entire frame are scanned first, and then the even numbered lines (see **illus.**). This process produces two distinct images per frame, representing two distinct samples of the image sequence at different points in time. The set of odd-numbered lines constitute the odd field, and the even-numbered lines make up the even field. All current television systems use interlaced scanning. One principal benefit of interlaced scanning is to reduce the scan rate (or the bandwidth). This is done with a relatively high field



Interlaced scanning. A television frame is divided into an odd field (containing odd-numbered scan lines) and an even field (containing even-numbered scan lines).

rate (a lower field rate would cause flicker), while maintaining a high total number of scan lines in a frame (lower number of lines per frame would reduce resolution on static images). Interlace cleverly preserves the high-detail visual information and, at the same time, avoids visible large-area flicker at the display due to temporal postfiltering by the human eye.

The North American Television Standard (NTSC) has 15,735 scan lines per second or 525 lines per frame, since there are 29.97 frames per second. For each scan line, a small period of time (16–18% of total line time), called blanking or retrace, is allocated to return the scanning beam to the left edge of the next scan line. European systems (PAL and SECAM) have 625 lines per frame, but only 25 frames per second. The larger number of lines results in better vertical resolution, whereas the larger number of frames in NTSC results in better motion rendition. *See* TELEVISION STANDARDS.

High-definition television. While there is no worldwide standard yet, all versions of high-definition television (HDTV) have approximately twice the horizontal and vertical resolution of conventional systems. In addition, high-definition television is digital, where the television scan lines are also sampled horizontally in time and digitized. Such sampling produces an array of pixels (picture elements). High-definition television has an array of approximately 1000 lines and as many as 2000 pixels per line. If the height-to-width ratio of the television raster is equal to the number of scan lines divided by the number of samples per line, the array is referred to as having square pixels; that is, the electron beam is spaced equally in the horizontal and vertical directions. This facilitates digital image processing as well as computer synthesis of images.

An agreement (not a standard) on high-definition formats has been reached and is in use, enabling the transition from analog to digital television. In early 2001, there were more than 150 stations broadcasting high-definition television. The format has 1920 active pixels per line and 1080 active (out of a total of 1125) lines in a frame. The frames may be interlaced or progressive, and the frame rate is 29.97 Hz. Progressive frames at a rate of 23.976 Hz are also permitted to accommodate film material. An alternative progressive-only format that provides additional temporal resolution at the expense of some spatial resolution has also been approved and is currently in use. This format has 1280 active pixels per line and 720 active (out of a total of 750) lines per frame, and the frame rates permitted are 23.976 Hz, 29.97 Hz, or 59.94 Hz. Both high-definition formats have square pixels and a 16 × 9 aspect ratio. *See* DATA COMPRESSION; IMAGE PROCESSING; TELEVISION.

Arun N. Netravali

Bibliography. A. N. Netravali and B. G. Haskell, *Digital Pictures*, 1994; Y. Ninomiya et al., An HDTV broadcasting system utilizing a bandwidth compression technique—MUSE, *IEEE Trans. Broadcasting*, BC-33:130, 1987; C. A. Poynton, *Digital Video and HDTV*: *Pixels*, *Pictures*, *and Perception*, 2001.

Television standards

The accepted criteria for a television system, including the image aspect ratio, number of lines per frame, type of scanning, original video signal bandwidth, transmission format and bandwidth, reception, demodulation, decoding, and sound system. The implementation of high-definition television (HDTV), where the image resolution and audio fidelity are significantly higher than for conventional television, has required new standards. *See* TELEVISION.

Early television. Black-and-white television dates back to the mid-1930s. In the United States the screen has 525 interlace lines in one frame. Within the 525 interlaced lines are 483 active lines of video information. Interlacing implies that the odd-numbered lines are transmitted in the first field in 1/60 second, followed by the even numbered lines in the second field in 1/60 s. Thus there are two fields per frame, and

30 frames/s, or 60 fields/s—just enough to eliminate any discernible flicker in the displayed image.

As an image is scanned and then later displayed, the intensity defines whether the picture element (pixel) is white, when the intensity is low, or black, when the intensity is high, or gray, when the intensity takes on a middle value. This intensity video signal is called the luminance. It takes 63.5 microseconds to traverse one line of an image. This time interval determines the line frequency as $1/(63.5\ \mu s) = 15.75$ kilohertz. The conventional television image has an aspect ratio (ratio of the image width to the image height) of 4:3. Using these television system parameters, the bandwidth of the video signal is approximately 4.2 megahertz. The resolution of a television image is the number of image lines that can be distinguished in the vertical and horizontal directions. The vertical resolution is the number of active lines, 483, although the slight misalignment of the raster scan reduces this by 30% to 340 lines. The horizontal resolution using the video signal bandwidth and the line frequency is approximately 450 lines. *See* BANDWIDTH REQUIREMENTS (COMMUNICATIONS); ELECTRONIC DISPLAY; TELEVISION SCANNING.

A complete television signal is created by including a monaural sound signal above the video signal using frequency modulation (FM) with a sound subcarrier of 4.5 MHz. The bandwidth of the resulting sound signal is approximately 80 kHz. The complete signal is then transmitted using vestigial sideband (VSB) amplitude modulation (AM) with a large carrier. The VSB-AM signal includes the entire upper sideband (4.5 MHz) and 1.25 MHz of the lower sideband. The monochrome television signal occupies a transmission bandwidth of 6 MHz including a frequency guard band of 250 kHz. The television carrier frequencies are assigned by the federal communications commission (FCC) in the very high frequency (VHF) ranges (54–72 MHz, 76–88 MHz, and 174–216 MHz), as well as in the ultrahigh frequency (UHF) range (470–806 MHz). After transmission, the VSB signal is demodulated, and the luminance and sound signals are recovered. *See* AMPLITUDE MODULATION; FREQUENCY MODULATION; SINGLE SIDEBAND; TELEVISION TRANSMITTER.

Color television standards. In the early 1950s, the National Television Systems Committee (NTSC) was formed to set standards for a color television signal (in the United States) that would be fully compatible with the existing monochrome signal. Any color can be formed as a linear combination, that is, a weighted sum, of red (R), green (G), and blue (B). The NTSC standard started with three image signals—R, G, and B—and matrixed these three primary color image signals as linear combinations into one luminance signal (the conventional black-and-white video signal, often called the Y-signal) and two chrominance signals that control hue and saturation. The chrominance signals are termed the in-phase (I) signal and the quadrature (Q) signal. Although the luminance signal retained the original 4.2-MHz bandwidth, the characteristics of human color perception allowed the I-signal to be limited to 1.5 MHz and the Q-signal to only 0.5 MHz.

The power spectrum of the luminance video signal, that is, the distribution of signal power versus frequency, displays distinct lines, or concentrations of power, at multiples of the horizontal line frequency (15.75 kHz). To avoid interference between the luminance signal and the chrominance signals, the color signals are multiplexed with a color subcarrier whose frequency is exactly midway between the 227th and the 228th harmonic of the horizontal line frequency. Using a color subcarrier of 3.58 MHz allows the spectral lines of the chrominance signal to be interleaved between the high-frequency (and low-amplitude) spectral lines of the luminance signal. To retain the original video signal bandwidth, the I-signal is amplitude-modulated with the cosine of the color subcarrier and then filtered to 4.2 MHz, while the Q-signal is amplitude-modulated with the sine of the color subcarrier.

The resulting composite color television signal contains the luminance signal plus the interleaved quadrature amplitude-modulated (QAM) chrominance signals with a 3.58-MHz color subcarrier plus the FM sound signal. A black-and-white television receiver can recover the luminance (Y-signal) and the sound, while a color television can also recover the chrominance (I- and Q-signals). Thus compatibility is achieved. The color television receiver converts the Y-, I-, Q-signals back to R-, G-, B-signals via a decoding matrix, which is just another set of linear combinations, before displaying the color image. Further developments of the sound signal resulted in stereo audio similar to commercial FM radio stations that use a left plus right (L + R) primary signal and a frequency-multiplexed left minus right (L − R) secondary signal. The total audio signal is then applied to the television FM generator. Modern color television receivers recover L + R along with L − R, and produce stereo sound by adding and subtracting the primary (L + R) and the secondary (L − R) audio signals. *See* TELEVISION RECEIVER.

While NTSC color television standards are used in North America, South America, and Japan, the European International Radio Consultative Committee (CCIR) system is used in England, Germany, Italy, and Spain. The color system employed with CCIR television is called Phase Alternate Line (PAL). A modified CCIR television standard is used in France and Russia, where the color system is called SECAM (Sequential Couleur à Mémoire). A list of some of the television standards is given in **Table 1**. These standards use interlaced scanning, so the field rate is twice the frame rate.

HDTV standards. The next generation of television, HDTV, is not compatible with the previous television systems. HDTV relies on digital technology, making it more amenable with computer displays, while taking full advantage of the power and efficiency of digital signal processing (DSP). In 1987, the Federal Communications Commission (FCC) chartered an advisory committee to recommend initial HDTV standards for the United States. This advisory committee

TABLE 1. Broadcast color television standards

Characteristics	NTSC	CCIR (PAL)	CCIR (SECAM)
Video bandwidth, MHz	4.2	5.0	5.0
Channel bandwidth, MHz	6.0	7.0	8.0
Total lines/frame	525	625	625
Frame rate, Hz	30	25	25
Line rate, kHz	15.75	15.625	15.625
Aspect ratio	4:3	4:3	4:3

was transformed into a group of eight organizations called the Grand Alliance (GA) in 1993 as the number of proposed HDTV schemes was reduced to four digital television systems. The desired characteristics of the HDTV system that would replace the NTSC system was expected to have a resolution that would approach the quality of a 35-mm film, that is, approximately twice the horizontal and twice the vertical resolution of conventional television, with a widescreen aspect ratio of 16:9. The target HDTV system standard was required to avoid interlace scanning artifacts, as well as chrominance artifacts and deliver digital multichannel audio. The end result was a 1996 FCC digital television (DTV) standard, and HDTV broadcasting commenced in the United States in 1998. The successor to the Grand Alliance is the Advanced Television Systems Committee (ATSC), an international organization that has established voluntary technical standards and recommended practices for HDTV.

The most challenging aspect of the DTV standard was to achieve the desired resolution with digital video and digital audio, while maintaining a transmission channel bandwidth of 6 MHz for terrestrial broadcast channels. This was achieved by using powerful digital compression algorithms to reduce the raw data rate from 1.5 Gbps (gigabits per second) to approximately 20 Mbps (megabits per second). The United States DTV standard that was adopted actually consists of multiple standards with different resolutions and frame rates. Details of the two HDTV formats are provided below along with the data compression and transmission specifications. *See* DATA COMPRESSION.

The transmitter of an HDTV system consists of a set of operations. It starts with the scanned luminance and chrominance video signals, and the wideband, multichannel audio signals, and ends with the radio-frequency (RF) signal that is broadcast to HDTV receivers. The main HDTV transmitter operations are video/audio coding and compression, data multiplexing into packets, data scrambling, channel coding (for error detection and correction), synchronization multiplexing, and digital modulation (for broadcast transmission). The HDTV receiver reverses the operations of the transmitter.

The two video formats included in the DTV standard use a 16:9 aspect ratio, but the horizontal and vertical resolutions differ, and several scanning frame rates are allowed. These are given in **Table 2**, where "active lines/frame" indicates the vertical resolution, and "pixels/scan line" designates the horizontal resolution. Progressive scanning is simply traversing each line of the image from top to bottom to complete one frame.

The HDTV video signal bandwidth specification is 24.9 MHz, and the audio signal bandwidth is 20 kHz, where five such audio signals can be encoded to provide multichannel Dolby digital surround sound. The necessary data compression is achieved for the video signal by employing a Motion Picture Experts Group (MPEG) encoding technique, called MPEG-2. This technique blocks the image into groups of 8×8 pixels within each frame, and executes a discrete cosine transform (DCT) on the 64 pixels resulting in a new set of coefficients, many of which are insignificant and need not be transmitted. In addition, frame-to-frame redundancy is removed by motion estimation and motion compensation via predictive coders.

Data multiplexing combines the digitized, compressed video and audio with ancillary data for such things as closed captions, resulting in transport packets that are scrambled to ensure random binary data. The data rate, with the synchronization signal included, is 19.39 Mbps. Channel coding consists of a Reed-Solomon (R-S) forward error correcting (FEC) code that increases the data rate to 21.5 Mbps, and rate 2/3 trellis coded modulation (TCM) that further increases the data rate by 50% to 32.3 Mbps. The inclusion of these channel coding operations guarantees that errors can be detected and corrected by the HDTV receiver. *See* DATA COMMUNICATIONS; INFORMATION THEORY; MULTIPLEXING AND MULTIPLE ACCESS.

To ensure that the broadcast signal fits into the specified 6-MHz transmission bandwidth, groups of 3 bits are converted to one of 8 levels, resulting in a digital signal that has a symbol or baud rate that is one-third of the 32.3 Mbps data rate, that is, 10.76 Mbps. This 8-level baseband signal is symbol-shaped to avoid intersymbol interference (ISI) using a sine-rolloff filter with a rolloff factor, $r = 0.0575$. The resulting baseband signal is amplitude-modulated using 8-level VSB, with a pilot

TABLE 2. HDTV image formats

Characteristics	HDTV 1	HDTV 2
Active lines/frame	1080	720
Pixels/scan line	1920	1280
Frame rate (Hz) and scan type	24, progressive 30, progressive 30, interlace	24, progressive 30, progressive 60, progressive

carrier signal, whose shaped upper sideband (5.69 MHz) is included, along with 5.4% of the lower sideband (0.31 MHz), such that the 6-MHz transmission bandwidth is maintained. *See* AMPLITUDE MODULATION; MODULATION.

In Europe, a different transmission standard has been adopted, in lieu of 8-VSB for digital television. The European system is referred to as the Digital Video Broadcasting (DVB) standard. It uses coded orthogonal frequency division multiplexing (COFDM) as a modulation scheme.

The FCC expects everyone to be using new HDTV receivers by the year 2011, at which time NTSC broadcasting will cease, and all NTSC color television receivers will need to be replaced or modified with some type of converter to be able to decode and display HDTV images. Joseph L. LoCicero

Bibliography. *Guide to the Use of ATSC Digital Television Standard*, Advanced Television Systems Committee, Washington, DC, Document A/54, October 4, 1995; G. N. Patchett, *Color TV Systems*, 1969; J. C. Whitaker, *HDTV: The Revolution in Digital Video*, 1999.

Television studio

A facility designed for the production of television programs, which may be broadcast live concurrently with the production or recorded for later broadcast. A television studio consists of the studio room, wherein the actual program takes place, and various support rooms, which include the control room, the equipment room, and the property room.

Studio room. The studio room is where the program action occurs and is analogous to a theatrical stage (**Fig. 1**). Studio rooms may be of almost any size, depending on use, but invariably provide certain facilities, such as a flexible lighting and scenery system, one or more cameras, one or more microphones for sound pickup, and a communications system to allow coordination during the program. The larger studios may provide seating for an audience.

Fig. 1. Television studio room. (*KDVR, Fox-31, Denver, CO*)

Most studios use a lighting grid suspended from a high ceiling that allows flexible placement of the various lighting fixtures. The grid is a steel framework that has electrical power outlets at regular intervals throughout its structure and allows the fixtures to be moved into any desired position. Portions of the grid may be lowered, or a catwalk may be provided for adjustments. Cabling to a lighting switch panel connects each electrical outlet. Jumper cables or multiple-pole switches are employed to assign each outlet to a particular dimmer circuit. Several outlets may be assigned to each dimmer, so that the lighting operator may control a number of lights with one control. Each dimmer control circuit may vary the brightness of its associated lights from full off to full on, usually by means of silicon controlled rectifiers grouped together in a convenient location and remotely controlled from the lighting control panel. Each fixture on the lighting grid may be equipped with various accessories, such as "barn-doors" (deflectors), scrims (diffusers), and gels (color filters), to provide the effects desired by the lighting director. *See* SEMICONDUCTOR RECTIFIER.

Automated lighting systems employ computers which have the ability to store in memory the various lighting calls used in a production. Once the desired sequence is programmed during rehearsal, it may be recreated flawlessly during the actual performance. An added benefit is that complicated lighting changes, requiring quick or simultaneous actions, are easily within the capability of the computer. *See* COMPUTER.

Studios may be equipped with various forms of scenery or sets. Often a fixed set may be used repeatedly for a series production or news programs where a familiar setting is required. Other productions may call for rapid changes of scenery, and the studio ceiling grid is used for hoisting undesired scenic elements out of camera view. The cyclorama, a scenic device often found in the studio, is essentially a curved wall which joins the floor so that no distinct corners are formed, the on-camera effect of which is the illusion of infinite space. This allows many visual effects to be created by use of special lighting. Powerful computers now allow the creation of virtual sets, which are combined with camera video to create the illusion of a large and complex physical set. Merely changing the computer program can redesign the set. News programs take particular advantage of this technology to enhance the viewing experience.

The well-designed television studio makes provision for plugging cameras, microphones, and intercom headsets into wall outlets placed in appropriate locations. It provides monitoring facilities for both video and audio, and has a studio announcing system so that the director can simultaneously address the entire production team from the control room.

Special sound-absorbing materials usually line the walls and ceilings to prevent unwanted acoustic reflections within the studio and to keep outside noises from intruding. The studio floor must be level and smooth, so that camera and microphone boom dollies can be rolled about the room noiselessly and without creating any jiggle in the camera's picture. The entire studio room must be provided with an air-conditioning system capable of noiselessly removing the considerable heat given off by the lighting system. *See* MICROPHONE; TELEVISION CAMERA.

Control room. The studio control room is the nerve center of the television production facility (**Fig. 2**). The room is often at the rear of the studio room in an acoustically isolated booth. Visual contact between the control room and the studio may be provided through soundproof windows, but this is not necessary. The program director, technical director, and sound engineer each have a station within the control room equipped with intercom headsets so that they can communicate quickly and efficiently.

The control room usually has a bank of video monitors with screens which display the output of each camera, videotape recorder, or special-effects generator, as well as a previous monitor which shows the director what the next shot will look like and a line monitor which shows the scene currently on the air. The technical director sits at a video switcher which allows him or her to put various cameras or other video sources on the air in response to orders from the program director. The video switcher is a device which can provide a variety of transitions between different sources, such as dissolves, fades, cuts, and wipes. It can also provide chroma key effects which can make the studio appear to be part of the prerecorded background scene. Modern switchers may incorporate computers to assist the technical director in complicated transitions or special effects.

The sound engineer usually sits behind the technical director, often in a sound-isolated booth. The sound engineer operates an audio mixer console which has every microphone used in the studio connected to a separate input. Other sources, such as turntables, audio tape recorders, tape cartridge machines, compact disk players and audio hard disk recorders (or audio servers), may also be connected to the audio mixer. The sound engineer controls the levels in accordance with the program director's instructions and ensures that program dialog can be heard above background sound effects. Equipment is often available that can alter the sound passing through it by means of equalization, echo chambers, and a wide variety of digital effects. The sound engineer is responsible for keeping the overall audio level within a standard range. *See* SOUND-RECORDING STUDIO.

Equipment room. The equipment room may be located some distance from the studio and control rooms. It contains the electronic equipment associated with the cameras, the video switcher, and special-effects generators. It may contain such equipment for several studios in large facilities. Video engineers operate the camera control units and other equipment by using waveform monitors and vectorscopes to ensure that video levels are correct. Therefore the technical director in the control room may select various video sources without large deviations in signal level or picture quality appearing on the air. The equipment room may also provide video- and audio-processing equipment to make final corrections to the picture and sound before the program is sent to a tape recorder, video hard disk recorder (also called a video server), the station transmitter, or a network of stations.

Fig. 2. Studio control room. (*KDVR, Fox-31, Denver, CO*)

Property room. The property (or prop) room is usually located adjacent to the studio. It is generally a large open area with shelves and bins suitable for storage of the scenic elements and props, such as chairs and tables, used in the production of television shows.

Earl F. Arbuckle, III

Bibliography. T. D. Burrows et al., *Television Production*, 6th ed., 1995; W. E. McCavitt, *Television Studio Operations*, 1980; National Association of Broadcasters, *NAB Engineering Handbook*, 8th ed., 1992; P. B. Seel and A. E. Grant, *Broadcast Technology Update: Production and Transmission*, 1997; J. Watkinson, *An Introduction to Digital Video*, 2d ed., 2001; J. C. Whitaker and K. B. Benson (eds.), *Standard Handbook of Video and Television Engineering*, 3d ed., 2000; A. Wurtzel et al, *Television Production*, 4th ed., 1995; H. Zettl, *Television Production Handbook*, 7th ed., 1999.

Television transmitter

An electronic device that converts audio and video signals into modulated radio-frequency (rf) energy which can be radiated from an antenna and received by a television receiver. The term can also refer to the entire television transmitting plant, consisting of the transmitter proper, associated visual and aural input

and monitoring equipment, transmission line, the antenna with its tower or other support structure, and the building in which the equipment is housed. In the United States, both analog NTSC (National Television Systems Committee) and digital 8-VSB transmitters are in service. The digital transmitters are used for what is termed high-definition television (HDTV).

An analog television transmitter can be thought of as two separate transmitters integrated into a common cabinet (**Fig. 1***a*). Video information is transmitted via a visual transmitter, while audio information is transmitted via an aural transmitter. Because video and audio have different characteristics, the two transmitters differ in terms of bandwidth, modulation technique, and output power level. Nevertheless, a common transmitting antenna is generally used, and the two transmitters feed this antenna via an rf diplexer or combiner. Some modern analog UHF transmitters actually combine the visual and aural signals at a very low level and then employ common high-power amplification systems. This requires careful engineering to avoid significant intermodulation products.

Fig. 1. Television transmitters. (*a*) Analog solid-state transmitter (*Harris Broadcast*). (*b*) Digital inductive-output tube (IOT) transmitter (*Thomcast Communications*).

A digital transmitter (Fig. 1*b*) accepts a single encoded digital bit stream that may contain video, audio, and data. In the United States, the digital terrestrial transmission standard is known as 8-VSB, which is an eight-level, vestigial sideband format. The FCC has mandated that all U.S. television stations convert to digital and terminate analog transmissions by 2006, although many view this as an optimistic timetable for the changeover.

Television stations are licensed to operate on a particular channel, but since it takes a very wide bandwidth to transmit a television picture, these channels are allocated over a broad range of frequencies. Channels 2 through 6 are low-band very-high-frequency (VHF) channels, while channels 7 through 13 are high-band VHF channels. Channels 14 through 69 are ultrahigh-frequency (UHF) channels. Each channel is 6 MHz wide. Because of the wide range of frequencies, television transmitters are designed to work in only one of the foregoing groups, and employ specific circuits which are most efficient for the channels involved. Nevertheless, every television transmitter, regardless of operating frequency, transmits a standard television signal in conformity with the regulations of the country in which it is operated. *See* RADIO SPECTRUM ALLOCATION.

Signal characteristics. In the United States the Federal Communications Commission (FCC) specifies the standard television signal in its rules This specification enables receiver manufacturers to market receivers that are compatible with all television transmitters.

For the NTSC standard, the FCC requires that a visual transmitter produce an amplitude-modulated (AM) carrier with an upper sideband extending to 4.2 MHz above the carrier and a lower sideband extending to only 0.75 MHz below the carrier. The lower sideband is restricted to this narrow bandwidth in order to conserve valuable frequency spectrum. Since both sidebands contain the same information, only one is required to transmit a picture. This is known as vestigial sideband transmission (**Fig. 2**). *See* AMPLITUDE MODULATION; SINGLE SIDEBAND.

FCC rules provide that the aural transmitter be frequency-modulated (FM) and that its carrier frequency must be 4.5 MHz above the visual carrier. This standard spacing between carriers allows use of a simplified sound receiver in most television sets. In the aural transmitter 100% modulation is defined as equal to 25 kilohertz deviation, and the transmitter must be capable of faithfully passing audio-modulating frequencies from at least 50 Hz to 15 kHz. *See* FREQUENCY MODULATION; TELEVISION STANDARDS.

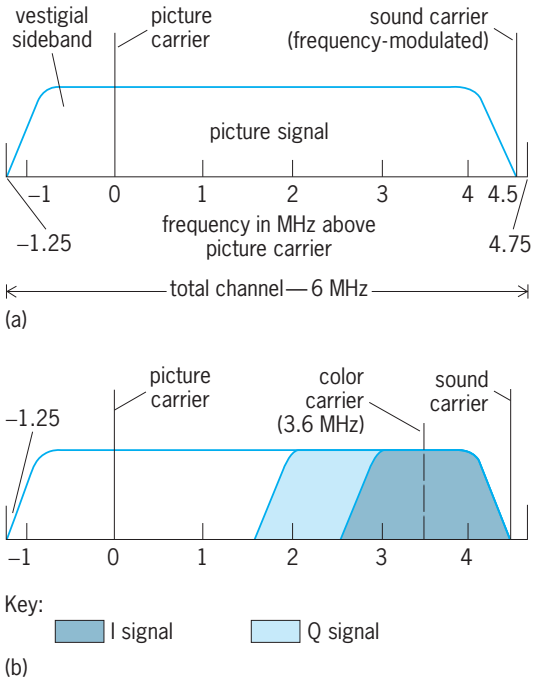


Fig. 2. Diagram of television channel showing portions occupied by color and monochrome signal components. (*a*) Monochrome television channel. (*b*) Color television channel. (*After G. W. Bartlett, ed., NAB Engineering Handbook, 6th ed., National Association of Broadcasters, 1975*)

Transmitter power. Television broadcast stations are limited to a specific effective radiated power (ERP) by the FCC. ERP is defined as the transmitter output power multiplied by transmitting antenna gain and an efficiency factor (less than 100%) due to the losses of the transmission line components between the transmitter and antenna.

To provide a consistent signal strength at the receiver, the FCC allows UHF stations to operate with more power (a maximum of 5 MW peak visual ERP in analog service, or 1 MW average ERP in digital service) than high-band VHF stations, and high-band VHF stations may operate with greater power (316 kW maximum) than low-band VHF stations (100 kW maximum). The maximum power that any station may utilize is reduced proportionately if its antenna height exceeds 1000 ft (305 m) above average terrain. Furthermore, due to its narrower bandwidth and other factors, the aural FM signal tends to carry better than the visual signal, so it is restricted to between 10 and 20% of the visual power by the FCC.

Because transmitting antennas usually have a power gain greater than unity, television transmitters need only provide a fraction of the ERP. Modern transmitters are rated at 10 W to 30 kW for the low-band VHF, and 10 W to 75 kW for high-band VHF; UHF transmitters are manufactured with outputs ranging from 100 W to more than 300 kW.

The lowest-power transmitters are generally configured as television translators, which are used to relay the signal of a high-power primary station into areas where terrain or other factors prevent viewers from being able to receive the primary station. The translator accepts the input on the primary channel and shifts its output to another channel so that one does not interfere with the other.

Transmitting antennas. Antenna gain is a function of design and the number of sections employed. Gain usually increases with physical size, due to increased radiating area, and may be increased further by stacking identical elements vertically; VHF stations usually operate with antenna gains of less than 5, while UHF stations commonly use gains of up to 50 in order to generate the much greater effective radiated power allowed by the FCC. The same antenna may be used for analog or digital transmission, sometimes simultaneously as in the case of multichannel, broadband antennas.

The horizontal radiation pattern of most television transmitting antennas is circular, providing equal radiated signal strength to all points of the compass. Higher-gain antennas achieve greater power in the direction of the horizon by reducing the power radiated at vertical angles above and below the horizon. Since this could result in weaker signals at some receivers close to the transmitter, beam tilt and null

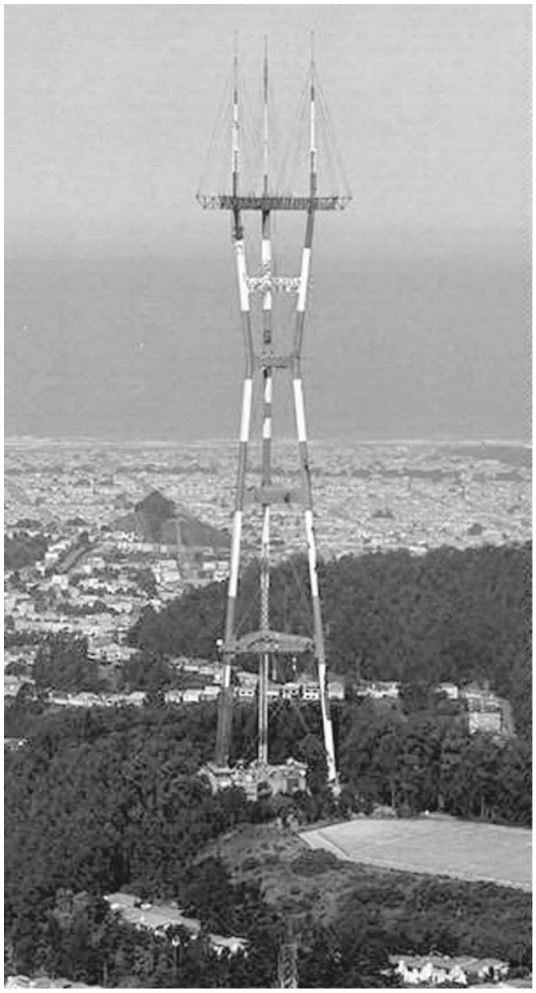


Fig. 3. Sutro Tower, with most of the television transmitters in San Francisco. (*KGO-TV, San Francisco*)

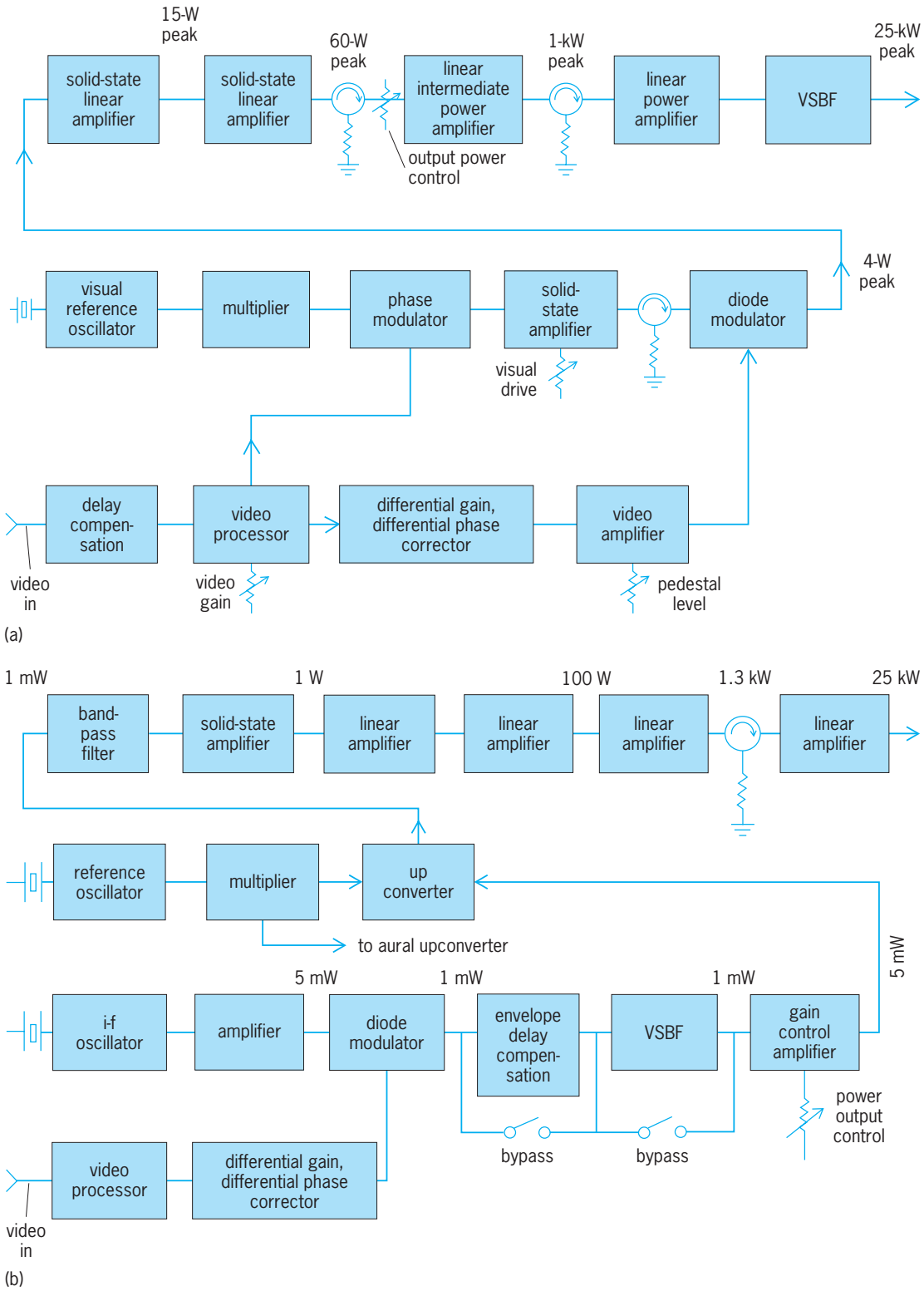


Fig. 4. Visual transmitters with modulation and vestigial-sideband filter (VSBF). (*a*) Transmitter with filtering of high-power signal at final frequency. (*b*) Transmitter with filtering of low-level, intermediate-frequency signal. (*After G. W. Bartlett, ed., NAB Engineering Handbook, 6th ed., National Association of Broadcasters, 1975*)

fill are often used to lower the angle of maximum radiated power.

Until about 1978, all television stations employed horizontally polarized antennas so that their signals would suffer less interference from impulse noise sources, such as automobile ignition systems, which tend to be vertically polarized. Some broadcasters then began utilizing circularly polarized antennas in the belief that television sets with "rabbit-ear" receiving antennas would obtain a better signal.

Because television signals travel in a "line of sight," transmitting antennas are usually placed as high as possible above ground with respect to the surrounding service area. Such locations minimize signal blockage or ghosting due to tall buildings and hills. It is also desirable to locate all of the transmitting antennas serving a given locality in the same place. This allows viewers to orient their receiving antennas in one direction for the best reception from all of the stations. In Los Angeles, for example, advantage is taken of a 5700-ft (1740-m) summit, Mount Wilson, to locate most of the transmitting antennas on individual 200- to 500-ft (60- to 150-m) towers. In New York City all of the transmitting antennas were mounted on a common 365-ft (111-m) mast atop the 1366-ft (416-m) World Trade Center's North Tower before the building was destroyed in a terrorist attack on September 11, 2001. In San Francisco, most of the television transmitters are located on the 977-ft (300-m) Sutro Tower (**Fig. 3**). In areas with neither mountains nor tall buildings, extremely tall towers have been erected, some of which exceed 2000 ft (600 m). Where many broadcasters are colocated, the impact of radio-frequency radiation (RFR) has become an issue. Federal guidelines require that human exposure to such radiation be limited to specific levels. Controversy exists as to the real impact of such exposure. *See* ANTENNA (ELECTROMAGNETISM).

Transmitter designs. There are two broad classes of VHF analog visual television transmitter design philosophy. The classical approach modulates the visual carrier at a moderate power level, amplifies the carrier to rated output power by means of high-power linear amplifiers, and then filters this high-power carrier to obtain the required vestigial-sideband signal (**Fig. 4***a*). The more contemporary approach, used by nearly all transmitter manufacturers, employs modulation at a very low power level of an intermediate-frequency (i-f) signal. The required vestigial-sideband filtering is imposed on this low-level signal, generally by means of a highly stable surface-acoustic-wave filter, whereupon the signal is upconverted to the carrier frequency and amplified by linear amplifiers to rated output power (Fig. 4*b*). *See* AMPLIFIER; SURFACE-ACOUSTIC-WAVE DEVICES.

Generally, modern analog and digital UHF transmitters employ large power tubes, called inductive output tubes (IOTs), to produce the large amounts of rf carrier power required. **Figure 5** shows such a tube in its circuit assembly.

Every analog transmitter contains a visual and aural exciter. This element determines the operating frequency of the visual and aural carriers and must be extremely stable, since the FCC requires that they be kept within 1 kHz of the assigned frequencies. The aural exciter produces the required frequency modulation by varying the frequency of the rf carrier oscillator at an audio rate (**Fig. 6**). Digital transmitters handle audio as simply an indistinguishable component of the digital bit stream applied to the digital exciter.

Regardless of the visual modulation approach taken, certain parameters must be kept within FCC tolerances. Suitable circuits must compensate nonlinearity of the rf power amplifier stages. Certain time delays, within the transmitter, called envelope delays, must be the complement of those found in the home television receiver, so that colors in the picture are properly superimposed. Flat frequency response over adequate bandwidth is necessary to ensure that the picture has good detail and accurate color rendition. The FCC, however, does not permit excessively wide response, so that a low-pass filter that attenuates video above 4.75 MHz by at least 20 dB is commonly inserted in the video input circuit. *See* ELECTRIC FILTER.

Analog transmission of color requires that the color subcarrier (3.58 MHz above the visual carrier) not be affected by changes in the luminance level of the picture. A change in color saturation (chrominance level) brought about by a change in luminance level is termed differential gain. A change in color hue (subcarrier phase) brought about by a change in luminance level is termed differential phase. Both parameters are to be minimized.

The FCC also strictly regulates the radiated rf carrier. Harmonics, or multiples, of the carrier frequency must be attenuated at least 60 dB below the peak power level. Harmonic filters are placed on the output of both the visual and aural transmitters, ahead of any aural-visual diplexers, to ensure

Fig. 5. Ultrahigh-frequency (UHF) inductive output tube (IOT) in circuit assembly. (*Marconi Applied Technology*)

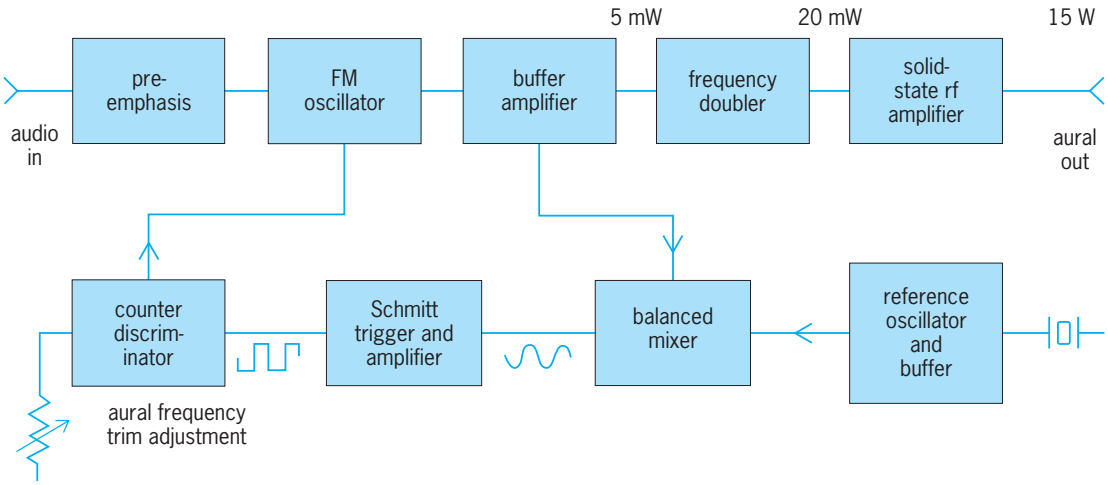


Fig. 6. Typical aural exciter. (*After G. W. Bartlett, ed., NAB Engineering Handbook, 6th ed., National Association of Broadcasters, 1975*)

compliance. In those transmitters which employ high-level vestigial-sideband filtering, a device known as a filterplexer often combines the functions of sideband filtering and aural-visual diplexing. Such filters are also used on the output of digital transmiters. *See* TELEVISION; TELEVISION CAMERA; TELEVISION CAMERA TUBE; TELEVISION NETWORKS; TELEVISION RECEIVER; TELEVISION SCANNING; TELEVISION STUDIO. Earl F. Arbuckle, III

Bibliography. R. Blair, *Digital Techniques in Broadcasting Transmission*, 2d ed., 2002; D. Christiansen (ed.), *Electronics Engineers' Handbook*, 4th ed., 1997; P. Dambacher, *Digital Terrestrial Television Broadcasting: Designs, Systems, and Operation*, 1998; P. Hodges, *An Introduction to Video Measurement*, 2d ed., 2001; P. B. Seel and A. E. Grant, *Broadcast Technology Update: Production and Transmission*, 1997; M. Silbergleid and M. J. Pescatore, *The Guide to Digital Television*, 3d ed., 2000; J. C. Whitaker and K. B. Benson (eds.), *Standard Handbook of Video and Television Engineering*, 4th ed., 2003; E. Williams et al., *National Association of Broadcasters Engineering Handbook*, 10th ed., Elsevier, 2007.

Tellurium

A chemical element, Te, atomic number 52, and atomic weight 127.60. There are eight stable isotopes of natural tellurium. Tellurium makes up approximately 10^{-9} % of the Earth's igneous rock. It is found as the free element, sometimes associated with selenium. It is more often found as the telluride sylvanite (graphic tellurium), nagyagite (black tellurium), hessite, tetradymite, altaite, coloradoite, and other silver-gold tellurides, as well as the oxide, tellurium ocher. *See* PERIODIC TABLE.

There are two important allotropic modifications of elemental tellurium, the crystalline and the amorphous forms. The crystalline form has a silver-white color and metallic appearance. This form melts at 841.6°F (449.8°C) and boils at 2534°F (1390°C). It has a specific gravity of 6.25, and a hardness of 2.5 on Mohs scale. The amorphous form (brown) has a specific gravity of 6.015. Tellurium burns in air with a blue flame, forming tellurium dioxide, TeO_2. It reacts with halogens, but not sulfur or selenium, and forms, among other products, both the dinegative telluride anion (Te^{2-}), which resembles selenide, and the tetrapositive tellurium cation (Te^{4+}) which resembles platinum(IV).

1																	18
1 H	2											13	14	15	16	17	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Tellurium is used primarily as an additive to steel to increase its ductility, as a brightener in electroplating baths, as an additive to catalysts for the cracking of petroleum, as a coloring material for glasses, and as an additive to lead to increase its strength and corrosion resistance. *See* SELENIUM. Stanley Kirschner

Bibliography. P. W. Atkins et al., *Inorganic Chemistry*, 4th ed., 2006; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; D. Sangeeta, *Inorganic Materials Chemistry Desk Reference*, 1997.

Telosporea

A class of the subphylum Sporozoa. These protozoa are divided into two subclasses, the Gregarinia and Coccidia. All members of the group are either intra- or extracellular parasites, and the life cycles have

both sexual and asexual phases. The spores lack a polar capsule and develop from an oocyst. The sporozoite is the usual infective stage which initiates the asexual phase in the life cycle. *See* COCCIDIA; GREGARINIA; PROTOZOA; SPOROZOA.

Elery R. Becker; Norman Levine

Temnopleuroida

An order of Echinacea with a camarodont lantern, smooth or sculptured test, tubercles imperforate or perforate (and usually crenulate), ambulacral plates of diademoid or echinoid type, and branchial slits which are usually shallow. The order includes a long phylogenetic series in which the original characters change considerably so that a concise but exact diagnosis is not possible. Following is an evolutionary summary of the three included families. (1) The Glyphocyphidae, known only from the Cretaceous and Eocene and probably ancestral to the other two families, were small forms with a sculptured test, perforate crenulate tubercles, and diademoid ambulacral plates. Their sculptured test links them with (2) the Temnopleuridae whose tubercles, however, are imperforate, though usually crenulate. This family arose in the Cretaceous and abounds today, especially in the tropics, on strandlines. Most of the order, so far, had shallow branchial slits, but transitional forms link them with (3) the Toxopneustidae, Tertiary and extant forms where the slits are deep and the sculpture tends to vanish. At the same time, the tubercles become imperforate and noncrenulate and the ambulacral plates change to the echinoid type. *See* ECHINACEA; ECHINODERMATA; ECHINOIDA; ECHINOIDEA.

Howard B. Fell

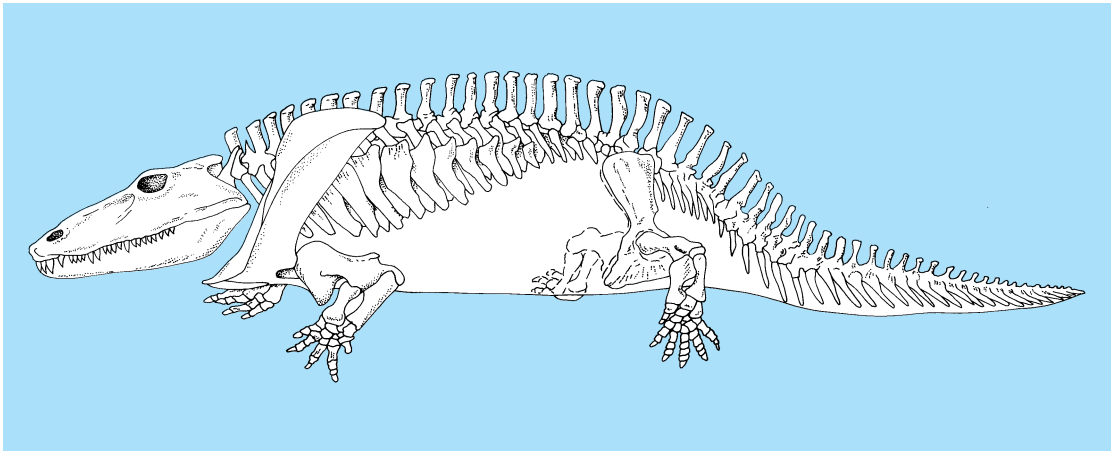
Temnospondyli

One of the largest recognized groups of early amphibians with about 180 described genera. The Temnospondyli first appear in the fossil record in the lower Carboniferous and reach substantial diversity in the later Carboniferous and Permian. The end-Permian extinction event wiped out several families, but the surviving lineages rediversified and temnospondyls remained numerous through the Triassic before declining in the later Mesozoic, the last survivor (*Koolasuchus*) being known from the mid-Cretaceous of Australia. *See* GEOLOGIC TIME SCALE.

General morphology. Earlier temnospondyls were superficially salamander-like with large flat heads, no necks, short limbs, and an undulating gait. In the context of other early amphibians, they are recognized by the large openings across the roof of the mouth (palatal vacuities) and by the semicircular embayments behind the cheek region for froglike large eardrums. They had short straight ribs which combined with the large skulls, suggest that they breathed by buccal pumping, ramming air into the lungs, and not using costal respiration. The individual vertebrae were of multipart construction with the body (centrum) made up of three separate bones, namely one large intercentrum anteriorly and two smaller pleurocentra posteriorly. The skull was massively constructed and survives intact in the fossil record even when the rest of the skeleton is lost.

Types and diversity. Carboniferous and Permian temnospondyls include a wide range of morphological types. As well as amphibious salamander-like forms, they evolved into larger superficially crocodile-like animals both amphibious and terrestrial (*Eryops*) [see **illustration**]. Others were specialized aquatic forms, retaining lateral-line canals to a large size and filling eel-like niches. Most were 30 cm (1 ft) to 1 m (3.3 ft) in length, with a few forms growing to 2 m (6 ft). Most of the Carboniferous and early Permian records are from North America and Europe (equatorial Euramerica in the late Paleozoic), but from the mid-Permian onward the group is found in all continents.

After the end-Permian extinction event, one subgroup of temnospondyls, the Stereospondyli, produced a new adaptive radiation of crocodile-like forms, mostly amphibious or aquatic and growing to 5 m (16.5 ft) [*Mastodonsaurus*]. In some of these



Common Permian temnospondyl *Eryops*; length is approximately 6 ft (2 m). (*After American Museum of Natural History*)

families, the vertebrae become further modified, with the main body of the vertebrae being reduced to a single large intercentrum. Many of the families forming this radiation were global in distribution. They include the Triassic trematosaurids, which are sometimes found in marine assemblages. At the end of the Triassic most temnospondyls vanish from the fossil record, but a few aquatic forms persisted in East Asia and Australia through the Jurassic and Early Cretaceous.

Relation to modern amphibians. The origin of the modern amphibian groups is controversial, but many workers consider the Temnospondyli to be the group from which they evolved. Two Carboniferous-Permian families—the Amphibamidae and the Branchiosauridae—comprise small temnospondyls which show resemblances to frogs and salamanders. The lower Permian amphibamid *Doleserpeton* has pedicellate bicuspid teeth like those of modern amphibians, and the branchiosaurid *Apateon* shows a pattern of skull bone development very like that in modern salamanders. *See* AMPHIBIA; ANTHRACOSAURIA; ICHTHYOSTEGA; LABYRINTHODONTIA; LISSAMPHIBIA.

Andrew Milner

Bibliography. J. A. Clack, *Gaining Ground*, Cambridge University Press, 2002; H. Heatwole and R. L. Carroll (eds.), *Amphibian Biology*, vol. 4, Surrey Beatty, Chipping Norton, NSW, Australia, 2000.

Temperature

A concept related to the flow of heat from one object or region of space to another. The term refers not only to the senses of hot and cold but to numerical scales and thermometers as well. Fundamental to the concept are the absolute scale and absolute zero and the relation of absolute temperatures to atomic and molecular motions.

Numbers for temperatures, such as 100°C and −15°F, have been used for only about 300 years. By the seventeenth century, science had developed to the point that, to fully describe the properties of matter, a numerical, quantitative scale of temperature differences was needed. For example, in 1756 Joseph Black in Scotland discovered that ice does not change temperature when it melts. Almost all substances so behave; also, the melting temperature depends on the purity of the substance. Thus one reason for devising a thermometer (literally, a meter for temperature) was that with it the composition of matter could be studied.

Temperature measurements are useful for studying molecular motions in material. **Figure 1** shows how the temperature of 1 g of H_2O, starting as ultracold ice, changes as heat is added at the constant rate of 1 cal/s (4.18 joules/s), assuming that no heat is lost to the surroundings. The graph plateaus illustrate Black's discovery that the temperature is constant during a phase change of solid to liquid or liquid to gas. *See* BOILING POINT; MELTING POINT.

Thermometers do not measure a special physical quantity. They measure length (as of a mercury column) or pressure or volume (with the gas thermometer at the National Bureau of Standards) or electrical voltage (with a thermocouple). The basic fact is that, if a mercury column has the same length when touching two different, separated objects, when the objects are placed in contact no heat will flow from one to the other. *See* THERMOMETER.

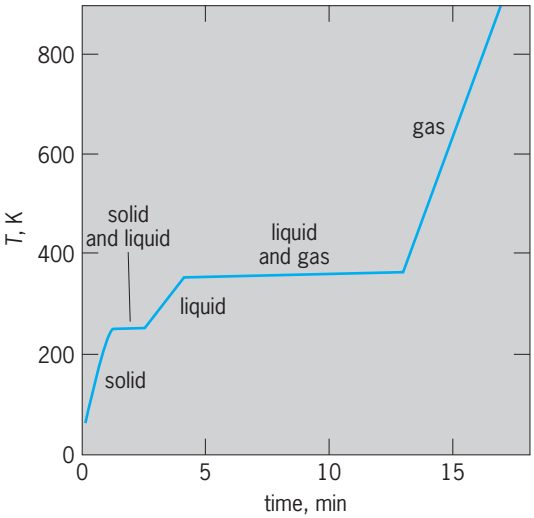


Fig. 1. Temperature of 1 g of H_2O, starting at 0 K, with a constant heat input of 1 cal/s (4.18 J/s). °F = (K × 1.8) − 459.67.

Empirical scales. The numbers on the thermometer scales are merely historical choices; they are not scientifically fundamental. The most widely used scales are the Fahrenheit (°F) and the Celsius (°C). The Centigrade scale with 0° assigned to ice water (ice point) and 100° assigned to water boiling under one atmosphere pressure (steam point) was formerly used, but it has been succeeded by the Celsius scale, defined in a different way than the Centigrade scale. However, on the Celsius scale the temperatures of the ice and steam points differ by only a few hundredths of a degree from 0° and 100°, respectively. **Figure 2** shows how the Celsius and Fahrenheit scales compare and how they fit onto the absolute scales. *See* ICE POINT.

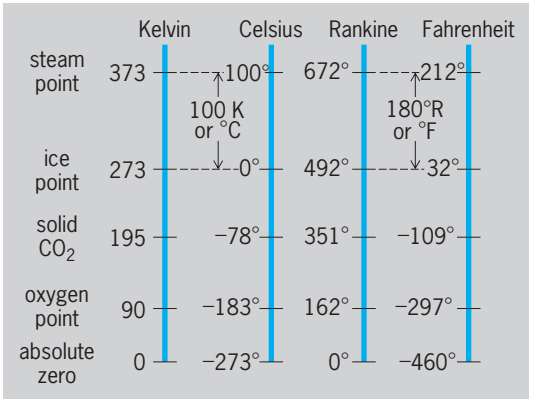


Fig. 2. Comparisons of Kelvin, Celsius, Rankine, and Fahrenheit temperature scales. Temperatures are rounded off to nearest degree. (*After M. W. Zemansky, Temperatures Very Low and Very High, Van Nostrand, 1964*)

These scales have one common value: $-40^\circ C = -40^\circ F$. This fact can be used to change a temperature from one scale to the other. Given a temperature in °C or in °F, add 40, multiply by 95 if converting from °C to °F or by $^5/_9$ if from °F to °C, then subtract 40. Example: Normal human body temperature is 98.6°F. To convert to °C, $98.6 + 40 = 138.6$; $138.6 \times {}^5/_9 = 77.0$; $77.0 - 40 = 37.0^\circ C$.

Absolute temperature scale. In 1848 William Thomson (Lord Kelvin), following ideas of Sadi Carnot, stated the concept of an absolute scale of temperature in terms of measuring amounts of heat flowing between objects. Most important, Kelvin conceived of a body which would not give up any heat and which was at an absolute zero of temperature. Experiments have shown that absolute zero corresponds to $-273.15^\circ C$ or $-459.7^\circ F$. Two absolute scales, shown in Fig. 2, are the Kelvin (K) and the Rankine (°R). *See* ABSOLUTE ZERO.

Interest in temperature and heat flow was stimulated in the early nineteenth century by efforts to improve the efficiency of steam engines. Out of this came the concept of a Carnot engine. This is not a real machine, but an imagined, ideal, frictionless system. A Carnot engine takes in heat Q_b from a higher temperature source at T_b (kelvins), does work W, and exhausts heat Q_l into a lower temperature source T_l. Two important deductions are: (1) The efficiency of the engine is $W/Q_b = 1 - T_l/T_b$. For example, a Carnot engine operated between the boiling point ($T_b = 373$ K) and the ice point ($T_l = 273$ K) of water has an efficiency of 0.268. A real engine would have an efficiency less than this, but the concept is nevertheless of great importance in engineering. (2) The ratio of temperatures equals the ratio of heats, namely, $T_b/T_l = Q_b/Q_l$. Lord Kelvin suggested that this be the basis for the absolute temperature scale. One special system (water at its triple point with ice, liquid, and water vapor present) is defined to have a particular value of absolute temperature (273.16 K). Any unknown temperature T_u, can be measured by operating a Carnot engine between 273.16 K and T_u, measuring the heats $Q_{273.16}$ and Q_u absorbed and rejected, and calculating $T_u = 273.16$ K $(Q_u/Q_{273.16})$. *See* CARNOT CYCLE; THERMODYNAMIC PRINCIPLES; TRIPLE POINT.

In practice, absolute temperatures are not measured this way. Instead low-density helium gas and dilute paramagnetic crystals, the most nearly ideal of real materials, allow measurement of temperatures virtually identical with those defined by a Carnot process. The advantages are that gas pressures and volumes and magnetic fields and magnetizations can be measured more conveniently and accurately than heat flows. *See* GAS THERMOMETRY.

The measurement of a single temperature with a gas or magnetic thermometer is a major scientific event done at a national standards laboratory. Only a few temperatures have been measured, including the freezing point of gold (1337.91 K or 1948.57°F), and the boiling points of sulfur (717.85 K or 832.46°F), oxygen (90.18 K or $-297.35^\circ F$), and helium (4.22 K or $-452.07^\circ F$). Various other types of thermometers (platinum, carbon, and doped germanium resistors; thermocouples) are calibrated at these temperatures and used to measure intermediate temperatures. *See* TEMPERATURE MEASUREMENT.

Kinetic temperature. An important aspect of the absolute temperature scale is its relation to the motions of atoms and molecules, whether vibrations as in solids and liquids, or straight path flights with collisions as in gases. There are two important facts here: (1) There is a definite distribution of motions. For example, in a gas, even though the motions are chaotic and a particular molecule changes velocity after each collision with another molecule, at any instant a definite number of molecules have a particular velocity. It cannot be said that the gas is at a definite temperature unless the molecules have this definite distribution of velocities, although different small portions of the gas may have definite, though different, temperatures. The same idea holds for the distribution of vibration frequencies in solids and liquids. (2) A body has a minimum amount of motion energy. It was supposed in the nineteenth century that this minimum was zero energy, but modern theories and experiments show that the minimum is greater than zero. A body in its lowest energy state cannot give out heat and is at absolute zero.

A system may have several degrees of freedom. The molecules of a gas, besides having straight-line motions, may rotate and vibrate and their electrons may be in different energy levels. When a system is in equilibrium, the energies stored in these different degrees of freedom are related to a common absolute temperature. In several cases, one can measure something related to a particular degree of freedom of a system. Then if the system is in equilibrium, its absolute temperature can be inferred.

Examples of this are measurements of temperatures in the Sun's corona and in the remote regions of the Milky Way Galaxy. The Sun's corona is so hot that atoms lose several electrons, that is, are multiply ionized. This greatly affects the wavelengths of light that these atoms emit. From the measured wavelengths that have been identified, the corona's temperature has been estimated at 2×10^6 K (3.6×10^6 °F).

Besides light, atoms and molecules can emit radio waves. The straight-line motions of atoms toward or away from Earth-based radio telescopes slightly shift the received wavelengths. From these Doppler shifts, temperatures from 1 to 100 K (-458 to $-280^\circ F$) have been found in the vast hydrogen atom clouds in the galaxy spiral arms. Rotations of OH molecules influence the populations of energy levels and thus affect the intensities of emitted radio signals. From these also, deep-space absolute temperatures are inferred. *See* KINETIC THEORY OF MATTER.

Negative absolute temperatures. Negative Celsius and Fahrenheit temperatures are readily accepted because 0° on these scales is arbitrarily set above absolute zero (Fig. 2). But the idea that a system could be at, say -50 K, was introduced only in the early 1950s. The concept applies only to systems with a finite number of energy levels, that is, those that can

store a finite amount of energy. Thus the translational energy of a gas or the vibration energy of a crystal cannot be at negative absolute temperatures. However, the energies of electron and nuclear magnetic moments (spins) in a magnetic field do have upper limits. Normally there are more spins in lower energy levels, in which case they are at positive absolute temperature. With special techniques one can arrange equal numbers of spins in the energy levels. Then the temperature is infinite kelvins. This does not mean infinite heat can be extracted from such a system. One can go further and cause there to be more spins in the higher energy levels than in the lower, and this situation is described as being at a negative absolute temperature.

The spin system is not in equilibrium with the crystal lattice vibrations, which remain at positive temperatures. But at temperatures around a few kelvins, the spins need minutes or hours to exchange heat with the lattice. Negative absolute temperatures are hotter than positive temperatures; this is reasonable since heat flows from the negative temperature spins to the positive temperature lattice. *See* NEGATIVE TEMPERATURE.

Extreme temperatures. For reasons involving both pure and applied science, researchers endeavor to achieve extremes of both low and high temperatures. At very low temperatures, phenomena that are well understood become frozen out and new predicted and unpredicted effects are sought. Experimenters have reported holding 4 lb (2 kg) of copper at about 50 microkelvins (9×10^{-5} °F above absolute zero) for a couple of days. Achieving and measuring such temperatures involve using the nuclear magnetic spin system. In the procedure to reach the 50-μK lattice temperature, the copper spins were brought down to 50 nanokelvins (9×10^{-8} °F above absolute zero). One object of such research is to see if nuclear spins will spontaneously align as electron spins do in ferromagnetic materials like iron. *See* LOW-TEMPERATURE PHYSICS; LOW-TEMPERATURE THERMOMETRY.

The highest equilibrium temperatures reached on Earth are around 10^7 K (2×10^7 °F) in experiments to achieve fusion of hydrogen nuclei. These temperatures have been sustained in very low-density gases like the Sun's corona for a few seconds. At these temperatures hydrogen nuclei have speeds of about 4.3×10^6 ft/s (1.3×10^6 m/s). Such speeds are necessary if the positively charged nuclei are to overcome their electric repulsion force. When the nuclei get close enough together, nuclear forces attract them to fuse and there is a net energy release. The goal of the research is to convert this energy into conventional electrical energy. *See* NUCLEAR FUSION.

Roland A. Hultsch

Bibliography. T. D. McGee, *Principles and Methods of Temperature Measurements*, 1988; B. W. Magum and G. T. Furukawa, *Guidelines for Realizing the Temperature Scale of 1990 (ITS-90)*, NIST Tech. Note 1265, 1990; K. Mendelssohn, *The Quest for Absolute Zero*, 2d ed., 1977; T. J. Quinn, *Temperature*, 2d ed., 1991; J. F. Schooley (ed.), *Temperature: Its Measurement and Control in Science and Industry*, vol. 5, 1982; M. W. Zemansky, *Temperatures Very Low and Very High*, 1964, reprint 1981.

Temperature adaptation

The ability of animals to survive and function at widely different temperatures is a result of specific physiological adaptations. Temperature is an all-pervasive attribute of the environment that limits the activity, distribution, and survival of animals. Ocean temperatures range from 28 to 86°F (−2 to 30°C), but considerably warmer temperatures are found near deep-sea hydrothermal vents (up to 662°F or 350°C), backwaters of desert streams (109°F or 43°C), and geothermal springs (194–212°F or 90–100°C). Air temperatures range from −94°F (−70°C) in polar regions to 176°F (80°C) at the desert surface. Although some bacteria and blue-green algae live at temperatures up to 230°F (110°C), life processes are generally restricted to the temperatures between 32 and 113°F (0 and 45°C), and most animals live within an even narrower range. Limits for reproduction and development are generally narrower than those for survival of adults.

Changes in temperature influence biological systems, both by determining the rate of chemical reactions and by specifying equilibria (**Fig. 1**). Because temperature exerts a greater effect upon the percentage of molecules that possess sufficient energy to react (that is, to exceed the activation energy) than upon the average kinetic energy of the system, modest reductions in temperature (for example, from 77 to 59°F or from 25 to 15°C, corresponding to only a 3% reduction in average kinetic energy) produce a marked depression (two- to threefold) in reaction rate. In addition, temperature specifies the equilibria between the formation and disruption of the noncovalent (electrostatic, hydrophobic, and hydrogen-bonding) interactions that stabilize both the higher levels of protein structure

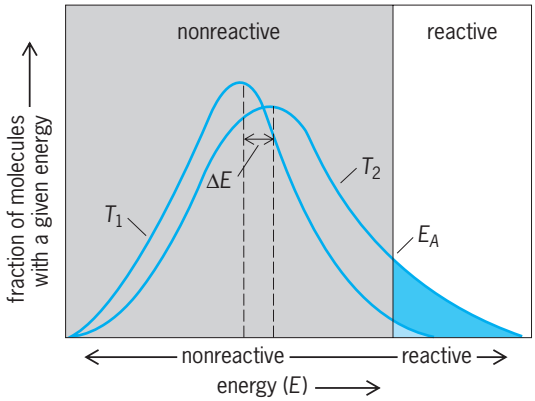


Fig. 1. Energy distribution profiles for a population of molecules at two different temperatures $T_2 > T_1$. The peaks in the two curves represent the average energy contents, and ΔE is the temperature-induced change in average energy content; E_A is the activation energy, and the shaded sections under the two curves show the change in the proportion of molecules with $E > E_A$.

and macromolecular aggregations such as biological membranes. Maintenance of an appropriate structural flexibility is a requirement for both enzyme catalysis and membrane function, yet cold temperatures constrain while warm temperatures relax the conformational flexibility of both proteins and membrane lipids, thereby perturbing biological function.

Animals are classified into two broad groups depending on the factors that determine body temperature. For ectotherms, body temperature is determined by sources of heat external to the body; levels of resting metabolism (and heat production) are low, and mechanisms for retaining heat are limited. Such animals are frequently termed poikilothermic or cold-blooded, because the body temperature often conforms to the temperature of the environment. In contrast, endotherms produce more metabolic heat and possess specialized mechanisms for heat retention. Therefore, body temperature is elevated above ambient temperature; some endotherms (termed homeotherms or warm-blooded animals) maintain a relatively constant body temperature. There is no natural taxonomic division between ecto- and endotherms. Most invertebrates, fish, amphibians, and reptiles are ectotherms, while true homeothermy is restricted to birds and mammals. However, flying insects commonly elevate the temperature of their thoracic musculature prior to and during flight (to 96°F or 36°C), and several species of tuna retain metabolic heat in their locomotory musculature via a vascular countercurrent heat exchanger. *See* THERMOREGULATION.

Endotherms. Homeotherms, by defending a constant body temperature, circumvent the problems associated with the maintenance of physiological function under varying body temperatures. The ability to regulate body temperature is dependent upon a suite of physiological adaptations involving the management of heat production, the distribution of heat within the body, and the exchange of heat with the environment (**Fig. 2**). When stressed by cold, homeotherms maintain body temperature by both shivering and nonshivering thermogenesis, that is, the production of heat by processes that do not involve muscle contraction. Both processes are strictly controlled, and the regulated heat production is just sufficient to maintain body temperature. Nonshivering thermogenesis is of particular interest, because the principal site of heat production occurs in brown adipose tissue, a tissue whose sole function is thermogenesis. Brown adipose tissue is a particularly important source of heat in young or cold-acclimated mammals and in arousing hibernators. Conversely, when stressed by heat, homeotherms lose heat to their environment by the evaporation of water from some body surface. In cattle, horses, and humans, high rates of evaporation are accomplished by sweating, whereas in species that do not sweat (dogs and cats), panting occurs. Insulation, in the form of subcutaneous adipose tissue and external pelage, is an adaptation to reduce the cost of thermoregulation in cold environments. Circulatory adaptations permitting the redistribution of blood flow are also important to thermoregulation. When body temperature rises, large volumes of blood are shunted through skin capillaries to increase heat transfer to the environment; conversely, peripheral blood flow is reduced in response to hypothermia in order to curtail heat loss. *See* ADIPOSE TISSUE; HIBERNATION AND ESTIVATION.

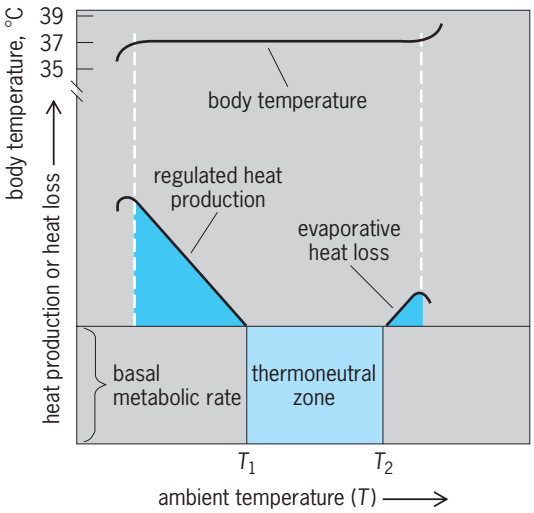


Fig. 2. Relationship between heat production and heat loss as a function of ambient temperature for a homeotherm. Thermoregulation can be achieved at minimal metabolic cost in the thermoneutral zone (defined by T_1 and T_2). Broken lines define the limits of ambient temperature over which a homeotherm can successfully defend (or maintain) its body temperature. °F = (°C × 1.8) + 32.

Ectotherms. Many ectotherms are essentially isothermal with their environment; even the muscles of actively swimming fish are within 1.8°F (1°C) of the water temperature. Consequently, geographic, seasonal, or diurnal fluctuations in temperature pose particular problems for the maintenance of physiological function in these animals. Yet, although metabolic rates are depressed by acute exposure to cold, many ectotherms (but not all—some ectotherms become torpid at low temperature as a means of conserving energy) remain active in the cold and exhibit similar levels of activity at seasonal extremes of temperature. Similarly, arctic and antarctic ectotherms perform as effectively near 32°F (0°C) as tropical species do at 86°F (30°C). Maintenance of similar rates of activity at widely different body temperatures is a clear indication that ectotherms can adapt to temperature extremes. At the cellular level, thermal adaptations are most commonly reflected in the lipid composition of cell membranes and the catalytic properties of enzymes.

Lipid-mediated adaptations. Membranes perform many vital cell functions, including regulating the exchange of material between the cell and its environment; storing energy in the form of transmembrane ion gradients; providing a matrix in which many metabolic processes are organized; and controlling the flow of information between the cell

and its surroundings by generating intracellular messengers in response to extracellular signals. Yet, the physical properties of phospholipids—the primary structural elements of biological membranes—are markedly temperature-dependent. With cooling, the acyl domain of phospholipids is transformed from a fluid to a gel phase. Such phase transitions significantly perturb membrane function, for in the gel phase the membrane is rigid, passive permeability is reduced, and the activity of membrane-associated enzymes declines. Conversely, at elevated temperatures membranes become hyperfluid and leaky to ions (loss of potassium from muscle cells is a contributing factor in heat death). *See* CELL MEMBRANES; OSMOREGULATORY MECHANISMS.

Ectotherms overcome these problems by restructuring their membranes so that lipids of appropriate physical properties are matched to the prevailing ambient temperature. As temperature drops, high-melting lipids are replaced by lower-melting ones; consequently, membranes remain fluid at cold temperatures. The melting point of membrane lipids is lowered primarily by increasing the degree of acyl chain unsaturation, which introduces a kink into the acyl chain and prevents close packing at low temperatures. Two metabolic adjustments contribute to this restructuring process: increased activities of acyl chain desaturases at cold temperatures, and the operation of a deacylation–reacylation cycle, which permits the acyl chain composition to be altered independently of the rest of the phospholipid molecule. In addition, phospholipids with small, as opposed to bulky, head groups also increase in abundance at low temperatures, and the resulting rise in the ratio of conically to cylindrically shaped lipids may offset the direct effects of temperature change upon lipid packing (that is, it may disrupt packing at low temperature). Differences in lipid composition between polar and tropical species resemble those between seasonally adapted individuals of temperate species.

As a consequence of lipid restructuring, membrane fluidity is relatively constant when compared at the respective growth temperatures to which an animal has become adapted, even though varying markedly with acute changes in temperature. This phenomenon is known as homeoviscous adaptation. *See* LIPID.

Protein-mediated adaptations. In addition to the homeoviscous regulation of membrane fluidity, ectoderms display other adaptations that permit function over a broad temperature range. These range from evolutionally fixed differences in the structure and function of specific proteins to seasonal adjustments in the rates and patterns of energy metabolism.

Enzyme structure and function vary interspecifically in a manner consistent with the conservation of catalytic rates and regulatory properties at different temperatures. The catalytic efficiency of enzymes is inversely correlated with habitat or cell temperature. For example, lactate dehydrogenase from an antarctic fish produces nearly twice as much product per minute per mole of enzyme as does the homologous

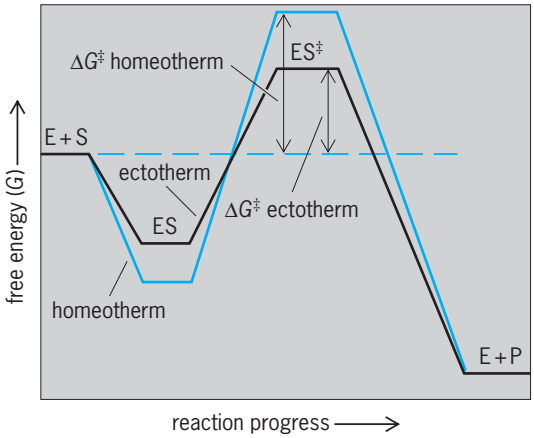


Fig. 3. Energy profiles for the reactions catalyzed by homologous enzymes from an ectotherm and a homeotherm. Note that the activation free-energy barrier is higher for the homeothermic than the ectothermic enzyme because of the tighter binding of substrate by the former. E = free enzyme; S = free substrate; P = free product; ES = enzyme substrate complex, not in the activated state; $ES^{\ddagger}$ = enzyme substrate complex in the activated state; $\Delta G^{\ddagger}$ = free energy of activation. (*After P. W. Hochachka and G. N. Somero, Biochemical Adaptation, Princeton University Press, 1984*)

enzyme from rabbit muscle when compared at 41°F (5°C). Such increased catalytic efficiencies typical of ectothermic enzymes are a reflection of a lowered activation energy. It has been postulated that differences in catalytic efficiency between homologous enzymes of ectotherms and endotherms reflect varying amounts of weak bond formation between the enzyme and substrate during the activation step of catalysis. The disruption of relatively few weak bonds in the formation of products by the enzymes of cold-adapted ectotherms would keep the free energy of the reaction low and the rate relatively high, thus providing an important mechanism for compensation of metabolic rates at low temperatures (**Fig. 3**). This hypothesis is substantiated by the observation that enzymes of homeotherms bind substrates as strongly at 98.6°F (37°C) as do those of antarctic species at 32°F (0°C). Such interspecific differences in enzyme function are presumed to reflect genetically fixed differences in primary structure.

In contrast to the evolutionary tailoring of enzymes to specific thermal environments, there is little evidence that seasonal acclimatization results in the production of environment-specific isozymes best suited to function at either warm or cold temperatures. Instead, seasonal acclimation generally results in altered levels of enzyme activity, which are presumed to reflect temperature-dependent differences in cellular enzyme content. In contrast to qualitative changes in the efficiency of an enzyme, this quantitative adaptation simply involves altered amounts of an identical enzyme at seasonal extremes of temperature. Acyl-chain desaturating enzymes are induced in cold-acclimated ectotherms; in addition, the activities of enzymes of aerobic (mitochondrial) metabolism are generally elevated at cold temperatures, providing an explanation for the thermal

compensation of metabolic rate. *See* CATALYSIS; COLD HARDINESS (PLANT); ENZYME. Jeffrey R. Hazel

Bibliography. R. C. Aloia, C. C. Curtain, and L. M. Gordon (eds.), *Physiological Regulation of Membrane Fluidity*, 1988; A. R. Cossins and K. Bowler, *Temperature Biology of Animals*, 1987; J. R. Hazel and C. L. Prosser, Molecular mechanisms of temperature compensation of poikilotherms, *Physiol. Rev.*, 54(3):620–677, 1974; P. W. Hochachka and G. N. Somero, *Biochemical Adaptation*, 1984; C. L. Prosser, *Adaptational Biology: Molecules to Organisms*, 1986.

Temperature inversion

The increase of air temperature with height; an atmospheric layer in which the upper portion is warmer than the lower. Such an increase is opposite, or inverse, to the usual decrease of temperature with height, or lapse rate, in the troposphere of about 3.3°F/1000 ft (6.5°C/km) and somewhat less on mountain slopes. However, above the tropopause, temperature increases with height throughout the stratosphere, decreases in the mesosphere, and increases again in the thermosphere. Thus inversion conditions prevail throughout much of the atmosphere much or all of the time, and are not unusual or abnormal. *See* AIR TEMPERATURE; ATMOSPHERE.

Inversions are created by radiative cooling of a lower layer, by subsidence heating of an upper layer, or by advection of warm air over cooler air or of cool air under warmer air. Outgoing radiation, especially at night, cools the Earth's surface, which in turn cools the lowermost air layers, creating a nocturnal surface inversion a few centimeters to several hundred meters thick. Over polar snowfields, inversions may be a kilometer or more thick, with differences of 54°F (30°C) or more. Solar warming of a dust layer can create an inversion below it, and radiative cooling of a dust layer or cloud top can create an inversion above it. Sinking air warms at the dry adiabatic lapse of 5°F/1000 ft (10°C/km), and can create a layer warmer than that below the subsiding air. Air blown onto cool water from warmer land or onto snow-covered land from warmer water can cause a pronounced inversion that persists as long as the flow continues. Warm air advected above a colder layer, especially one trapped in a valley, may create an intense and persistent inversion.

Inversions effectively suppress vertical air movement, so that smokes and other atmospheric contaminants cannot rise out of the lower layer of air. California smog is trapped under an extensive subsidence inversion; surface radiation inversions, intensified by warm air advection aloft, can create serious pollution problems in valleys throughout the world; radiation and subsidence inversions, when horizontal air motion is sluggish, create widespread pollution potential, especially in autumn over North America and Europe. *See* AIR POLLUTION; SMOG.

Arnold Court

Temperature measurement

Measurement of the hotness of a body relative to a standard scale. The fundamental scale of temperature is the thermodynamic scale, which can be derived from any equation expressing the second law of thermodynamics. Efforts to approximate the thermodynamic scale as closely as possible depend on relating measurements of temperature-dependent physical properties of systems to thermodynamic relations expressed by statistical thermodynamic equations, thus in general linking temperature to the average kinetic energy of the measured system. Temperature-measuring devices, thermometers, are systems with properties that change with temperature in a simple, predictable, reproducible manner. *See* TEMPERATURE; THERMODYNAMIC PRINCIPLES.

Temperature scale. In the establishment of a useful standard scale, assigned temperature values of thermodynamic equilibrium fixed points are agreed upon by an international body (the General Conference of Weights and Measures), which updates the scale about once every 20 years. Thermometers for interpolating between fixed points and methods for realizing the fixed points are prescribed, providing a scheme for calibrating thermometers used in science and industry.

The scale now in use is the International Temperature Scale of 1990 (ITS-90). Its unit is the kelvin, K, arbitrarily defined to be 1/273.16 of the thermodynamic temperature T of the triple point of water (the state in which the liquid, solid, and vapor phases coexist). The scale extends upward from 0.65 K. For temperatures above 273.15 K, it is common to use International Celsius Temperatures, t_{90} (rather than International Kelvin Temperatures, T_{90}), having the unit degree Celsius, with symbol °C. The degree Celsius has the same magnitude as the kelvin. Temperatures, t_{90} are defined as $t_{90}/^\circ\mathrm{C} = T_{90}/\mathrm{K} - 273.15$, that is, as differences from the ice-point temperature at 273.15 K. The ice point is the state in which the liquid and solid phases of water coexist at a pressure of 1 atm (101,325 pascals). [The Fahrenheit scale, with symbol °F, still in common use in the United States, is given by $t_F/^\circ\mathrm{F} = t_{90}/^\circ\mathrm{C} \times 1.8) + 32$, or $t_F/^\circ\mathrm{F} = (T_{90}/\mathrm{K} \times 1.8) - 459.67$.] The ITS-90 is defined by 17 fixed points (**Table 1**). Between 0.65 and 5.0 K, the ITS-90 is defined in terms of the vapor pressure of ^{3}He and ^{4}He; between 3.0 and 24.5561 K, by interpolating constant-volume gas thermometry (using either ^{3}He or ^{4}He); between 13.8033 K and 961.78°C, by platinum resistance thermometry; and above 961.78°C, by radiation thermometry. There are overlapping ranges of vapor pressure and interpolating constant-volume gas thermometry, and of interpolating constant-volume gas thermometry and platinum resistance thermometry, with the different definitions having equal status. For radiation thermometry, based on Planck's radiation formula, the silver, gold, or copper point may be used as the reference temperature. Below 0.65 K, no internationally agreed upon scale exists, but the scale may be extended into this range in the future. At present,

TABLE 1. Defining fixed points of the International Temperature Scale of 1990 (ITS-90)

Equilibrium state*	Temperature	
	T_{90}, K	t_{90}, °C
Vapor pressure equation of helium	3 to 5	−270.15 to −268.15
Triple point of equilibrium hydrogen†	13.8033	−259.3467
Vapor pressure point of equilibrium hydrogen† (or constant volume gas thermometer point of helium)	≈17	≈−256.15
Vapor pressure point of equilibrium hydrogen† (or constant volume gas thermometer point of helium)	≈20.3	≈−252.85
Triple point of neon	24.5561	− 248.5939
Triple point of oxygen	54.3584	−218.7916
Triple point of argon	83.8058	−189.3442
Triple point of mercury	234.3156	−38.8344
Triple point of water	273.16	0.01
Melting point of gallium	302.9146	29.7646
Freezing point of indium	429.7485	156.5985
Freezing point of tin	505.078	231.928
Freezing point of zinc	692.677	419.527
Freezing point of aluminum	933.473	660.323
Freezing point of silver	1234.93	961.78
Freezing point of gold	1337.33	1064.18
Freezing point of copper	1357.77	1084.62

*The triple point is the equilibrium temperature at which the solid, liquid, and vapor phases coexist. The freezing point and the melting point are the equilibrium temperatures at which the solid and liquid phases coexist under a pressure of 101,325 Pa, 1 standard atmosphere. The isotopic composition is that naturally occurring.

† Equilibrium hydrogen is hydrogen with the equilibrium distribution of its ortho and para states at the corresponding temperatures. Normal hydrogen at room temperature contains 25% para and 75% ortho hydrogen.

temperatures below 0.65 K are determined by magnetic thermometry, nuclear orientation thermometry, and noise thermometry. In some cases, a ^{3}He melting-curve thermometer is used. *See* LOW-TEMPERATURE THERMOMETRY.

Primary thermometers. These are devices which relate the thermodynamic temperature to statistical mechanical formulations incorporating the Boltzmann constant k_B or the Boltzmann factor $\exp(-E_i/k_BT)$, where E_i is the energy of the ith state of the system (**Table 2**). However, the fixed points of at and below 419.527°C of the ITS-90 are all based on one or more types of gas thermometry, with those above 419.527°C being determined by spectral radiation pyrometry referenced to gas thermometry at a temperature near 460°C. *See* GAS THERMOMETRY.

Secondary thermometers. These are used as reference standards in the laboratory because primary thermometers are often too cumbersome. Since this is especially true for realization of the ITS-90 below the silver point, it is necessary to establish standard or secondary thermometers referenced to one or more fixed points for interpolation between fixed points. Over the range 13.8033 K to 961.78°C, the platinum resistance thermometer with specified characteristics is the prescribed standard; above 961.78°C, optical pyrometry and spectroscopic methods are used. In the low-temperature range, the rhodium-iron resistance thermometer is an excellent candidate for interpolation device between about an 0.3 and 35 K, to overlap the platinum thermometer, the vapor-pressure scales of the helium isotopes, and the interpolating constant-volume gas thermometer. The germanium resistance thermometer also is a suitable interpolating device between about 0.05 and 30 K. *See* PYROMETER; THERMOCOUPLE; THERMOMETER.

Lower-order thermometers. These are used for most practical purposes and, when fairly high accuracy is required, can usually be calibrated against reference standards maintained at national standards laboratories, such as the U.S. National Institute of Standards and Technology, or against portable reference devices (sealed freezing or melting point cells). Examples of these practical thermometers include wide varieties of resistance thermometers (employing pure metals or semiconductors, the latter including carbon thermometers, germanium thermometers, and thermistors), thermoelectric thermometers (thermocouples), liquid-in-glass thermometers, vapor-pressure thermometers, magnetic thermometers, and capacitance thermometers. To obtain the highest accuracy when a calibrated thermometer is put into use, the conditions must reproduce those established during the calibration. Special care must be taken to ensure that the thermometer is in good thermal contact with the body whose temperature is to be measured; that sufficient time is allowed before measurement is made for the thermometer to equilibrate its temperature with that of the body; and that extraneous heat leaks are eliminated. *See* THERMISTOR.

Temperature indicators. Often it is useful to ascertain when or if a certain temperature has been reached or exceeded. Such qualitative information can be acquired by the use of temperature indicators, which are either irreversible or reversible. Examples of irreversible indicators are pyrometric ceramic cones, which can be formulated to soften and bend over a specific range of temperatures, and metallic

TABLE 2. Primary thermometry methods

Method	Approximate useful range of T, K	Principal measured variables	Relation of measured variables to T	Remarks
Gas thermometry	1.3–950	Pressure P and volume V	Ideal gas law plus correction: $PV \propto k_B T$ plus corrections	Careful determination of corrections necessary, but capable of high accuracy
Acoustic interferometry	1.5–3000	Speed of sound W	$W^2 \propto k_B T$ plus corrections	
Magnetic thermometry		Magnetic susceptibility	Curie's law plus corrections: $\chi \propto 1/k_B T$ plus corrections	
1. Electron paramagnetism	0.001–35			
2. Nuclear paramagnetism	0.000001–1			
Gamma-ray anisotropy or nuclear orientation thermometry	0.01–1	Spatial distribution of gamma-ray emission	Spatial distribution related to Boltzmann factor for nuclear spin states	Useful standard for $T < 1$ K
Thermal electric noise thermometry		Mean square voltage fluctuation $\overline{V}^2$	Nyquist's law: $\overline{V}^2 \propto k_B T$	Other sources of noise serious problem for $T > 4$ K
1. Josephson junction point contact	0.001–1			
2. Conventional amplifier	4–1400			
Radiation thermometry (visual, photoelectric, or, photodiode)	500–50,000	Spectral intensity J at wavelength λ	Planck's radiation law, related to Boltzmann factor for radiation quanta	Needs blackbody conditions or well-defined emittance
Infrared spectroscopy	100–1500	Intensity I of rotational lines of light molecules	Boltzmann factor for rotational levels related to I	Also Doppler line broadening ($\propto\sqrt{k_B T}$) useful; principal applications to plasmas and astrophysical observations; proper sampling, lack of equilibrium, atmospheric absorption often problems
Ultraviolet and x-ray spectroscopy	5000–2,000,000	Emission spectra from ionized atoms—H,He, Fe, Ca, and so on	Boltzmann factor for electron states related to band structure and line density	

pellets or notched rods, which melt or deform at known temperatures. Examples of reversible indicators are certain paints or crayons and liquid-crystal materials, which change color, sometimes exceedingly sharply, at predeterminable temperatures. *See* LIQUID CRYSTALS.

B. W. Mangum

Bibliography. R. P. Benedict, *Fundamentals of Temperature, Pressure and Flow Measurements*, 3d ed., 1984; J. R. Leigh, *Temperature Measurement and Control*, 1988; T. D. McGee, *Principles and Methods of Temperature Measurement*, 1988; T. J. Quinn, *Temperature*, 2d ed., 1990; F. Schooley, *Thermometry*, 1988.

Tempering

The reheating of previously quenched alloy to a predetermined temperature below the critical range, holding the alloy for a specified time at that temperature, and then cooling it at a controlled rate, usually by immediate rapid quenching, to room temperature. Tempering has a long history, beginning as an empirical process to impart toughness. Consequently, the term is broadly applied to any process that toughens a material. *See* ALLOY.

In alloys, if the composition is such that cooling produces a supersaturated solid solution, the resulting material is brittle. Heating the alloy to a temperature only high enough to allow the excess solute to precipitate out and then rapidly cooling the saturated solution fast enough to prevent further precipitation of grain growth result in a microstructure combining hardness and toughness. *See* SOLID SOLUTION.

With steel, precipitation of carbide out of the supersaturated unstable martensite or interstitial solution of carbon in iron results in improved ductility. To be most successful, the tempering must be carried out by slow heating to avoid steep temperature gradients, stress relief being one of the objectives. Properties produced by tempering depend on the temperature to which the steel is raised and on its alloy composition. For example, if hardness is to be retained, molybdenum or tungsten is used in the alloy.

For tool steel, to obtain great hardness, the finished or nearly finished tool may be heated to a critical temperature dependent on the particular alloy. While at this temperature below the transformation range, the steel is refined; that is, it forms a solid solution with small grain size. After being held at the refining temperature long enough to reach thermal equilibrium throughout, the steel is rapidly cooled or quenched so as to pass quickly through temperature regions in which grain growth can occur. For steel

with 0.83% carbon the refining temperature is slightly above 1290°F (700°C); the faster this alloy is cooled the harder it becomes. *See* HEAT TREATMENT (METALLURGY). Frank H. Rockett

Bibliography. American Society for Testing and Materials, *Temper Embrittlement of Alloy Steels*, 1972; L. Coudurier, D. W. Hopkins, and I. Wilkomirsky, *Fundamentals of Metallurgical Processes*, 1985; J. V. Woodworth, *Steel Hardening, Tempering, and Forging Made Easy*, 1991.

Temporary structure (engineering)

A structure erected to aid in the construction of a permanent project. Temporary structures are used to facilitate the construction of buildings, bridges, tunnels, and other above- and below-ground facilities by providing access, support, and protection for the facility under construction, as well as assuring the safety of the workers and the public. Temporary structures either are dismantled and removed when the permanent works become self-supporting or completed, or are incorporated into the finished work. Temporary structures are also used in inspection, repair, and maintenance work.

There are many types of temporary structures, including cofferdams; earth-retaining structures; tunneling supports; underpinning; diaphragm/slurry walls; roadway decking; construction ramps, runways, and platforms; scaffolding; shoring; falsework; concrete formwork; bracing and guying; site protection structures such as sidewalk bridges, boards and nets for protection against falling objects, barricades and fences, and signs; and all sorts of unique structures that are specially conceived, designed, and erected to aid in a specific construction operation.

These temporary works have a primary influence on the quality, safety, speed, and profitability of all construction projects. More failures occur during construction than during the lifetimes of structures, and most of those construction failures involve temporary structures. However, codes and standards do not provide the same scrutiny as they do for permanent structures. Typical design and construction techniques and some industry practices are well established, but responsibilities and liabilities remain complex and present many contractual and legal pitfalls.

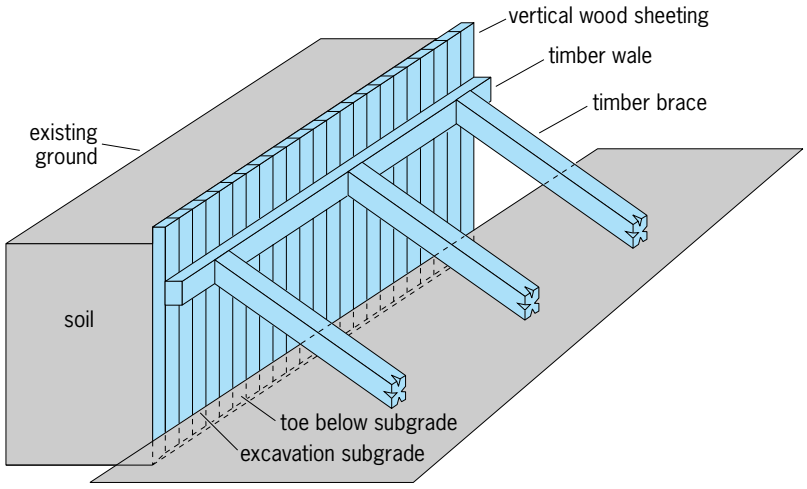


Fig. 1. Excavation sheeting and bracing. (*After R. T. Ratay, Handbook of Temporary Structures in Construction, 2d ed., McGraw-Hill, 1996*)

Cofferdams. Cofferdams are temporary enclosure walls designed to keep water and soil out of the work area, thus allowing the construction of the permanent facility in a dry environment. They enable the construction of bridge piers, intake structures, pump houses, locks, and dams below the surrounding water level by encircling the work area with an impervious wall. After the cofferdam is constructed, the area within it is pumped out to provide the dry construction site. Pumping usually continues at a moderate rate during the construction activities to remove any seepage from the surrounding body of water. *See* COFFERDAM.

Earth-retaining structures. The primary purpose of temporary earth-retaining structures is to permit the safe digging of a hole or trench in the ground and to provide a safe work area below the ground level. Another purpose is the mitigation of subsidence of the surrounding ground and settlement of nearby structures, pavements, and utilities. There are several common types of temporary earth retaining structures, including vertical wood sheeting and bracing, steel sheet piling, and soldier piles and horizontal sheeting. *See* RETAINING WALL.

Vertical wood sheeting and bracing. This type is the oldest and, under certain conditions, the most economical method of retaining earth. Its most common uses are in trench excavations, where the sheeting is made up of vertical planks, 2 to 3 in. (5 to 7.5 cm) thick, that are lowered or driven down and held in position against the soil by one or more lines of heavy horizontal timber wales (construction members), which in turn are pressed against the sheeting by horizontal or inclined bracing struts (**Fig. 1**). Successive levels of sheeting and bracing are installed as the excavation progresses downward.

Steel sheet piling. This type consists of a row of interlocking sheet piles, installed prior to excavation by driving the sheet piles one after another into the soil to a depth several feet below the bottom of the future excavation so as to provide embedment for the sheet piling to act as a cantilever retaining wall. Variations on the cantilever type are the braced and tied-back systems where, in addition to the anchorage by embedment at the bottom, the sheet piling is supported at one or more levels along its height. *See* CANTILEVER.

Soldier piles and horizontal sheeting. This system includes vertical steel beams, spaced 6 to 10 ft (1.8 to 3 m) apart, that act as vertical cantilevers supporting the sheeting panels that hold back the soil. The system is installed by driving the soldier piles (also known as soldier beams) to a depth several feet below the bottom of the future excavation, so as to provide embedment for their cantilever action, and then placing sheeting planks to span horizontally between soldiers as the excavation progresses. Variations on the cantilever type are the braced and tied-back systems where, in addition to the anchorage by embedment

at their bottoms, the solider piles are supported at one or more levels along their heights.

Tunneling supports. The basic types of temporary supports used in the construction of tunnels and shafts include steel liner plates; for soft ground, steel ribs and lagging; and for rock tunneling, rock bolting and grouting. There are a number of industry standards with many variations to suit given conditions. In most cases, temporary supports used during construction of tunnels and shafts remain in place after construction and become parts of the permanent supports. *See* TUNNEL.

Underpinning. Underpinning is a temporary or permanent support installed below an existing foundation to provide greater depth, a different load path, or increased bearing capacity, necessitated by some desired alteration of a foundation. The purpose is to relieve the foundation load at a particular location, transfer it to a temporary support, and transmit the load to the soil at a lower elevation. Common underpinning methods use pits and piers or driven or augered piles. Typically, a pit is dug below a part of the foundation, and a pier or pile is installed to support the foundation at that location. The load is transferred to the new support by drypacking or wedging between the bottom of the existing foundation and the top of the new pier. Excavation sheeting and concrete formwork, when used for constructing the piers, may be left or removed. The process is repeated under an adjacent area until the entire length of the existing foundation is underpinned.

Diaphragm/slurry walls. This term refers to a reinforced concrete wall constructed below ground level and utilizing the slurry method of trench stabilization. A trench with the width of the intended thickness of the concrete wall is excavated. As the digging progresses, the excavated material is immediately and continually replaced by a bentonite slurry mix. The lateral pressure created by the slurry prevents the walls of the trench from collapsing. After the trench is excavated to its final elevation, caged reinforcing steel is lowered through the slurry. The concrete, placed by the tremie method, fills the trench from the bottom up, displacing the lighter-weight bentonite slurry. Once the concrete cures and the wall attains its intended strength, the soil is excavated from one side, making the wall a temporary, and then later a permanent, retaining or foundation wall.

Roadway decking. Temporary roadway decking is the support and travel surface constructed (within the limits of streets and highways) to allow traffic over open-cut or belowground work areas. Once excavation within the roadway is completed, a support structure is constructed on which the surface decking is placed (**Fig. 2**). An exception is the simplest form of temporary decking: a very heavy steel plate placed over a relatively small excavation, set directly on the pavement without structural supports. The most common material for decking is heavy timber beams, similar in appearance to railroad ties, followed by heavy precast concrete slabs.

Construction ramps, runways, and platforms. Construction ramps, runways, and platforms provide vehicular access to and within a construction site. A construction ramp is utilized to provide access between the foundation or basement and the street level, or between two levels of the project. A runway is a strip of level roadway that provides a track for wheeled vehicles; it is usually a narrow extension into the construction site. A platform is a raised horizontal surface used by cranes, vehicles, equipment, and workers for convenient positioning. These facilities are constructed of earth, timber, steel, precast concrete, or combinations thereof.

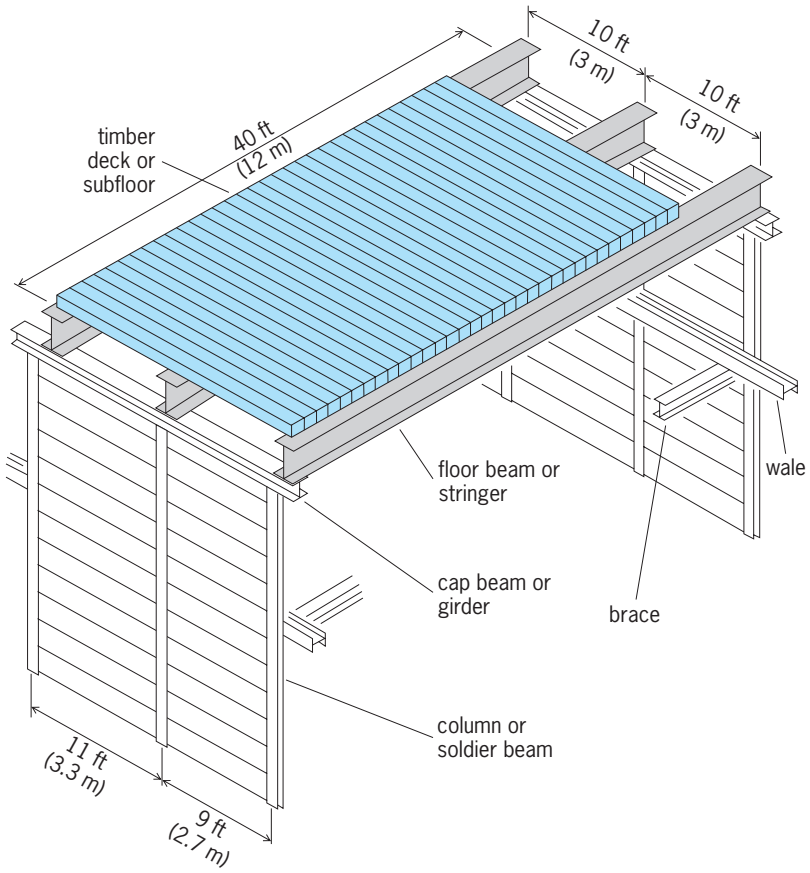


Fig. 2. Typical roadway decking with steel beams and timber decking. (*After R. T. Ratay, Handbook of Temporary Structures in Construction, 2d ed., McGraw-Hill, 1996*)

Scaffolding. Scaffolding is a structure used to provide access to a location of work that is too high for workers to reach. Scaffolds are made of aluminum or steel tubes assembled into structural systems, with wood or aluminum planking as the working platform. (In some Asian countries bamboo rods tied by lashings are used instead of metal tubes.) The two common types are the tube-and-coupler and the sectional scaffolds that can be assembled to hundreds of feet in height.

Tube-and-coupler. These scaffolds are assembled at the time and place of their use from five basic structural elements: the vertical posts, which rise from the ground; the horizontal bearers, which attach to the posts and support the working platforms; the runners, which attach to the posts directly below the bearers and provide longitudinal connections along the length of the scaffold; the diagonal bracing connected to the posts to provide stability in the

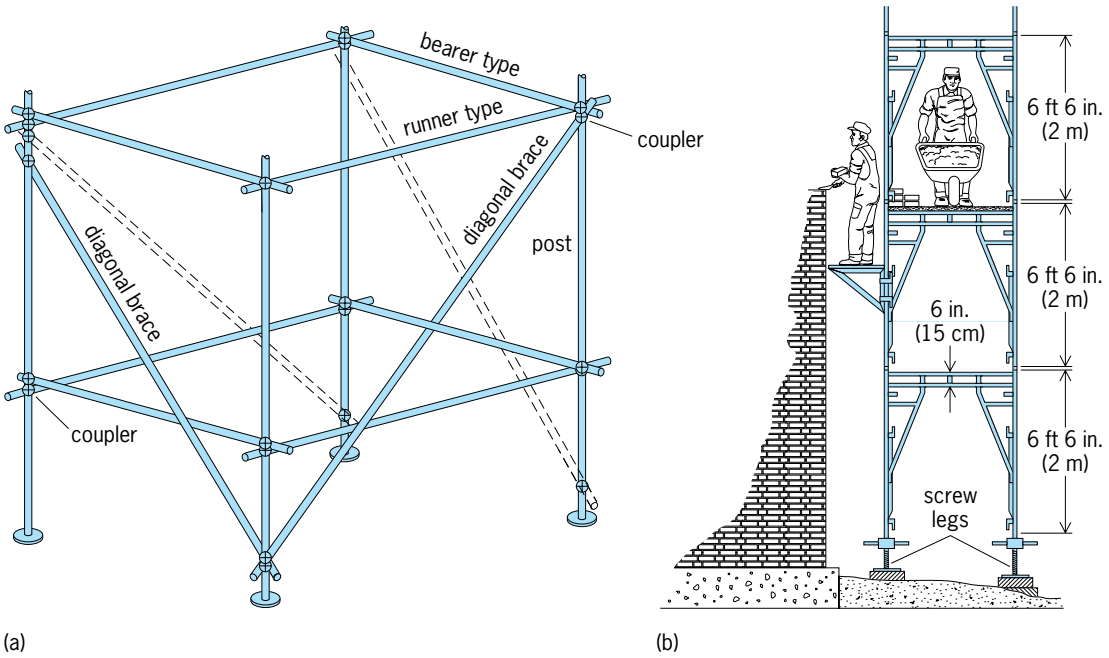


Fig. 3. Scaffolds. (*a*) Basic assembly of tube-and-coupler type. (*b*) Sectional walk-through type. (*After R. T. Ratay, Handbook of Temporary Structures in Construction, 2d ed., McGraw-Hill, 1996)*

longitudinal direction; and the couplers through which the parts connect (**Fig. 3***a*). Stability in the short direction is usually provided by anchoring against the permanent structure.

Sectional. These scaffolds have two basic components: prefabricated sectional frames that contain both posts and bearers (Fig. 3*b*), and vertical x-bracing that connects the frames and provides stability along the length of the scaffolding. Frames are stacked one on top of another up to the desired height. Some stability in the short direction is provided by the rigid frames themselves, but safety regulations require external bracing or guying above certain heights. Important additional parts of a stationary scaffold are the screw legs with base plates for solid founding and proper leveling of the assembly.

In addition to stationary ground-based scaffolds, there are rolling scaffold towers on casters, movable suspended scaffolds, and other special-purpose arrangements.

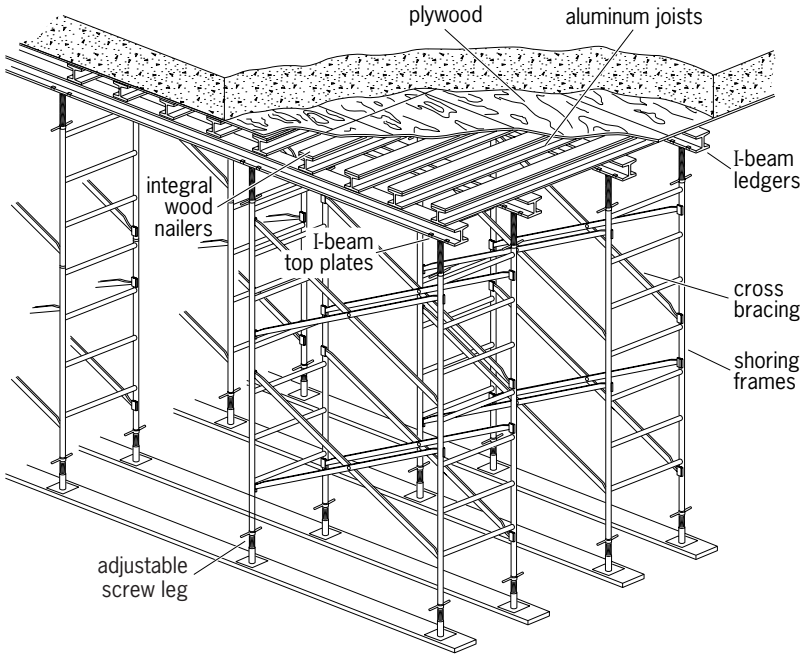


Fig. 4. Shoring towers used as falsework. (*After R. T. Ratay, Handbook of Temporary Structures in Construction, 2d ed., McGraw-Hill, 1996*)

Shoring. Shoring is a structure assembled for the purpose of temporarily supporting a load, usually during some construction, repair, or maintenance work. It may consist of a simple block, a single post, a single shoring tower, or an extensive system of columns, beams, braces, and towers. The traditional shoring material is wood, but manufactured shoring components made of tubes of structural steel are the most common. So-called shoring towers are used extensively because of their adaptability to load and size requirements. These towers are assembled at the site from two or more prefabricated steel shoring frames that are connected with x-bracing for lateral stability. Units of braced frames are stacked one on top of another to the desired height. Important additional components are the base (for solid founding and proper alignment of the shoring) and the head (for proper engagement of the supported load). Steel-tube shoring may look like scaffolding but it is different. While the function of scaffolding is to support the relatively light weights of workers and materials, shoring usually is used for heavy loads.

Falsework. Shoring that is used to support the formwork for fresh concrete or other materials is known as falsework (**Fig. 4**). Its purpose is to hold

the formwork in place until the concrete attains enough strength and stiffness to be self-supporting without excessive deformation. Falsework must be well founded, well braced, and tight against the formwork to prevent movement. After the removal of the formwork, the falsework is usually replaced by reshoring to provide continued support to the hardened concrete for an additional time period. Reshoring components are usually single wood or steel pipe posts.

Concrete formwork. Formwork is the mold into which the fresh concrete for a floor, beam, column, pier, and so forth is placed and remains until hardened. The formwork thus contains and temporarily supports the concrete and defines its permanent shape. Although the materials most frequently used for such applications are lumber and plywood because of their economy, availability, and workability in the job site, more durable materials such as aluminum, steel, or reinforced plastic are common for reusable forms. Forms may be custom-built at the site or constructed from prefabricated modular units. Tying the forms together, anchoring them, and supporting the framework is critical to maintain required shape and to resist movement due to very high liquid-concrete pressures. Premature or careless removal of the forms can create permanent damage to the built structure. *See* CONCRETE.

Bracing and guying. Temporary bracing and guying of a structure during construction provide lateral stability (resistance to horizontal forces). A brace is usually a rigid strut working in compression; a guy is a taut cable working in tension. In many instances, jacks are attached to braces, and turnbuckles to guy cables, to allow tightening or loosening in order to facilitate installation, alignment, and removal. Temporary bracing and guying may be installed either within or external to the permanent structure (**Fig. 5**), and remain until the construction of the permanent structure advances to the point which it can provide its own lateral stability. Thereafter the braces and guys are usually removed, but they may remain incorporated into the permanent structure. Because braces and guys are normally at an incline, the axial forces within them have horizontal and vertical components that must be included in calculating the construction design loads on the permanent structure.

Site protection structure. Hazards to life and property on and around a construction site are significant. Government regulations impose strict rules for safety at construction sites. A number of temporary structures are employed to warn and protect the workers and the public. Sidewalk bridges (shields) are strong platforms built on shoring posts or frames to protect the public from falling construction debris while maintaining pedestrian traffic on the sidewalk. Fall protection platforms and nets, used to catch falling objects and protect workers, people, and property below, are wooden planks and outstretched nets installed on outriggers near the top level of the construction; they are moved up periodically as the construction progresses. Barricades and fences protect people, equipment, and materials; delineate areas of the site; and organize the flow of traffic. Signs provide identification, information, and warning. Special-purpose temporary structures are employed at the adjacent areas to provide dust protection, blast protection, noise abatement, heat protection, and wind enclosures and to facilitate other construction-related activities.

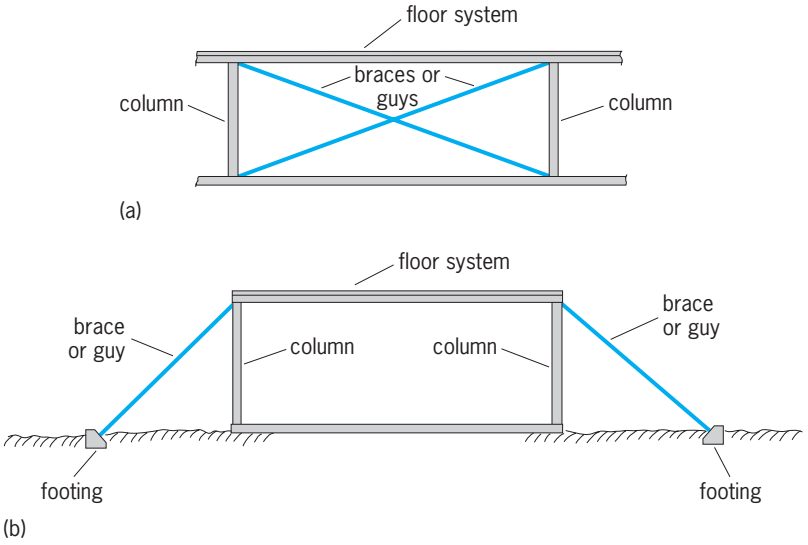


Fig. 5. Bracing and guying for lateral stability during construction. (*a*) Internally opposed braces or guys. (*b*) Externally opposed braces or guys. (*After R. T. Ratay, Handbook of Temporary Structures in Construction, 2d ed., McGraw-Hill, 1996*)

Robert T. Ratay

Bibliography. R. T. Ratay, *Forensic Structural Engineering Handbook*, McGraw-Hill, 2000; R. T. Ratay, *Handbook of Temporary Structures in Construction*, 2d ed., 1996.

Tendon

A cord connecting a muscle to another structure, often a bone. A tendon is a passive material, lengthening when the tension increases and shortening when it decreases. This characteristic contrasts with the active behavior of muscle. *See* MUSCLE.

Away from its muscle, a tendon is a compact cord. It can slide, in the direction of its length, freely through the body. At the muscle, it spreads into thin sheets called aponeuroses, which lie over and sometimes within the muscle belly. The large surface area of the aponeuroses allows the attachment of muscle fibers with a total cross-sectional area that is typically 50 times that of the tendon. The angle of attachment is usually 5–10°.

A tendon lengthens when subjected to a longitudinal tensile force. Stiffness increases with stress, reaching a constant value once the crimp has been straightened out. A tendon is a good spring in the sense that about 93% of the energy required in stretching is returned on release. Being constructed of partially independent, crimped fibers, tendons

can bend and buckle easily and without damage. The main exception to this flexibility is in the legs of some birds (for example, turkeys) where the tendons are calcified and thus are stiffer, especially against bending.

Structure. Tendons are living tissues that contain cells. In adult tendons, the cells occupy only a very small proportion of the volume and have a negligible effect on the mechanical properties. Like other connective tissues, tendon depends on the protein collagen for its strength and rigidity. The arrangement of the long, thin collagenous fibers is essentially longitudinal, but incorporates a characteristic waviness known as crimp. The fibers lie within a matrix of aqueous gel. Thus, tendon is a fiber-reinforced composite (like fiberglass), but its collagen is much less stiff than the glass and its matrix is very much less stiff than the resin. *See* COLLAGEN.

Where tendons go around corners, the matrix on the inside of the bend may be unusually stiff. This specialization makes such parts of tendon somewhat like cartilage. A more extreme variation is the incorporation of hard, calcified lumps, called sesamoid bones (for example, the kneecap), within a tendon.

Function. Tendons transmit force. They allow the force from the muscle to be applied in a restricted region. For example, the main muscles of the fingers are in the forearm, with tendons to the fingertips. If the hand had to accommodate these muscles, it would be too plump to be functional. Tendon extension can also be significant in the movement of a joint. For example, the tendon which flexes a human thumb joint is about 7 in. (170 mm) long. The maximum force from its muscle stretches this tendon about 0.1 in. (2.9 mm), which corresponds to rotation of the joint through an angle of about 21°. *See* JOINT (ANATOMY).

Some tendons save energy by acting as springs. In humans, the Achilles tendon reduces the energy needed for running by about 35%. This tendon is stretched during the first half of each step, storing energy which is then returned during takeoff. This elastic energy transfer involves little energy loss, whereas the equivalent work done by muscles would require metabolic energy in both stages.

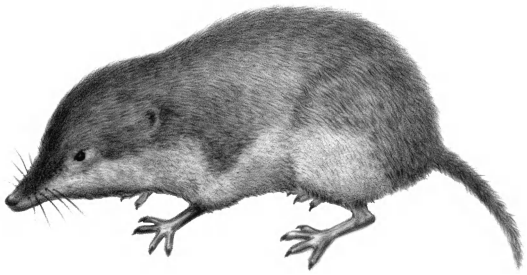
When a joint is misplaced, its ligaments may be overloaded and torn (for example, a sprained ankle). Tendons, in general, are much less at risk. Since they are in series with muscles, the maximum force which can be applied to a tendon is limited by the force its muscle can exert. Only those tendons which serve as springs are thin enough for the stress to reach far into the linear region, and are at risk from an excessive tensile load. For this reason, the Achilles tendon is the main human tendon prone to rupture. *See* CONNECTIVE TISSUE; MUSCULAR SYSTEM; SKELETAL SYSTEM.

Robert F. Ker

Bibliography. R. McN. Alexander, *Exploring Biomechanics: Animals in Motion*, 1992; V. C. Mow and W. C. Hayes (eds.), *Basic Orthopaedic Biomechanics*, 2d ed., 1997.

Tenrec

An insectivorous mammal indigenous to Madagascar. There are 30 species in 10 genera. These animals are nocturnally active and feed on insects, worms, and mollusks. All tenrecs are essentially primitive unspecialized mammals, with poor vision (see **illustration**). The digits are clawed, and the first digit is not



The tenrec is a small insectivorous mammal.

opposable to the others. Some species, such as the tailless or common tenrec (*Tenrec ecaudatus*), become dormant during the hot dry season. The largest tenrecs are 12–16 in. (30–40 cm) long. The body of the tenrec is covered with a mixture of hair, spines, and bristles; the tail is rudimentary; and the toes may be separate or webbed, depending on the species. These animals exude a strong, offensive odor. The dental formula is I 2/3 C 1/1 Pm 3/3 M 4/3, for a total of 40 teeth. Tenrecs are adapted to climbing and swimming. The female is prolific, with litters of 12–20 young, and has 22 mammary glands to feed the offspring. *See* DENTITION; SCENT GLAND.

The hedgehog or spring tenrec (*Setifer tenrec*) has stiff hairs that tend to be erected when the animal is disturbed. Some species may curl up into a ball when disturbed, as does the hedgehog. The rice tenrecs (*Oryzorictes talpoides* and *O. tetradactylus*) are fossorial, molelike species which do considerable damage to crops when they burrow for insects. *Microgale*, with 19 species, is the largest tenrec genus. *See* INSECTIVORA; MAMMALIA.

Charles B. Curtin

Bibliography. R. M. Nowak, *Walker's Mammals of the World*, Johns Hopkins University Press, 1999.

Tensor analysis

The systematic study of tensors which led to an extension and generalization of vectors, begun in 1900 by two Italian mathematicians, G. Ricci and T. Levi-Civita, following G. F. B. Riemann's proposal concerning a generalization of Euclidean geometry. The principal aim of the tensor calculus (absolute differential calculus) is to construct relationships which are generally covariant in the sense that these relationships or laws remain valid in all coordinate systems. The differential equations for the geodesics in a Riemannian space are covariant expressions; they yield a description of the geodesics which is valid

for all coordinate systems. On the other hand, Newton's equations of motion require a preferred coordinate system for their description, namely, one for which force is proportional to acceleration (an inertial frame of reference). Thus Albert Einstein was led to a study of Riemannian geometry and the tensor calculus in order to construct the general theory of relativity.

Arithmetic, or vector, n-space. The coordinates of a point in a three-dimensional Euclidean space are given by the triple of numbers (x,y,z) or (x^1,x^2,x^3), with $x = x^1$, $y = x^2$, $y = x^2$, $z = x^3$. The superscripts in x^i, $i = 1, 2, 3$, are simply labels which enable one to distinguish and order the various elements in the triple of numbers. The totality of all number triples of the form (x^1,x^2,x^3) with the x^i, $i = 1, 2, 3$, real, yields the arithmetic 3-space, designated as V_3. A simple generalization yields the arithmetic n-space, V_n. This space or manifold consists of all n-tuples of the form $(x^1,x^2, \ldots, x^n)$, the x^i, $i = 1, 2, \ldots, n$, taken as real numbers. Now a space of n dimensions is defined as any set of objects which can be put in 1:1 reciprocal correspondence with the arithmetic n-space. The 1:1 correspondence between the elements or points of the n-space can be chosen in many ways, and in general, the choice depends on the nature of the physical problem. In the special theory of relativity, an event is specified by the three space coordinates and the time, so that each event corresponds to a four-tuple $(x^1,x^2,x^3, x^4 = ct)$.

Let a point P of an n-space correspond to the n-tuple $(x^1,x^2, \ldots, x^n)$. Now consider the n equations, (1), and assume that one can solve for each x^i, yielding Eq. (2). It is assumed that Eqs. (1) and

$$y^i = y^i(x^1, x^2, \ldots, x^n) \quad i = 1, 2, \ldots, n \tag{1}$$

$$x^i = x^i(y^1, y^2, \ldots, y^n) \quad i = 1, 2, \ldots, n \tag{2}$$

(2) are single valued, and that the partial derivatives $\partial y^i/\partial x^i$, $\partial x^i/\partial y^i$, where $i, j = 1, 2, \ldots, n$, exist and are continuous. The n-space of which P is an element can also be put into 1:1 correspondence with the arithmetic n-space $(y^1,y^2, \ldots, y^n)$. Thus, one has a new description of every point in the n space. The points or elements of the n-space have not changed; one merely has a new description (coordinate system) of these points. Thus, Eq. (1) is called a transformation of coordinates.

The algebra of tensors is simplified by use of Einstein's summation convention. From the calculus, Eq. (1) yields Eq. (3). The index α occurs just

$$\frac{\partial y^i}{\partial y^j} = \sum_{\alpha=1}^{n} \frac{\partial y^i}{\partial x^\alpha} \frac{\partial x^\alpha}{\partial y^j} = \frac{\partial y^i}{\partial x^\alpha} \frac{\partial x^\alpha}{\partial y^j} \tag{3}$$

twice in the last expression of Eq. (3), thus indicating a summation over this index. In a space of three dimensions, the expression $S = a_{\alpha\beta}x^\alpha x^\beta$ requires a double summation because the indices α and β occur exactly twice. Thus,

$$S = \sum_{\beta=1}^{3} \sum_{\alpha=1}^{3} a_{\alpha\beta}x^\alpha x^\beta$$

and Einstein found it convenient to remove the summation signs. A further convenience occurs if the Kronecker delta is introduced. Thus, δ^i_j, $i,j = 1, 2, \ldots, n$, is a set of quantities whose numerical values are given by Eq. (4).

$$\delta^i_j = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases} \qquad i, j = 1, 2, \ldots, n \tag{4}$$

In particular

$$\delta^1_1 = \delta^2_2 = \cdots = \delta^n_n = 1$$
$$\delta^1_2 = \delta^1_3 = \cdots = \delta^{n-1}_n = 0$$

and for an n-space

$$\delta^\alpha_\alpha = \delta^1_1 + \delta^2_2 + \cdots + \delta^n_n = n$$

Equation (3) may be written

$$\delta^i_j = \frac{\partial y^i}{\partial x^\alpha} \frac{\partial x^\alpha}{\partial y^i}$$

This expression yields

$$\left|\frac{\partial y^i}{\partial x^j}\right| \left|\frac{\partial x^\alpha}{\partial y^\beta}\right| = 1$$

applying the rule for the product of two determinants, with

$$\left|\frac{\partial y^i}{\partial x^j}\right|$$

the Jacobian of the transformation in Eq. (1), and the

$$\left|\frac{\partial x^i}{\partial y^j}\right|$$

Jacobian of the inverse transformation in Eq. (2).

Contravariant vectors. In an n-space the locus of elements given by Eq. (5) represents a space curve Γ with λ a parameter. The n-tuple in notation (6), designated by $dx^i/d\lambda$, is defined to be a tangent element to the space curve, Eq. (5). Under an allowable coordinate transformation, Eq. (1), the space curve Γ can be represented by Eq. (7). The components of the

$$x^i = x^i(\lambda) \quad i = 1, 2, \ldots, n \quad \lambda_0 \leq \lambda \leq \lambda_1 \tag{5}$$

$$\left(\frac{dx^1}{d\lambda}, \frac{dx^2}{d\lambda}, \ldots, \frac{dx^n}{d\lambda}\right) \tag{6}$$

$$y^i = y^i\left(x^1(\lambda), x^2(\lambda), \ldots, x^n(\lambda)\right) = y^i(\lambda) \tag{7}$$

$$i = 1, 2, \ldots, n$$

tangent vector for the $y = (y^1,y^2, \ldots, y^n)$ coordinate

system are given by notation (8), which represent the elements of the n-tuple, notation (6).

$$\frac{dy^i}{d\lambda} \qquad i = 1,\ 2,\ \ldots, n \tag{8}$$

Certainly the x-coordinate system is no more important than the y-coordinate system as a description of the space curve Γ. One cannot say that $dx^i/d\lambda$ is the tangent vector; nor can one say that $dy^i/d\lambda$ is the tangent vector. If one considers all allowable coordinate systems, one obtains the entire class of tangent elements, each element (n-tuple) claiming to be the tangent vector for its particular coordinate system. It is the abstract collection of all these tangent elements that is specified as the tangent vector. To find what relationship exists between the components of the tangent vector when described by two different coordinate systems, Eq. (7) can be used to give Eq. (9).

$$\frac{dy^i}{d\lambda} = \frac{\partial y^i}{\partial x^\alpha}\frac{dx^\alpha}{d\lambda} \qquad i = 1,\ 2,\ \ldots,\ n \tag{9}$$

Note that at any point each component $dy^i/d\lambda$ depends linearly on every component of $dx^i/d\lambda$. Moreover, Eq. (10) holds.

$$\frac{dx^i}{d\lambda} = \frac{\partial x^i}{\partial y^\alpha}\frac{dy^\alpha}{d\lambda} \qquad i = 1,\ 2,\ \ldots,\ n \tag{10}$$

One can now form the following generalization: Any set of numbers $A^i(x^1,x^2,\ \ldots,\ x^n)$, $i = 1,\ 2,\ \ldots,\ n$, which transform according to the law shown by Eq. (11) under the coordinate transformation

$$\overline{A}^i(\overline{x}^1, \overline{x}^2,\ \ldots, \overline{x}^n) = A^\alpha(x^1, x^2,\ \ldots,\ x^n)\frac{\partial \overline{x}^i}{\partial x^\alpha} \tag{11}$$

$$i = 1,\ 2,\ \ldots,\ n$$

$\overline{x}^i = \overline{x}(\overline{x}^1,\ x^2,\ \ldots,\ x^n)$, $i = 1,\ 2,\ \ldots,\ n$, is said to be a contravariant vector. The vector is not just the set of components in any coordinate system, but is rather the abstract element which is represented in each coordinate system x by the set of numbers $A^i(x)$. The $\overline{A}^i(\overline{x})$ depend linearly on the $A^\alpha(x)$; therefore the sum and difference of two vectors in an n-space are also vectors.

One immediately sees that the law of transformation for a contravariant vector is transitive. Let

$$\overline{A}^i = A^\alpha\frac{\partial \overline{\overline{x}}^i}{\partial x^\alpha} \qquad \overline{\overline{A}}^i = \overline{A}^\alpha\frac{\partial \overline{\overline{x}}^i}{\partial \overline{x}^\alpha}$$

so that

$$\overline{\overline{A}}^i = \overline{A}^\beta\frac{\partial \overline{\overline{x}}^i}{\partial \overline{x}^\beta} = A^\alpha\frac{\partial \overline{x}^\beta}{\partial x^\alpha}\frac{\partial \overline{\overline{x}}^i}{\partial \overline{x}^\beta} = A^\alpha\frac{\partial \overline{\overline{x}}^i}{\partial x^\alpha}$$

which proves the statement.

If the components of a contravariant vector are known in one coordinate system, the components are known in all allowable coordinate systems from Eq. (11). A coordinate transformation does not yield a new vector; it merely changes the components of the same vector. Thus, a vector is said to be invariant under a coordinate transformation. Any object which is not changed under a coordinate transformation is called an invariant.

If $\phi(x^1,x^2,\ \ldots,\ x^n)$ is a scalar point function $\phi(x^1,x^2,\ldots,x^n) = \phi(x^1(\overline{x}),x^2(\overline{x}),\ \ldots,\ x^n(\overline{x})) = \overline{\phi}(\overline{x}^2,\ \overline{x}^2,\ \ldots,\ \overline{x}^n)$, then Eq. (12) can be derived.

$$\frac{\partial \overline{\phi}}{\partial \overline{x}^i} = \frac{\partial \phi}{\partial x^\alpha}\frac{\partial x^\alpha}{\partial \overline{x}^i} \tag{12}$$

The n-tuple

$$\left(\frac{\partial \phi}{\partial x^1}, \frac{\partial \phi}{\partial x^2}, \ldots, \frac{\partial \phi}{\partial x^n}\right)$$

yields the components of a covariant vector, called the gradient of ϕ.

More generally Eq. (13) holds, if one says that the

$$\overline{A}_i(\overline{x}^1, \overline{x}^2,\ \ldots,\ \overline{x}^n) = A(x^1, x^2, \ldots,\ x^n)\frac{\partial x^\alpha}{\partial \overline{x}^i} \tag{13}$$

$$i = 1,\ 2,\ \ldots,\ n$$

$A_\alpha(x)$ are the components of a covariant vector.

The law of transformation, Eq. (13), differs from that of Eq. (11) because in general,

$$\frac{\partial x^\alpha}{\partial \overline{x}^i} \neq \frac{\partial \overline{x}^i}{\partial x^\alpha}$$

However, for the group of orthogonal transformations $\overline{x}^i = a^i_\alpha x^\alpha$ such that $\mathbf{A} = \|a^i_j\|$ satisfies $\mathbf{A}^T = \mathbf{A}^{-1}$, one can show that $\partial \overline{x}^i/\partial x^\alpha = \partial x^\alpha/\partial \overline{x}^i$, and for this reason no distinction occurs between contravariant and covariant vectors in elementary vector analysis.

An important scalar invariant can be formed from a contravariant vector A^i and a covariant vector B_i. From their laws of transformation it follows that Eq. (14) holds.

$$\overline{A}^i\overline{B}_i = A^\alpha B_\beta\frac{\partial \overline{x}^i}{\partial x^\alpha}\frac{\partial x^\beta}{\partial \overline{x}^i} = A^\alpha B_\beta\delta_\alpha{}^\beta$$

$$= A^\alpha B_\alpha = A^i B_i \tag{14}$$

Thus, the expression

$$A^iB_i = \sum_{i=1}^{n} A^iB_i$$

is invariant both in form and in its numerical value under a transformation of coordinates. One calls $A^\alpha B^\alpha$ the scalar, or inner, product, of the two vectors A^i,B_i.

Tensors. The contravariant and covariant vectors discussed above are special cases of differential invariants called tensors. The components of a tensor are of the form

$$T^{a_1a_2\ \ldots\ a_r}_{b_1b_2\ \ldots\ b_s}(x^1, x^2,\ \ldots,\ x^n)$$

and indices $a_1,a_2,\ \ldots,\ a_r,b_1,b_2,\ \ldots,\ b_s$ run through the integers $1,\ 2,\ \ldots,\ n$. The components transform according to the rule shown by Eq. (15).

$$\overline{T}^{a_1a_2\ldots a_r}_{b_1b_2\ldots b_s} = \left|\frac{\delta x}{\delta \overline{x}}\right|^N T^{\alpha_1\alpha_2\ldots\alpha_r}_{\beta_1\beta_2\ldots\beta_s}\frac{\partial \overline{x}^{\alpha 1}}{\partial x^{\alpha 1}}$$

$$\cdots\frac{\partial \overline{x}^{ar}}{\partial x^{\alpha r}}\frac{\partial x^{\beta 1}}{\partial \overline{x}^{b1}}\cdots\frac{\partial x^{\beta s}}{\partial \overline{x}^{bs}} \tag{15}$$

The exponent N of the Jacobian

$$\left|\frac{\partial x}{\partial \overline{x}}\right|$$

is called the weight of the tensor field. For $N = 0$ the tensor field is absolute; otherwise it is of weight N. A tensor density occurs for $N = 1$. The number of indices is $r + s$, the rank of the tensor. Vectors are tensors of rank one. The tensor of Eq. (15) is contravariant of order r and covariant of order s.

An important property of tensors is immediately evident from Eq. (15). If every component of a tensor vanishes in one coordinate system, the components vanish in all coordinate systems.

Two tensors are said to be of the same kind if they have the same order in their contravariant and covariant indices and if they are of the same weight. Additional tensors can be constructed as follows:

1. The sum and difference of two tensors of the same kind yields a new tensor of the same kind. This is apparent from the linear property of Eq. (15).

2. The product of two tensors is a new tensor. This can be shown for a special case. Let

$$\overline{T}^a_b = \left|\frac{\partial x}{\partial \overline{x}}\right| T^\alpha_\beta \frac{\partial x^\beta}{\partial \overline{x}^b}\frac{\partial \overline{x}^a}{\partial x^\alpha} \qquad \overline{S}^c = \left|\frac{\partial x}{\partial \overline{x}}\right|^3 S^\gamma \frac{\partial \overline{x}^c}{\partial x^\gamma}$$

Then

$$\overline{W}^{ac}{}_b \equiv \overline{T}^a_b \overline{S}^c = \left|\frac{\partial x}{\partial \overline{x}}\right|^4 (T^\alpha_\beta S^\gamma)\frac{\partial \overline{x}^a}{\partial x^\alpha}\frac{\partial \overline{x}^c}{\partial x^\gamma}\frac{\partial x^\beta}{\partial \overline{x}^b}$$

$$= \left|\frac{\partial x}{\partial \overline{x}}\right|^4 W^{\alpha\gamma}_\beta \frac{\partial \overline{x}^a}{\partial x^\alpha}\frac{\partial \overline{x}^c}{\partial x^\gamma}\frac{\partial x^\beta}{\partial \overline{x}^b}$$

To illustrate the concept of contraction, consider the absolute tensor

$$\overline{A}^{ij}_{kl} = A^{\alpha\beta}_{\sigma\tau}\frac{\partial \overline{x}^i}{\partial x^\alpha}\frac{\partial \overline{x}^j}{\partial x^\beta}\frac{\partial x^\sigma}{\delta \overline{x}^k}\frac{\partial x^\tau}{\partial \overline{x}^l}$$

Replacing k by i and summing yields

$$\begin{aligned}\overline{B}^j_l \equiv \overline{A}^{ij}_{il} &\equiv A^{\alpha\beta}_{\sigma\tau}\frac{\partial \overline{x}^i}{\partial x^\alpha}\frac{\partial \overline{x}^j}{\partial x^\beta}\frac{\partial x^\sigma}{\partial \overline{x}^i}\frac{\partial x^\tau}{\partial \overline{x}^l}\\ &= A^{\alpha\beta}_{\sigma\tau}\frac{\partial x^\sigma}{\partial x^\alpha}\frac{\partial \overline{x}^j}{\partial x^\beta}\frac{\partial x^\tau}{\partial \overline{x}^l}\\ &= A^{\alpha\beta}_{\sigma\tau}\delta^\sigma_\alpha\frac{\partial \overline{x}^j}{\partial x^\beta}\frac{\partial x^\tau}{\partial \overline{x}^l}\\ &= A^{\sigma\beta}_{\sigma\tau}\frac{\partial \overline{x}^j}{\partial x^\beta}\frac{\partial x^\tau}{\partial \overline{x}^l} = B^\beta_\tau\frac{\partial \overline{x}^j}{\partial x^\beta}\frac{\partial x^\tau}{\partial \overline{x}^l}\end{aligned}$$

so that $B^\beta_\tau \equiv A^{\sigma\beta}_{\sigma\tau}$ is a mixed tensor. In general, one equates a certain covariant index with a contravariant index, sums on this repeated index, and obtains a new tensor whose rank is two less than that of the original tensor. This process of producing a new tensor is called the method of contraction.

A few examples of tensors can be listed.

1. The Kronecker delta is a mixed absolute tensor because

$$\delta^i_j\frac{\partial x^j}{\partial \overline{x}^\alpha}\frac{\partial \overline{x}^\beta}{\partial x^i} = \frac{\partial x^i}{\partial \overline{x}^\alpha}\frac{\partial \overline{x}^\beta}{\partial x^i} = \frac{\partial \overline{x}^\beta}{\partial \overline{x}^\alpha} = \overline{\delta}^\beta_\alpha$$

2. If A_i and B_i are absolute covariant vectors, then

$$\overline{C}_{ij} \equiv \overline{A}_i\overline{B}_j = A_\alpha B_\beta\frac{\partial x^\alpha}{\partial \overline{x}^i}\frac{\partial x^\beta}{\partial \overline{x}^j} = C_{\alpha\beta}\frac{\partial x^\alpha}{\partial \overline{x}^i}\frac{\partial x^\beta}{\partial \overline{x}^j}$$

so that $C_{\alpha\beta} \equiv A_\alpha B_\beta$ are the components of an absolute covariant tensor of rank two.

3. Let ϕ_i be the components of an absolute covariant vector. From $\overline{\phi}_i = \phi_\alpha(\partial x^\alpha/\partial \overline{x}^i)$, it follows that

$$\begin{aligned}\overline{F}_{ij} &\equiv \frac{\partial \overline{\phi}_i}{\partial \overline{x}^j} - \frac{\partial \overline{\phi}_j}{\partial \overline{x}^i} = \left(\frac{\partial \phi_\alpha}{\partial x^\beta} - \frac{\partial \phi_\beta}{\partial x^\alpha}\right)\frac{\partial x^\alpha}{\partial \overline{x}^i}\frac{\partial x^\beta}{\partial \overline{x}^j}\\ &= F_{\alpha\beta}\frac{\partial x^\alpha}{\partial \overline{x}^i}\frac{\partial x^\beta}{\partial \overline{x}^j}\end{aligned}$$

so that

$$F_{\alpha\beta} \equiv \frac{\partial \phi_\alpha}{\partial x^\beta} - \frac{\partial \phi_\beta}{\partial x^\alpha}$$

are the components of an absolute covariant tensor. Note that $F_{\alpha\beta} = -F_{\beta\alpha}$, so that the tensor is skew symmetric. If the $\phi_i, i = 1, 2, 3, 4$, are the components of the electromagnetic vector potential, the F_{ij} are those of the electromagnetic field tensor.

Line element of Riemannian geometry. A simple generalization of the euclidean metric, or line element, $ds^2 = dx^2 + dy^2 + dz^2$, led Riemann to consider an n-space such that a metric is imposed on the space which yields the invariant distance ds between two nearby points whose coordinates differ by dx^i. The line element in a riemannian n-space is given by Eq. (16).

$$ds^2 = g_{\alpha\beta}(x^1, x^2, \ldots, x^n)\,dx^\alpha dx^\beta \qquad (16)$$

$$g_{\alpha\beta} = g_{\beta\alpha}$$

Under a coordinate transformation, $x^\alpha = x^\alpha(\overline{x})$, Eq. (16) becomes Eq. (17).

$$ds^2 = g_{\alpha\beta}\frac{\partial x^\alpha}{\partial \overline{x}^\sigma}\frac{\partial x^\beta}{\partial \overline{x}^\tau}\,d\overline{x}^\sigma\,dx^\tau = \overline{g}_{\sigma\tau}\,d\overline{x}^\sigma d\overline{x}^\tau \qquad (17)$$

$$\overline{g}_{\sigma\tau} = g_{\alpha\beta}\frac{\partial x^\alpha}{\partial \overline{x}^\sigma}\frac{\partial x^\beta}{\partial \overline{x}^\tau}$$

Hence the $g_{ij}(x)$ are the components of an absolute covariant tensor of rank two. One can show that the $g^{ij}(x)$ defined by $g^{ij}g_{jk} = \delta^i_k$ are the components of an absolute contravariant tensor, provided $|g_{ij}| \neq 0$. Furthermore, it can be shown that

$$|\overline{g}_{ij}| = |g_{ij}|\left|\frac{\partial x}{\partial \overline{x}}\right|^2$$

Thus, if A^i are the components of an absolute contravariant vector, then $|g_{\alpha\beta}|^{1/2}A^i$ are the components of a contravariant vector density.

Geodesics in a Riemannian space. If a space curve in a Riemannian n-space is given by $x^i = x^i(\lambda)$, $\lambda_0 \leqq \lambda \leqq \lambda_1$, the distance along the curve between the end points is given by Eq. (18).

$$L = \int_{\lambda_0}^{\lambda_1}\left(g_{\alpha\beta}\frac{dx^\alpha}{d\lambda}\frac{dx^\beta}{d\lambda}\right)^{1/2}d\lambda \qquad (18)$$

The geodesic path between two fixed points of a Riemannian space is that curve joining the two points which yields a minimum value for the integral

of Eq. (18). The same path is obtained if one minimizes, as shown in notation (19).

$$\int_{P_0}^{P_1} \left(g_{\alpha\beta} \frac{dx^\alpha}{ds} \frac{dx^\beta}{ds} \right) ds \tag{19}$$

Applying the Euler-Lagrange equations of the calculus of variations,

$$\frac{d}{ds}\left(\frac{\partial L}{\partial \dot{x}^i}\right) = \frac{\partial L}{\partial x^i}$$

with $L = g_{\alpha\beta}\dot{x}^\alpha \dot{x}^\beta$, yields the differential equations of the geodesics, Eq. (20), with Eq. (21) defining a term.

$$\frac{d^2x^i}{ds^2} + \Gamma^i_{jk}(x)\frac{dx^j}{ds}\frac{dx^k}{ds} = 0 \tag{20}$$

$$\Gamma^i_{jk} = \frac{1}{2} g^{i\sigma}\left(\frac{\partial g_{\sigma j}}{\partial x^k} + \frac{\partial g_{k\sigma}}{\partial x^j} - \frac{\partial g_{jk}}{\partial x^\sigma}\right) \tag{21}$$

The important elements Γ^i_{jk} are called the Christoffel symbols of the second kind. The elements $\{i,jk\} = g_{i\sigma}\Gamma^\sigma_{jk}$ are the Christoffel symbols of the first kind. Under a coordinate transformation, the Christoffel symbols transform according to the rule expressed in Eq. (22).

$$\overline{\Gamma}^i_{jk}(\overline{x}) = \Gamma^\alpha_{\beta\gamma}(x)\frac{\partial x^\beta}{\partial x^j}\frac{\partial x^\gamma}{\partial \overline{x}^k}\frac{\partial \overline{x}^i}{\partial x^\alpha} + \frac{\partial^2 x^\alpha}{\partial \overline{x}^j \partial \overline{x}^k}\frac{\partial \overline{x}^i}{\partial x^\alpha} \tag{22}$$

Thus, the Christoffel symbols are not the components of a mixed tensor. However, if the $\Gamma^i_{jk}(x)$ are given in an x-coordinate system, their values can be computed in an $\overline{x}$-coordinate system from Eq. (22).

Covariant differentiation. From Eq. (11) follows Eq. (23), so that $\partial A^\alpha/\partial x^\beta$ are not the components

$$\frac{\partial \overline{A}^i}{\partial \overline{x}^j} = \frac{\partial A^\alpha}{\partial x^\beta}\frac{\partial \overline{x}^i}{\partial x^\alpha}\frac{\partial x^\beta}{\partial \overline{x}^j} + A^\alpha \frac{\partial^2 \overline{x}^i}{\partial x^\beta \partial x^\alpha}\frac{\partial x^\beta}{\partial \overline{x}^j} \tag{23}$$

of a mixed tensor. This is not surprising; in a euclidean space using coordinates other than rectangular coordinates, the unit vectors change directions from point to point. A differentiation process which yields a new tensor can be introduced by combining Eqs. (22) and (23). It can be shown, Eq. (24), that

$$A^i_{\ ,j} \equiv \frac{\partial A^i}{\partial x^j} + A^i_\alpha \Gamma^i_{\alpha j} \tag{24}$$

the terms are the components of a mixed tensor. The term $A^i_{\ ,j}$ is called the covariant derivative of A^i (the comma denoting covariant differentiation).

In a euclidean space using rectangular coordinates, the Γ^i_{jk} vanish from Eq. (21), and the covariant derivative reduces to the ordinary partial derivative.

$$ds^2 = \sum_{i=1}^{n} (dx^i)^2$$

The intrinsic derivative of A^i along a path $x^i = x^i(s)$, with s arc length, is given by Eq. (25).

$$\frac{\delta A^i}{\delta s} \equiv A^i_{\ ,j}\frac{dx^j}{ds} = \frac{dA^i}{ds} + A^\alpha \Gamma^i_{\alpha j}\frac{dx^j}{ds} \tag{25}$$

The covariant derivative of the general tensor given by Eq. (15) is Eq. (26).

$$\begin{aligned} T^{\alpha_1\alpha_2\cdots\alpha_r}_{\beta_1\beta_2\cdots\beta_s,j} &= \frac{\partial T^{\alpha_1\alpha_2\cdots\alpha_r}_{\beta_1\beta_2\cdots\beta_s}}{\partial x^j} + T^{\mu\alpha_2\cdots\alpha_r}_{\beta_1\beta_2\cdots\beta_s}\Gamma^{\alpha_1}_{\mu j} \\ &+ \cdots + T^{\alpha_1\alpha_2\cdots\alpha_{r-1}\mu}_{\beta_1\beta_2\cdots\beta_s}\Gamma^{\alpha_r}_{\mu j} - T^{\alpha_1\alpha_2\cdots\alpha_r}_{\mu\beta_2\cdots\beta_s}\Gamma^{\mu}_{\beta_1 j} \\ &- \cdots - T^{\alpha_1\alpha_2\cdots\alpha_r}_{\beta_1\beta_2\cdots\beta_{s-1}\mu}\Gamma^{\mu}_{\beta_s j} - NT^{\alpha_1\alpha_2\cdots\alpha_r}_{\beta_1\beta_2\cdots\beta_s}\Gamma^{\mu}_{\mu j} \end{aligned} \tag{26}$$

A few examples of covariant differentiation now follow:

1. Let A_i be an absolute covariant vector. The curl of A_i is defined by Eq. (27).

$$\begin{aligned} \text{curl } A_i &= A_{i,j} - A_{j,i} \\ &= \frac{\partial A_i}{\partial x^j} - A_\alpha \Gamma^\alpha_{ij} - \frac{\partial A_j}{\partial x^i} + A_\alpha \Gamma^\alpha_{ji} \\ &= \frac{\partial A_i}{\partial x^j} - \frac{\partial A_j}{\partial x^i} \end{aligned} \tag{27}$$

2. The divergence of A^i is defined by Eq. (28) with

$$\begin{aligned} \text{div } A^i = A^\alpha_{\ ,\alpha} &= \frac{\partial A^\alpha}{\partial x^\alpha} + A^\alpha \Gamma^\beta_{\alpha\beta} \\ &= \frac{1}{\sqrt{|g|}}\frac{\partial}{\partial x^\alpha}(=\sqrt{|g|}\,A^\alpha) \end{aligned} \tag{28}$$

$|g| = |g|_{ij}|$. The proof of Eq. (28) is omitted.

3. If ϕ is an absolute scalar, the gradient of ϕ is defined by $\phi_{,i} = \partial\phi/\partial x^i$. The associated vector of $\phi_{,i}$ is $g^{\alpha\beta}\phi_{,\beta} = g^{\alpha\beta}(\partial\phi/\partial x^\beta) = A^\alpha$. Applying Eq. (28) to A^α yields the Laplacian of ϕ given by Eq. (29).

$$\text{Lap } \phi = \frac{1}{\sqrt{|g|}}\frac{\partial}{\partial x^\alpha}\left(\sqrt{|g|}g^{\alpha\beta}\frac{\partial\phi}{\partial x^\beta}\right) \tag{29}$$

See CALCULUS OF VECTORS; RIEMANNIAN GEOMETRY.

Harry Lass

Bibliography. R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis and Applications*, 1983, reprint 1988; C. T. Dodson and T. Poston, *Tensor Geometry: The Geometric Viewpoint and Its Uses*, 2d ed., 1991; A. W. Joshi, *Matrices and Tensors in Physics*, 3d ed., 1995; D. C. Kay, *Schaum's Outline of Tensor Calculus,* 1988; J. L. Synge and A. Schild, *Tensor Calculus*, 1969, reprint 1978; R. Wasserman, *Tensors and Manifolds: With Applications to Mechanics and Relativity*, 1992; E. C. Young, *Vector and Tensor Analysis*, 2d ed., 1992.

Terbium

Element number 65, terbium, Tb, is a very rare metallic element of the rare-earth group. Its atomic weight is 158.924, and the stable isotope ^{159}Tb makes up

1																	18
1 H	2											13	14	15	16	17	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

100% of the naturally occurring element. *See* PERIODIC TABLE.

The common oxide, Tb_4O_7, is brown and is obtained when its salts are ignited in air. Its salts are all trivalent and white in color and, when dissolved, give colorless solutions. The higher oxides slowly decompose when treated with dilute acid to give the trivalent ions in solution. Although the metal is attacked readily at high temperatures by air, the attack is extremely slow at room temperatures. The metal has a Néel point at about 229 K and a Curie point at about 220 K. For properties of the metal *see* RARE-EARTH ELEMENTS

Frank H. Spedding

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; K. A. Gschneidner Jr., J.-C. Bünzli, and V. K. Pecharsky (eds.), *Handbook on the Physics and Chemistry of Rare Earths*, 2005.

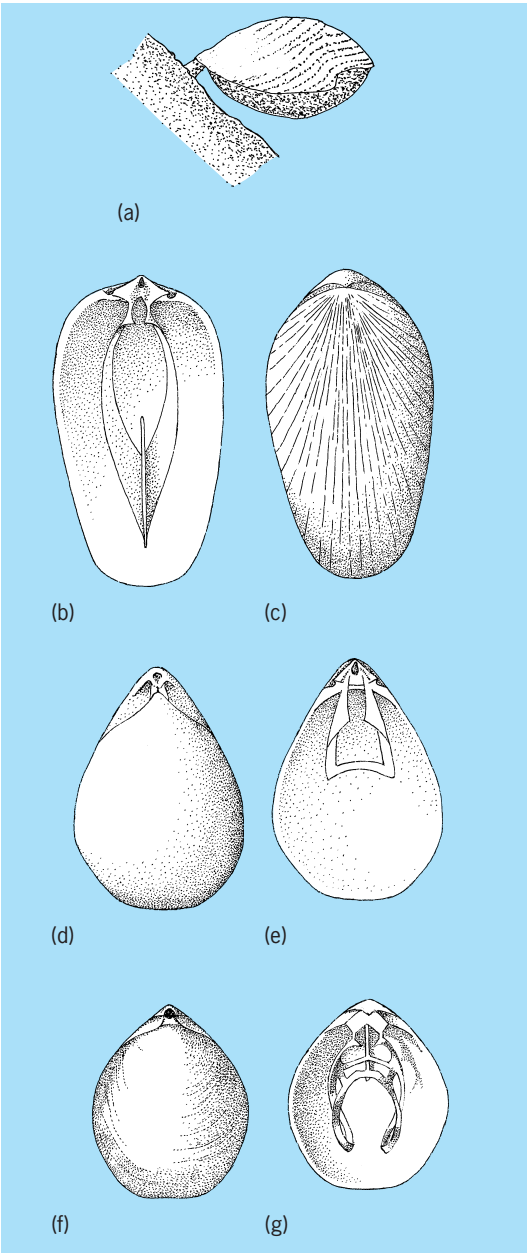
Terebratulida

An order of articulated brachiopods consisting of a group of sessile, suspension-feeding, marine, benthic, epifaunal bivalves with representatives occurring from the Early Devonian Era. It is most diverse and abundant group of living brachiopods, which probably exhibits maximum diversity in present-day seas.

The shells are biconvex and usually smooth, although some show radial ribbed ornamentation. The valves articulate about a hinge structure and posterior edges of the valves are not coincident with the hinge axis (nonstrophic condition). The shell is calcareous and punctate. A fleshy pedicle usually attaches the animal to the substrate (**illus.** *a*), but in thecidine brachiopods the ventral valve is cemented to the substrate. Terebratulids have a pair of adductor muscles and a pair of diductor muscles, and a dorsal pair and a ventral pair of adjustor muscles.

The tentacular feeding organ (lophophore) occupies the mantle cavity as a looped structure (the ptycholophe) in the smaller forms or as a looped and coiled structure (the plectolophe) in the larger forms. In the smaller forms the lophophore is supported by a calcareous ridge, but in the larger forms by a calcareous loop. Both of these structures emanate from the dorsal (brachial) valve. The digestive system consists of a mouth, pharynx, esophagus, stomach, digestive diverticula, pylorus, and blind intestine. Digestive diverticula usually occur as a pair of posterior lobes which open to the stomach through one of three pairs of ducts.

The excretory system consists of a pair of ciliated funnels (metanephridia) which during spawning function as gonoducts and allow the discharge of gametes from the coelom into the mantle cavity. Some solid waste may also be ejected through the nephridiopores, enmeshed in mucus, while the main



Terebratulida. (*a*) Living *Calloria*, attached to the substrate by the pedicle. (*b, c*) *Rensselaeria* (Centronellida), dorsal view and brachial interior; (*d, e*) *Cranaena* (Terebratulidina), dorsal view and brachial interior; (*f, g*) *Laqueus* (Terebratellidina), dorsal view and brachial interior. (*After R. C. Moore, ed., Treatise on Invertebrate Paleontology, pt. H, Geological Society of America and University of Kansas Press, 1965*)

excretory product, ammonia, is voided through the tissues of the mantle and lophophore.

Terebratulids possess an open circulatory system of blood vessels and coelomic canals, containing fluid which is coagulable and has free cellular inclusions consisting of blood cells and coelomocytes. They have a central nervous system containing unsheathed nerves, but no differentiated sense organs. Members of this order are known to be gonochoristic or hermaphroditic and produce lecithotrophic larvae, which in some species are brooded between the tentacles of the lophophore or in specialized brood chambers.

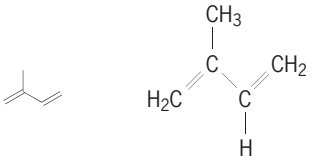
Three suborders are recognized: Centronellida, which have a primitive calcareous lophophore loop support of oval form (illus. *b* and *c*); Terebratulidina, which have a short calcareous lophophore loop support (illus. *d* and *e*); and Terebratellidina, which have a long calcareous lophophore loop support (illus. *f* and *g*). *See* RHYNCHONELLIFORMEA; BRACHIOPODA.

Mark A. James

Bibliography. M. A. James et al., Biology of living brachiopods, *Adv. Mar. Biol.*, 28:175–387, 1992.

Terpene

A class of natural products having a structural relationship (**1***a*) to isoprene (**1***b*). Over 5000 struc-



(**1***a*) (**1***b*)

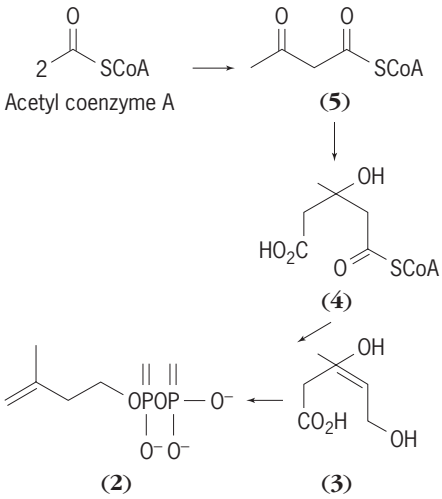
turally determined terpenes are known; many of these have also been synthesized in the laboratory. Historically terpenes have been isolated from green plants, but new compounds structurally related to isoprene continue to be isolated from other sources as well, so the class is also referred to as terpenoids, reflecting the biochemical origin without specification of the natural source.

Classification. Terpenes are classified according to the number of isoprene units of which they are composed, as is shown below.

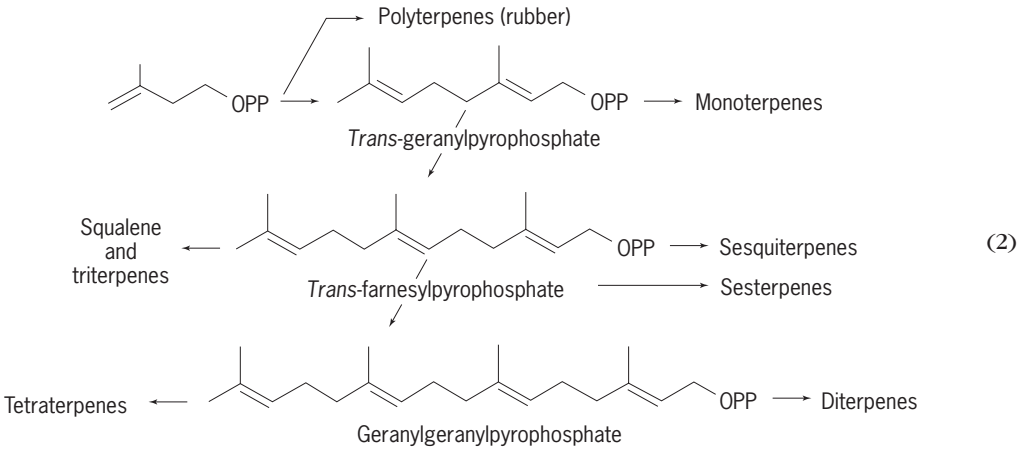
5	hemi-	25	ses-
10	mono-	30	tri-
15	sesqui-	40	tetra-
20	di-	$(5)_n$	poly-

Although they may be named according to the systematic nomenclature and numbering systems set by the International Union of Pure and Applied Chemistry for all organic compounds, it is often easier to refer to terpenes by their common names, which usually reflect the botanical or zoological name of their source.

Biogenesis. The function of terpenes in plants and other organisms is not clear, although they sometimes possess toxic properties linked to the protection of the species. Terpenes are products of secondary metabolism [reaction (1)]; the key building



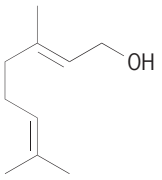
block, isopentenylpyrophosphate (**2**), arises from mevalonic acid (**3**) via hydroxymethyl glutarate (**4**).



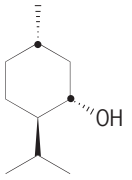
The starting point of this metabolic pathway is believed to be the condensation of two molecules of acetic acid to form acetoacetyl coenzyme A (**5**). Other biogenetic pathways available for the production of hydroxymethyl glutarate (**4**) depend on the particular organism.

Isopentenylprophosphate (**2**) has two reactive sites and can polymerize in a variety of ways to form larger terpenes. Its head-to-tail dimerization is the most commonly encountered process, although head-to-head and tail-to-tail linkages are also found. The multitude of structural types in terpenes arises from intramolecular rearrangements of several basic skeletons formed by the cyclizations of the linear precursors shown in reactions (2), where OPP represents oxygen-alkylated pyrophosphate. *See* POLYMER.

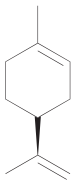
Monoterpenes. Acyclic or cyclic C_{10} hydrocarbons and their oxygenated derivatives are known as monoterpenes. The most common structural types are biogenetic derivatives of geraniol (**6**), the main constituent of geranium oil. Menthol (**7**) is the chief constituent of peppermint oil; limonene (**8**) composes over 90% of lemon oil; oil of rosemary contains α-pinene (**9**); camphor (**10**) is the main component of sage oil; and other interesting monoterpenes are iridoids, substances which have been isolated from ants, such as iridodial (**11**).



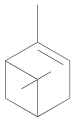
(**6**)



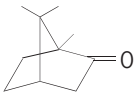
(**7**)



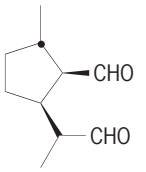
(**8**)



(**9**)

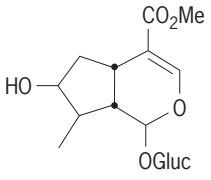


(**10**)

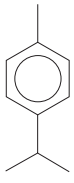


(**11**)

Loganin (**12**; Me = CH_3 and Glu = β-glucose) is the biogenetic precursor of some alkaloids. Its combination with tryptophane, for example, commences the biogenesis of yohimbine alkaloids. Cyclic monoterpenes can be easily oxidized to fully aromatic ring systems such as *p*-cymene (**13**), the constituent of ajowan oil.



(**12**)

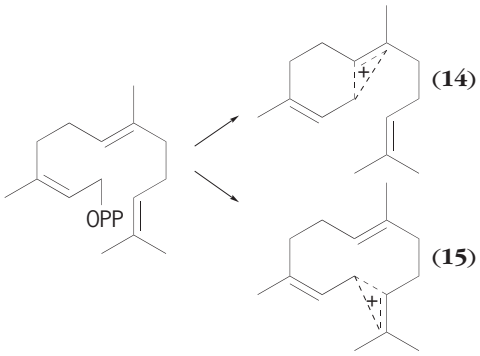


(**13**)

Monoterpenes are widely used in the flavor and perfume industries because of their attractive odors, low molecular weights, and high volatilities. Most are synthesized rather than extracted from plant sources.

Monoterpenes have been well studied, and most of them have been prepared in the laboratory. *See* CAMPHOR; MENTHOL; PINE TERPENE.

Sesquiterpenes. These are C_{15} hydrocarbons or their oxygenated analogs. The major categories are mentioned below. They arise from the cyclization of farnesylpyrophosphate and subsequent rearrangements of the resulting carbonium ions (**14** and **15**), as in reaction (3). Almost all known sesquiterpenes

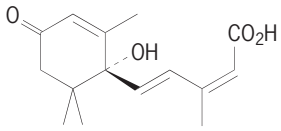


(3)

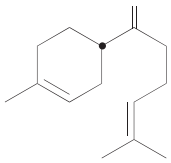
can be derived from these two cations.

Many sesquiterpenes are important constituents of the characteristic aromas of plant products, and many others have interesting or useful physiological properties.

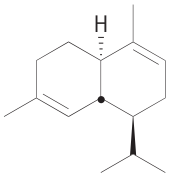
Farnesanes and eudesmanes. These are mono- and bicylic terpenes derived from the alkylative cyclization of farnesylprophosphate. They can be exemplified by the structures of abscisic acid (**16**), a plant growth regulator, zingiberine (**17**), the constituent of ginger oil, cadinene (**18**) from the oil cubebs, and α-santonine (**19**) from plants of the *Artemisia*.



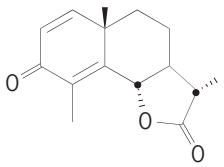
(**16**)



(**17**)

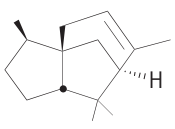


(**18**)

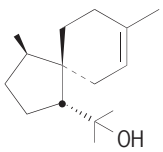


(**19**)

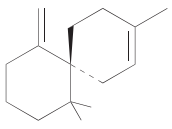
Acoranes, cedranes, chamigrane. These sesquiterpenes are characterized by spirocyclic skeletons. They are obtained from various species of wood, such as the Alaskan cedar. Examples are α-cedrane (**20**), β-acorenol (**21**), and β-chamigrene (**22**).



(**20**)

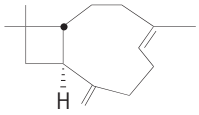


(**21**)

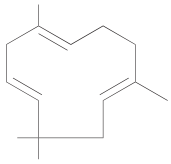


(**22**)

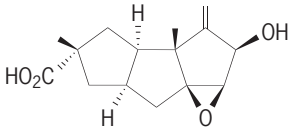
Caryophyllanes, illudanes, humulenes. These are biogenetically related macrocyclic terpenes; they are exemplified by caryophyllene (**23**) and humulene (**24**), which occur in hops. Related to illudanes is hirsutic acid (**25**).



(23)



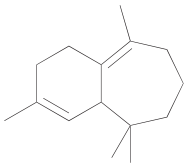
(24)



(25)

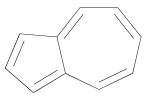
Himachalenes and longifolanes. These contain fused six- and seven-membered rings, and are derived from decalin skeletons via rearrangements; longifolene (**26**) and β-himachalene (**27**) are examples.

(26)

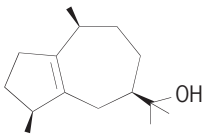


(27)

Perhydroazulenes. The fastest-growing group of identified sesquiterpenenes are the perhydroazulenes. The structures of these compounds were formerly determined by their oxidation to derivatives of the blue hydrocarbon azulene (**28**). A representative member of this class is quaiol (**29**) from quaiacum wood oil.

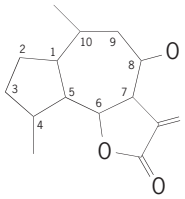


(28)



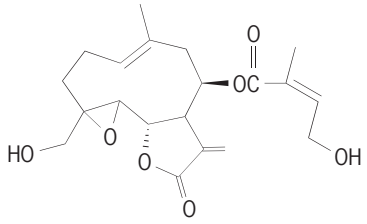
(29)

Quaianes and quaianolides. These constitute the largest single group of sesquiterpenes. New species continue to be isolated from fungi, marine organisms, or plants. The sesquiterpene lactones or quaianolides have remarkable cytoxic properties and are represented by a general structure (**30**).



(30)

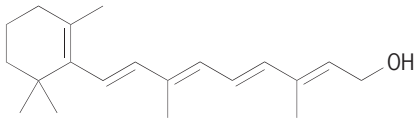
Germacranes. These are macrocyclic terpene lactones, exemplified by eupassopin (**31**). The quaianes, quaianolides, and germacranes have been shown to include a number of compounds with promising antitumor activity.



(31)

Diterpenes. These are C_{20} hydrocarbons, and their oxidized derivatives are composed of four isoprene units. They are synthesized by organisms from geranylgeranalpyrophosphate and are utilized in protective coatings of higher plants. Extraction of resins of various coniferous species yields a number of diterpenes. Their skeletons do not exhibit the variety shown by sesquiterpenoids. Abietic acid derivatives are the principal constituent of rosin, a resin obtained mainly from pine with a variety of industrial uses. Diterpene frameworks are found in many complex alkaloids of the delphinium or atisine types. Nitrogen is incorporated into the diterpene nucleus via a carboxylic acid function.

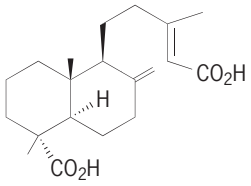
Mild antibiotic activity is associated with some diterpenes. The gibberellins are plant growth promoters. The best-known open-chain diterpene is vitamin A (**32**). The structure of other diterpenes



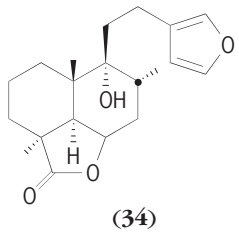
(32)

varies from bicyclic to tetracyclic skeletons.

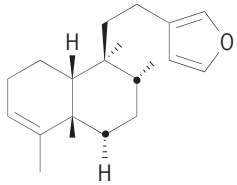
Labdanes and clerodanes. These are bicyclic resin acids isolated as the "bitter principles" of bark, roots, and stems. Agathic acid (**33**), marubiin (**34**) from horehound, and clerodane (**35**) are representatives.



(33)



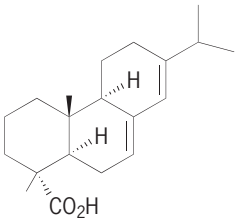
(34)



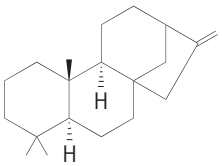
(35)

Tricyclic and tetracyclic diterpenes are typified by the structures of abietic acid (**36**; actually an artifact of isolation and not a natural product) from pine

resin, and phyllocladene (**37**), a typical diterpene constitutent of many essential oils.

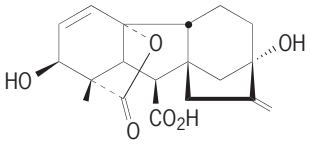


(**36**)



(**37**)

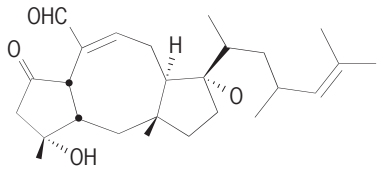
Gibberellins. The metabolites of a rice fungus, known as gibberellins, are exemplified by the structure of gibberellic acid (**38**).



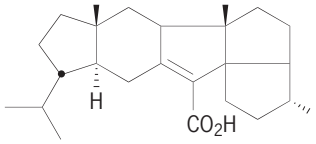
(**38**)

See GIBBERELLIN.

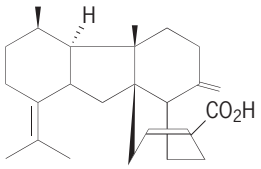
Sestquiterpenes. These are C_{25} compounds derived from geranylfarnesol pyrophosphate. They can be isolated from protective coatings and waxes of insects or from various fungal sources. The first-known member of the sesquiterpenes was isolated in 1965. Some examples are ophiobolin A (**39**), retigeranic acid (**40**), and gastardic acid (**41**).



(**39**)



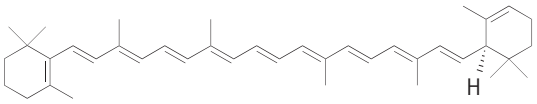
(**40**)



(**41**)

Triterpenes. These have 30 carbon atoms and are composed of six isoprene units. They form the largest group of terpenes. *See* TRITERPENE.

Tetraterpenes. These are C_{40} or higher terpenes. They have a large number of polyene units and therefore do not possess the stereochemical and structural complexity of lower terpenes. Biogenetically, tetraterpenes arise by the dimerization of C_{20} units in a tail-to-tail fashion, with concomitant cyclizations restricted to the termini. Most known tetraterpenes are carotenoid pigments, exemplified by α-carotene (**42**). Carotenes and various oxygenated analogs,



(**42**)

found in the chloroplasts of all green plants and in some algae, serve as accessory pigments for photosynthesis. Vitamin A may be regarded as a degraded carotene. *See* CAROTENOID; VITAMIN A.

Polyterpenes. These are various natural rubbers from *Hevea*, *Guayule*, and other plants and usually consist of 500–5000 cis-linked isoprene units; guttapercha contains 100 trans-linked isoprene residues. These polymers arise by a genetic error characteristic of only about 1% of the plant kingdom, whereby species possess the ability to channel all of their mevalonate biogenesis into polymerization. *See* ESSENTIAL OILS; ROSIN; RUBBER.

Tomas Hudlicky

Bibliography. K. F. Albizati et al., *Synthesis of Marine Natural Products*, vol. 1: *Terpenoids*, 1992; Chemical Society, London, *Terpenoids and Steroids*, vols. 1–8, Specialist Periodical Reports, 1978; J. S. Glasby, *Encyclopedia of the Terpenoids*, 1983; J. B. Harborne and F. A. Tomas-Barberan, *Ecological Chemistry and Biochemistry of Plant Terpenoids*, 1991; T.-L. Ho, *Carbocyclization: Implications in Terpene Synthesis,* 1988.

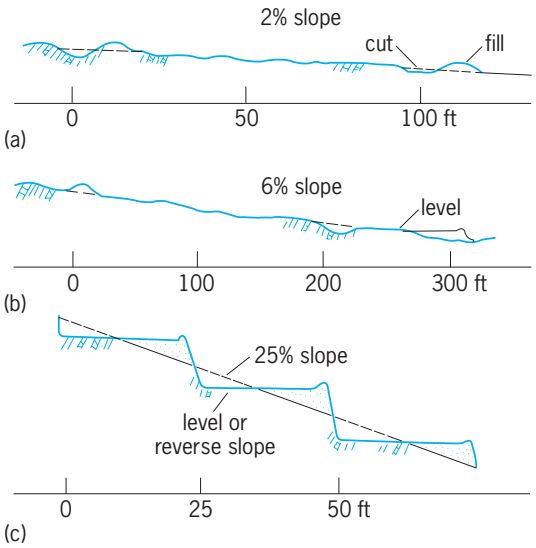
Terracing (agriculture)

A method of shaping land to control erosion on slopes of rolling land used for cropping and other purposes. In early practice the land was shaped into a series of nearly level benches or steplike formations. Modern practice in terracing, however, consists of the construction of low-graded channels or levees to carry the excess rainfall from the land at nonerosive velocities. The physical principle involved is that, when water is spread in a shallow stream, its flow is retarded by the roughness of the bottom of the channel and its carrying, or erosive, power is reduced. Since direct impact of rainfall on bare land churns up the soil and the stirring effect keeps it in suspension in overland flow and rills, terracing does not prevent sheet erosion. It serves only to prevent destruction of agricultural land by gullying and must be supplemented by other erosion-control practices, such as grass rotation, cover crops, mulching, contour farming, strip cropping, and increased organic matter content.

In areas of low rainfall and absorbent soils, nearly level terraces are used to retain the runoff and conserve soil moisture, thus preventing wind erosion of the soil. *See* EROSION; SOIL CONSERVATION.

Types of terraces. The two major types of terraces are the bench and the broadbase (see **illus.**).

Bench. This is essentially a steep-land terrace and consists of an almost vertical retaining wall, called a riser, or a steep vegetative slope to hold the nearly level surface of the soil for cultivation, orchards, vineyards, or landscaping. The bench terrace has



Types of terraces. (*a*) Broadbase. (*b*) Conservation bench. (*c*) Bench. 1 ft = 0.3 m. (*After Soil and Water Conservation Engineering, 2d ed., The Ferguson Foundation Agricultural Engineering Series, John Wiley and Sons, 1966*)

been in use all over the world for the past several thousand years, particularly in Europe, Australia, and Asia. Bench terraces are adaptable to slopes of 25–30% and are costly to construct. For the most part, bench terraces have been abandoned for general farming because they are not adaptable to the efficient use of modern-day farm equipment.

Modern mechanized farming greatly increased the need for erosion control; at the same time, however, modern machinery furnished the power for increased terracing. The early United States practice of constructing hillside ditches across the slopes of fields to prevent up- and downhill gullying was followed by the development of the more easily controlled levee or ridge. This was called a narrow-base ridge terrace and was, in some cases, vegetated and developed into a bench terrace. By plowing to this narrow levee and maintaining the drainage channel, its base was widened until it became known as a broadbase terrace.

Broadbase. This terrace has the distinguishing characteristic of farmability; that is, crops can be grown on this terrace and worked with modern-day machinery. These terraces are constructed either to remove or retain water and, based on their primary function, are classified either as graded or level.

A graded broadbase terrace is constructed from the upper side only with a variable or uniform grade in the channel to remove water at an acceptable rate. This is sometimes called the Nichols or channel terrace.

A level broadbase terrace is constructed on the contour and from both sides. It is recommended in areas where the soil is permeable to prevent overtopping. This terrace is sometimes called the Mangum or ridge terrace.

Another terrace in use today in semiarid regions where maximum moisture conservation is needed is the Zingg conservation bench terrace. It consists of an earthen embankment and a very broad, flat channel resembling a level bench. The runoff store in the channel comes from the sloping area above the bench which extends upslope to the next terrace.

The design of any type terrace must take into consideration factors such as soil characteristics, the cropping system to be used, soil management practices, climatic conditions, and others. All influence the effectiveness of the terrace system.

Terrace outlets. Since terrace channels concentrate rainfall on hillsides, outlets are a major feature of any successful terrace system. There are many different schemes for outlet construction. Masonry structures, such as storm sewers, concrete flumes, or drop inlets, may be used on steep land. Vegetation, such as grass or other thick-growing crops, may be used on gentle slopes. Modern practice in outlet construction calls for the use of natural channels with careful shaping and vegetating before large concentrations of water are turned into them.

Parallel terraces. Parallel terraces have been a major change in the layout of the terracing system. The object of parallel terracing is to facilitate the use of modern farm machinery by the elimination of point rows. Point rows in a field cause an excessive loss of productive time because of the turning of equipment. They also cause the destruction of crops and terraces by making it necessary for machinery to pass over them in their course of turning. Parallel terraces are made possible by a relatively small amount of land forming, such as moving land from high to low spots in the field and smoothing. This modification allows the efficient use of 4-, 6-, and 8-row cultivating equipment.

Construction equipment. Terracing equipment may be classified according to the method by which it moves the soil: lift and roll, throw, scrape and push, and carry.

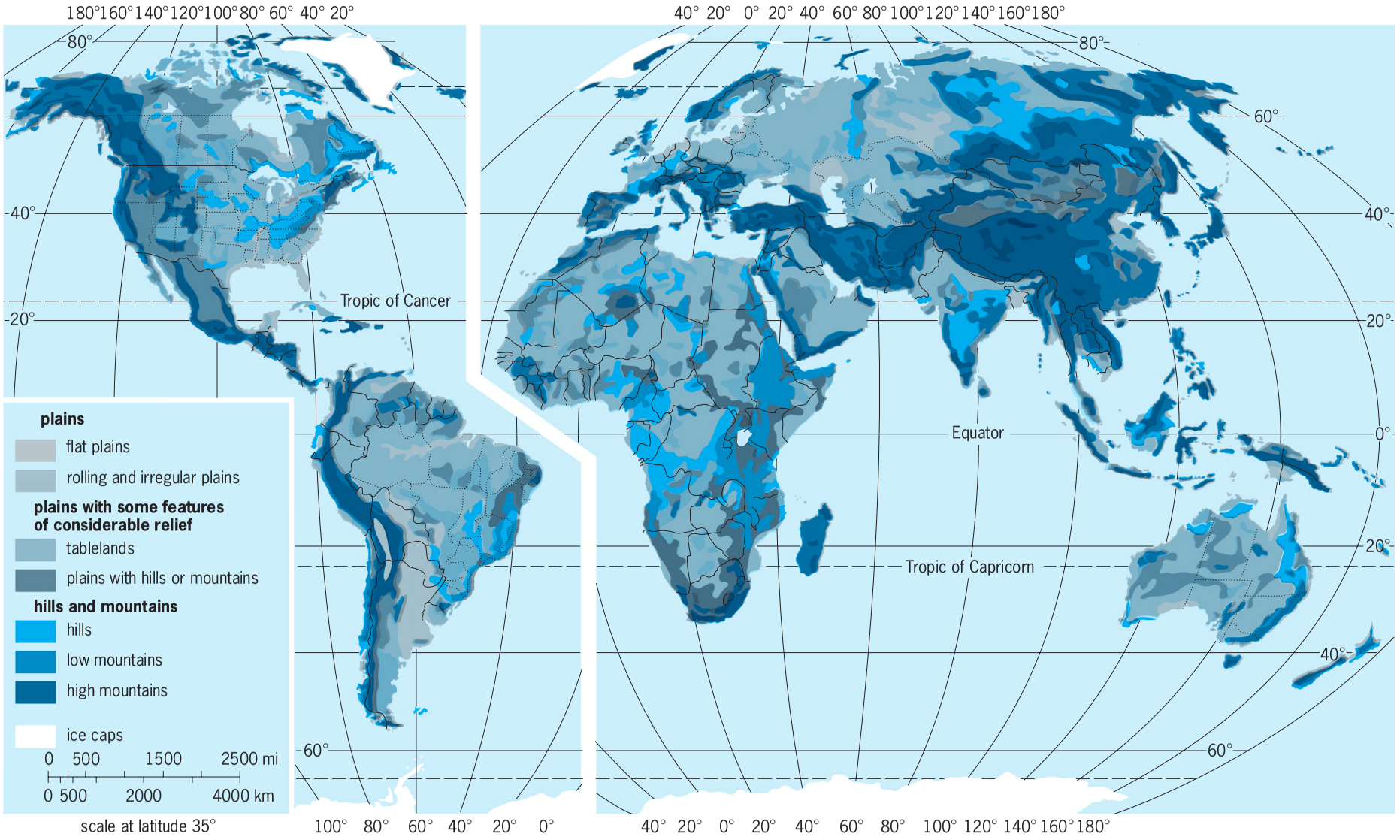
The ideal terracing machine will meet certain requirements, such as (1) displacement of soil laterally to the desired position in the ridge; (2) placing the topsoil on or near the top of the ridge; (3) high work speed; (4) operation on slopes up to 15–20% and (5) low initial and operating cost. *See* AGRICULTURAL SOIL AND CROP PRACTICES; LAND DRAINAGE (AGRICULTURE).

Charles B. Ogburn

Terrain areas

Subdivisions of the continental surfaces distinguished from one another on the basis of the form, roughness, and surface composition of the land. These areas of distinctive landforms are the product of various combinations and sequences of events involving both deformation of the Earth's crust and surficial erosion and deposition by water, ice, gravity, and wind. The pattern of landform differences is strongly reflected in the arrangement of such other features of the natural environment as climate, soils, and vegetation. These regional associations must be carefully reckoned with by anyone planning activities as diverse as agriculture, transportation, city development, and military operations.

The **illustration** distinguishes among eight



Distribution of terrain classes over the Earth. The subdivisions are intended to bring out only the most striking contrasts among land surfaces. Percentage of area occupied by each class is given in the text. (*After G. T. Trewartha, A. H. Robinson, and E. H. Hammond, Elements of Geography, 5th ed., McGraw-Hill, 1967*)

classes of terrain, on the basis of steepness of slopes, local relief (the maximum local difference in elevation), cross-sectional form of valleys and divides, and nature of the surface material. Approximate definitions of terms used and percentage figures indicating the fraction of the world's land area occupied by each class are as follows: (1) flat plains: nearly level land, slight relief, 4%; (2) rolling and irregular plains: mostly gently sloping, low relief, 30%; (3) tablelands: upland plains broken at intervals by deep valleys or escarpments, moderate to high relief, 5%; (4) plains with hills or mountains: plains surmounted at intervals by hills or mountains of limited extent, 15%; (5) hills: mostly moderate to steeply sloping land of low to moderate relief, 8%; (6) low mountains: mostly steeply sloping, high relief, 14%; (7) high mountains: mostly steeply sloping, very high relief, 13%; and (8) ice caps: surface material, glacier ice, 11%.

The continents differ considerably. Australia, the smoothest continent, has only one-fifth of its area occupied by hill and mountain terrain as against one-third of North America and more than one-half of Eurasia. Antarctica is largely ice covered; the only other great ice cap is on Greenland.

North America, South America, and Eurasia are alike in that most of their major mountain systems are linked together in extensive cordilleran belts. These form a broken ring about the Pacific basin, with an additional arm extending westward across southern Asia and Europe. The principal plains of Eurasia and the Americas lie on the Atlantic and Arctic sides of the cordilleras, but are in part separated from the Atlantic by lesser areas of rough terrain.

Most of Africa and Australia, together with the eastern uplands of South America and the peninsulas of Arabia and India, show great similarity to one another. They lack true cordilleran belts, and are composed largely of upland plains and tablelands, locally surmounted by groups of hills and mountains, and in many places descending to the sea in rough, dissected escarpments. *See* CONTINENT; CORDILLERAN BELT; HILL AND MOUNTAIN TERRAIN; PLAINS.

Edwin H. Hammond

Bibliography. *Hammond World Atlas*, 4th ed., 2002; T. L. McKnight and D. Hess, *Physical Geography: A Landscape Appreciation*, 8th ed., 2004; A. H. Strahler and A. Strahler, *Introducing Physical Geography*, 4th ed., 2005.

Terrestrial coordinate system

The perpendicular intersection of two curves or two lines, one relatively horizontal and the other relatively vertical, is the basis for finding and describing terrestrial location. The Earth's graticule, consisting of an imaginary grid of east-to-west-bearing lines of latitude and north-to-south bearing lines of longitude, is derived from the Earth's shape and rotation, and is rooted in spherical geometry. The development of latitude and longitude likely originated during the classical period in Greece in the second century B.C. Hipparchus of Rhodes is thought to have been instrumental in its development. Plane coordinate systems, equivalent to horizontal X and vertical Y coordinates, are based upon cartesian geometry and differ from the graticule in that they have no natural origin or beginning for their grids.

Latitude and longitude. The Earth, which is essentially a sphere, rotates about an axis that defines the geographic North and South poles. The poles serve as the reference points on which the system of latitude and longitude is based (see **illus.**). *See* LATITUDE AND LONGITUDE.

Latitude is arc distance (angular difference) from the Equator and is defined by a system of parallels, or lines that run east to west, each fully encompassing the Earth. The Equator is the parallel that bisects the Earth into the Northern and Southern hemispheres, and lies a constant 90° arc distance from both poles. As the only parallel to bisect the Earth, the Equator is considered a great circle. All other parallels are small circles (do not bisect the Earth), and are labeled by their arc distance north or south from the Equator and by the hemisphere in which they fall. Parallels are numbered from 0° at the Equator to 90° at the poles. For example, 42°S describes the parallel 42 degrees arc distance from the Equator in the Southern Hemisphere. For increased location precision, degrees of latitude and longitude are further subdivided into minutes ($1^\circ = 60'$) and seconds ($1' = 60''$). *See* EQUATOR; GREAT CIRCLE, TERRESTRIAL.

Longitude is defined by a set of imaginary curves extending between the two poles, spanning the Earth. These curves, called meridians, always point to true geographical north (or south) and converge at the poles. In the present-day system of longitude, meridians are numbered by degrees east or west of the beginning meridian, called the Prime Meridian or the Greenwich Meridian, which passes through the Royal Observatory in Greenwich, England. The Prime Meridian was selected at an international conference in Washington, DC, in 1884 and assigned a longitude of 0°.

Since the Earth is a fundamentally a sphere, its circumference describes a circle containing 360°, the arc distance through which the Earth rotates in 24 hours. The arc distance from the Prime Meridian describes the location of any meridian (see illus.). The 180° meridian is commonly referred to as the International Dateline. Together, the Prime Meridian and the International Dateline describe a great circle that bisects the Earth, as do all other meridian circles. The west half of the Earth, located between the Prime Meridian and the International Dateline, comprises the Western Hemisphere, and the east half on the opposite side forms the Eastern Hemisphere. Meridians within the Western Hemisphere are labeled with a W, and meridians within the Eastern Hemisphere are labeled with an E. A complete description of longitude includes an angular measurement and a hemispheric label. For example, 78° W is the meridian 78° west of the Prime Meridian. Neither the 0° meridian (Prime) nor the 180° meridian (Dateline) is given a hemispheric suffix because they divide the two hemispheres, and therefore do not belong to either one.

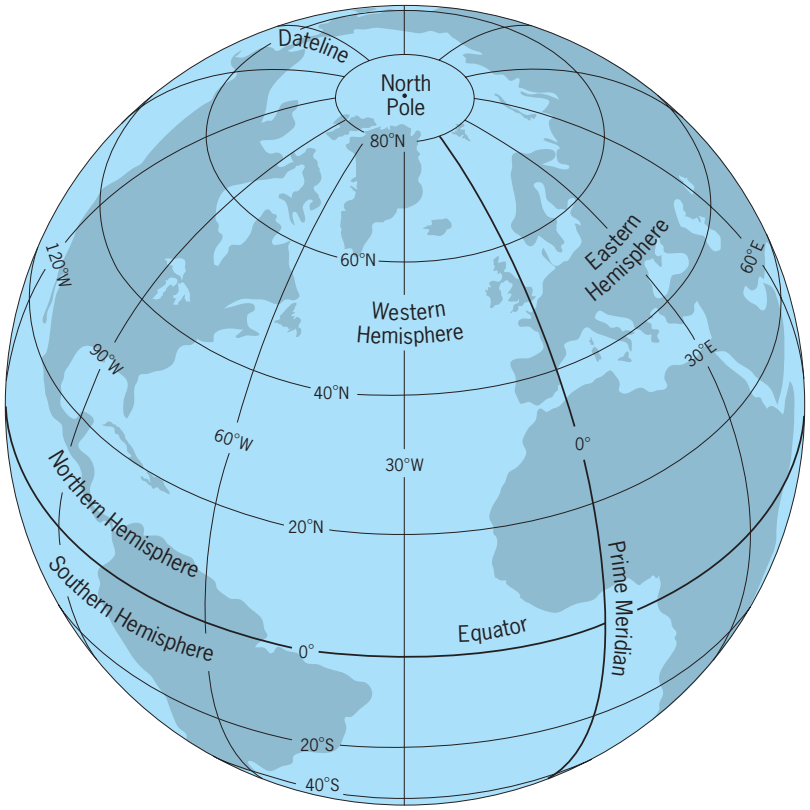
Interaction of latitude and longitude. As parallels approach the poles, their circumferences shorten, while meridians converge as they approach the poles. Consequently, the length of a degree of latitude is shorter the closer it is to a pole. At the Equator, the length of a degree of longitude is greatest, and is equal to 69.17 mi (110 km). At 50° a degree of longitude equals 44.5 mi (71.7 km), and at 80° a degree of longitude equals 12.0 mi (19.4 km). The length of one degree of longitude at given latitude equals the cosine of the latitude times the length of a degree of longitude at the Equator (69.17 mi). In comparison, one degree of latitude is approximately 69 mi everywhere on the Earth because all meridians are of equal length.

The complete description of a location using the graticule describes a point where a parallel intersects a meridian. For example, the description 42°S, 78° W provides sufficient information to find that particular location on a map of the Earth.

Determining latitude and longitude. The empirical determination of latitude and longitude has a long history. For thousands of years it was common navigational knowledge that latitude can be measured through the observation of fixed stellar bodies, such as the Sun or North Star, using a sextant or any device that measures arc distance from the Earth's horizon to the star. No such means for measuring longitude existed until an accurate chronometer (clock) was invented in 1774. Before this, historical navigation maps had great discrepancies between the accuracy of latitude and longitude, where N-S locations showed good accuracy, but E-W locations were often in error. Unlike latitude, longitude has no natural starting point on the Earth. Longitude must be determined with respect to the difference in local time—the actual solar time at any location—between arbitrarily chosen Prime Meridian and the observation position, knowing that the Earth rotates at a relatively constant speed (1 degree of longitude in 4 minutes). *See* CELESTIAL NAVIGATION; CHRONOMETER; NAVIGATION; SEXTANT; TIME.

In the twentieth century, radio signals from Greenwich indicating the exact time at the Prime Meridian increased the accuracy of longitude determination. Today, precise measurement of latitude, longitude, and altitude is made possible through GPS (Global Positioning System) technology. GPS includes a constellation of 24 satellites orbiting the Earth which continually transmit location information and time to hand-held receivers on the Earth. Some GPS receivers are capable of determining Earth positions within centimeters. *See* SATELLITE NAVIGATION SYSTEMS.

Plane coordinates. Coordinate system alternatives to the graticule evolved in the early twentieth century because of the complexity of using spherical geometry in determining latitude, longitude, and direction. Plane (two-dimensional) or cartesian coordinate systems presume, with a small amount of error, that a relatively nonspherical Earth exists in smaller areas. Plane coordinates are superimposed upon these small areas, with coordinates being determined by the equivalent of a grid composed of a number of parallel vertical lines (X) and a complementary set of parallel horizontal lines (Y). All plane coordinate systems select a false origin for the grid because, unlike the graticule, there is no natural starting point to choose as an origin.



Earth's graticule. Meridians of longitude run from north to south, but are measured east or west of the Prime Meridian. Parallels run from east to west, but are measured north or south of the Equator.

State plane coordinate system (SPC). This system, used only in the United States, partitions each state into zones, each of which has its own coordinate system that slightly overlaps adjacent zones to facilitate connections to adjacent zones. The number of zones designated in each state is determined by the size of the state. Illinois, for example, has two zones, whereas Texas has five. Zone boundaries follow, with exceptions made for political subdivisions and natural features, either meridians or parallels depending on the shape of the state. All measurements are made in feet from a zone's false origin.

Universal Transverse Mercator (UTM) system. The UTM system is a worldwide coordinate system in which locations are expressed using metric units. The basis for the UTM system is the Universal Transverse Mercator map projection. This projection becomes vastly distorted in polar areas above 80°, and for this reason the UTM system is confined to extend from 84°N to 80°S. The UTM system partitions the Earth into 60 north-south elongated zones, each having a width of 6° of longitude. UTM measurements (in meters) are made in reference to a false origin located at the western edge of the zone and the Equator in the Northern Hemisphere or 90°S in the Southern Hemisphere. *See* MAP PROJECTIONS.

A number of other coordinate systems are in use

today. Foremost among these are the U.S. Public Land Survey System, the Universal Polar Stereographic (UPS) system, and the World Geographic Reference (GEOREF) system. Stephen Lavin

Bibliography. L. M. Bugayevskiy and J. P. Snyder, *Map Projections*, 1995; J. Campbell, *Map Use and Analysis*, 4th ed., 2000; A. H. Robinson et al., *Elements of Cartography*, 6th ed., 1995; N. J. W. Thrower, *Maps and Civilization*, 2d ed., 1999.

Terrestrial ecosystem

A community of organisms and their environment that occurs on the land masses of continents and islands. Terrestrial ecosystems are distinguished from aquatic ecosystems by the lower availability of water and the consequent importance of water as a limiting factor. Terrestrial ecosystems are characterized by greater temperature fluctuations on both a diurnal and seasonal basis than occur in aquatic ecosystems in similar climates, because water has a high specific heat, a high heat of vaporization, and a high heat of fusion compared with the atmosphere, all of which tend to ameliorate thermal fluctuations. The availability of light is greater in terrestrial ecosystems than in aquatic ecosystems because the atmosphere is more transparent than water. Gases are more available in terrestrial ecosystems than in aquatic ecosystems. Those gases include carbon dioxide that serves as a substrate for photosynthesis, oxygen that serves as a substrate in aerobic respiration, and nitrogen that serves as a substrate for nitrogen fixation. Terrestrial environments are segmented into a subterranean portion from which most water and ions are obtained, and an atmospheric portion from which gases are obtained and where the physical energy of light is transformed into the organic energy of carbon-carbon bonds through the process of photosynthesis.

Extent. The Earth has an estimated surface area of 197,272,000 mi^2 (510,934,000 km^2) of which about 29%, or 57,200,000 mi^2 (148,100,000 km^2), is occupied by land. About 1,540,000 mi^2 (4,000,000 km^2) of the land surface is occupied by fresh-water ecosystems of lakes, streams, and marshes; so terrestrial ecosystems occupy 55,660,000 mi^2 (144,150,000 km^2), or 28.2%, of Earth's surface.

Earth and the other planets are believed to be about 4.5 billion years old; the earliest fossil life forms are about 3.5 billion years old. The first terrestrial organisms appeared in the Silurian Period, about 425 million years ago. Therefore, terrestrial ecosystems have an age no greater than 9.4% of the total age of the Earth and 12.1% of the duration of life on Earth.

Principal organisms. Although they are comparatively recent in the history of life and occupy a much smaller portion of Earth's surface than marine ecosystems, terrestrial ecosystems have been a major site of adaptive radiation of both plants and animals. Major plant taxa in terrestrial ecosystems are members of the division Magnoliophyta (flowering plants), of which there are about 225,000 species in the class Magnoliopsida (dicots) and 50,000 species in the class Liliopsida (monocots), and the division Pinophyta (conifers), of which there are about 500 species. Members of the division Bryophyta (mosses and liverworts), of which there are about 24,000 species, are also important in some terrestrial ecosystems. Major animal taxa in terrestrial ecosystems include the classes Insecta (insects) with about 900,000 species, Aves (birds) with 8500 species, and Mammalia (mammals) with approximately 4100 species. *See* SYSTEMATICS; PLANT TAXONOMY; TAXONOMY.

Organisms in terrestrial ecosystems have adaptations that allow them to obtain water when the entire body is no longer bathed in that fluid, means of transporting the water from limited sites of acquisition to the rest of the body, and means of preventing the evaporation of water from body surfaces. They also have traits that provide body support in the atmosphere, a much less buoyant medium than water, and other traits that render them capable of withstanding the extremes of temperature, wind, and humidity that characterize terrestrial ecosystems. Finally, the organisms in terrestrial ecosystems have evolved many methods of transporting gametes in environments where fluid flow is much less effective as a transport medium.

Energy and chemical flow. The organisms in terrestrial ecosystems are integrated into a functional unit by specific, dynamic relationships due to the coupled processes of energy and chemical flow. Those relationships can be summarized by schematic diagrams of trophic webs, which place organisms according to their feeding relationships. The base of the food web is occupied by green plants, which are the only organisms capable of utilizing the energy of the Sun and inorganic nutrients obtained from the soil to produce organic molecules. The total living mass present in an ecosystem at any given time is referred to as biomass. The change in biomass through time is referred to as net productivity. Productivity due to the green plants and microorganisms is termed primary productivity while that due to animals is secondary productivity. Productivity will always diminish from stage to stage in the trophic web due to the laws of thermodynamics. This limitation is summarized by the equation $P_g = P_n + R$, where P_g is gross productivity, or total energy entering a trophic level, P_n is the net productivity of that trophic level, and R is the respiratory cost of maintaining that trophic level. For green plants, P_g is the energy initially captured in photosynthesis, R is the cost of maintaining the organisms, and P_n is the energy remaining to produce a biomass increment. Since it is only this biomass increment that is available to higher trophic levels, net productivity must always diminish at progressively higher trophic levels. *See* BIOLOGICAL PRODUCTIVITY; BIOMASS; FOOD WEB; PHOTOSYNTHESIS.

Terrestrial food webs can be broken into two segments based on the status of the plant material that enters them. Grazing food webs are associated with the consumption of living plant material by herbivores. Detritus food webs are associated with the consumption of dead plant material by detritivores.

Area and productivity of the various terrestrial ecosystems*

Type	Area occupied, 10^6 km^2	Mean net primary productivity, g/(m^2)(year)	Total annual productivity, 10^9 tons/year
Tundra	8.0	140	1.1
Desert	42.0	40	1.7
Temperate grassland	9.0	600	5.4
Woodland and shrubland	8.5	700	6.0
Cultivated land	14.0	650	9.1
Boreal forest	12.0	800	9.6
Savanna	15.0	900	13.5
Temperate forest	12.0	1250	14.9
Tropical forest	24.5	2000	49.4

*1 km^2 = 0.386 mi^2; 1 g/m^2 = 0.029 oz/yd^2; 1 ton = 0.9 metric ton.
source: After R. H. Whittaker, *Communities and Ecosystems,* 2d ed., Macmillan, 1975.

The relative importance of those two types of food webs varies considerably in different types of terrestrial ecosystems. Grazing food webs are more important in grasslands, where over half of net primary productivity may be consumed by herbivores. Detritus food webs are more important in forests, where less than 5% of net primary productivity may be consumed by herbivores. *See* SOIL ECOLOGY.

Energy flow is unidirectional in ecosystems, with a portion of that energy dissipated at each trophic level. As energy is dissipated, nutrients are released back into the environment. Nutrient flow, therefore, is cyclic. Energy flow and nutrient flow are impossible to separate functionally since they are intimately coupled. Energy flow will depend upon the rate at which nutrients are recycled, and nutrient flow will depend upon the availability of energy. *See* BIOGEOCHEMISTRY; ECOLOGICAL ENERGETICS.

Seasonality. Because of the marked seasonality of most of the terrestrial ecosystems, due to either temperature fluctuations in temperate locations or rainfall fluctuations in tropical locations, there are corresponding fluctuations in net primary productivity. The plants may sometimes be able to complete their life cycle in a brief time period, as do annual plants in the desert, or they may undergo periods of dormancy, as do deciduous trees in temperate forests. Similarly, animals either may hibernate during unfavorable periods or may migrate to other locations where conditions are more favorable. Many birds, for example, migrate from temperate latitudes during summer to tropical latitudes during winter in the temperate zones. Others may even migrate between the respective summer seasons in north and south temperate zones. Those animals that remain active in terrestrial ecosystems during periods of low primary productivity must have adaptations that allow them to subsist during unfavorable periods, often storing fat during productive periods and depleting those reserves during unfavorable periods.

Types. There is one type of extensive terrestrial ecosystem due solely to human activities and eight types that are natural ecosystems. Those natural ecosystems reflect the variation of precipitation and temperature over Earth's surface (see **table**). The smallest land areas are occupied by tundra and temperate grassland ecosystems, and the largest land area is occupied by tropical forest. The most productive ecosystems are temperate and tropical forests, and the least productive are deserts and tundras. Cultivated lands, which together with grasslands and savannas utilized for grazing are referred to as agroecosystems, are of intermediate extent and productivity. Because of both their areal extent and their high average productivity, tropical forests are the most productive of all terrestrial ecosystems, contributing 45% of total estimated net primary productivity on land. Due to their importance in productivity, tropical forests are believed to play an important role in the global carbon budget. There is increasing concern that deforestation of tropical ecosystems associated with agricultural development may contribute to increasing atmospheric carbon dioxide concentrations that may lead to climatic change. *See* DESERT; ECOLOGICAL COMMUNITIES; ECOSYSTEM; FOREST AND FORESTRY; GRASSLAND ECOSYSTEM; SAVANNA; TUNDRA.

S. J. McNaughton

Bibliography. W. B. Clapham, Jr., *Natural Ecosystems*, 2d ed., 1983; E. J. Kormondy, *Concepts of Ecology*, 3d ed., 1984; E. P. Odum, *Basic Ecology*, 1983; R. E. Ricklefs, *Ecology*, 3d ed., 1989; R. H. Young, S. Cousins, and D. Green (eds.), *Landscape Ecology and Geographical Information Systems*, 1993.

Terrestrial radiation

Electromagnetic radiation emitted from the Earth and its atmosphere. Terrestrial radiation, also called thermal infrared radiation or outgoing longwave radiation, is determined by the temperature and composition of the Earth's atmosphere and surface. The atmosphere is composed of two groups of gases, one with a nearly permanent concentration, consisting principally of nitrogen (N_2) and oxygen (O_2) molecules, and another with variable concentrations of other gases. Although considered to be permanent constituents, carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), and carbon monoxide (CO) have been observed to increase in association with anthropogenic activities. One of the principal variable gases is water (H_2O) vapor, the major compound that modulates the hydrological cycle involving evaporation, cloud formation, and precipitation. Water vapor concentration decreases rapidly with

latitude, almost following an exponential function. Ozone (O_3) concentration also varies significantly with space and time and occurs principally at altitudes of about 15–30 km (10–20 mi). A significant variable gas is a mixture of chlorofluorocarbons (CFCs) produced by industrial activities. *See* ATMOSPHERE; ATMOSPHERIC GENERAL CIRCULATION; HYDROLOGY; RADIATIVE TRANSFER.

The atmosphere also contains aerosol particles ranging in size from about 10^{-3} to 20 micrometers that are known to be produced by natural processes as well as by human activity. Aerosol concentrations generally decrease rapidly with height. Some aerosols promote condensation and ice nuclei, upon which cloud particles may form. Clouds are global in nature and regularly cover more than 50% of the Earth. There are various types of clouds. Some clouds, such as high-level cirrus in the tropics and low-level stratus in the Arctic and near coastal areas, are climatologically persistent. Some clouds generate precipitation. *See* AEROSOL; CLOUD; CLOUD PHYSICS.

The temperature structure of the Earth and the atmosphere is a result of numerous physical, chemical, and dynamic processes. In a one-dimensional context, the temperature structure is determined by the balance between radiative and convective processes. From the surface to a height of about 10 km (6 mi), the temperature decreases at a typical rate of about 6.5°C (12°F) per kilometer; this region is referred to as the troposphere, where major weather events occur. Variability of the global temperature field is modulated by the transport of various types of energy by the winds, the radiative energy emitted from the Sun and received by the Earth, and the radiative energy emitted from the Earth and the atmosphere. *See* HEAT BALANCE, TERRESTRIAL ATMOSPHERIC; SOLAR RADIATION; SUN; TROPOSPHERE; WIND.

Emissivities. The Earth's surface emits electromagnetic radiation according to the laws that govern a blackbody or a graybody. A blackbody absorbs the maximum radiation and at the same time emits that same amount of radiation so that thermodynamic equilibrium is achieved, defining a uniform temperature. The rate at which emission takes place is a function of temperature and wavenumber (cm^{-1}) or wavelength (μm) according to Planck's law. A graybody is characterized by incomplete absorption and emission and is said to have emissivity less than unity. The thermal infrared emissivities from water and land surfaces are normally 90–95%. It is usually assumed that the Earth's surfaces are approximately black in the analysis of infrared radiative transfer. Exceptions include snow and some sand surfaces whose emissivities are wavelength-dependent and could be less than 90%. *See* BLACKBODY; GRAYBODY; HEAT RADIATION; PLANCK'S CONSTANT.

Absorption and emission of radiation by atmospheric molecules are more complex and require a fundamental understanding of quantum mechanics. A molecule, composed of atoms, can rotate, or revolve, about an axis through its center of gravity, and therefore has rotational energy. The atoms of the molecule are bounded by certain forces like springs in which the individual atoms can vibrate about their equilibrium positions relative to one another. The molecule therefore also has vibrational energy. It is also possible for the energy of a molecule to change due to a change in the energy state of the electrons of which it is composed. Thus the molecule has electronic energy. These three forms of energy are said to be quantized, and assume only discrete values. Absorption and emission of radiation take place when the molecules undergo transitions from one energy state to another, and these transitions are generally governed by selection rules in quantum mechanics. Rotational energy changes are relatively small, and many of the rotational energy levels are populated at terrestrial temperatures. Changes in vibrational energy are much larger than the minimum changes in rotational energy. Thus, vibrational transitions never occur alone, but are coupled with simultaneous rotational transitions, producing a group of absorption lines known as the vibrational–rotational band in the thermal infrared spectrum of the Earth's atmosphere associated with atmospheric gases. These lines have finite widths and overlap one another. The line shape is determined by molecular collision in the lower atmosphere and by random thermal motion in the upper atmosphere. The width of the line is primarily dependent upon pressure but is also affected by temperature. *See* INFRARED SPECTROSCOPY; QUANTUM MECHANICS; SELECTION RULES (PHYSICS).

The water molecule is composed of two hydrogen atoms and one oxygen atom that form an isosceles triangle that is obtuse, referred to as an asymmetric top configuration. This configuration produces the pure water rotational band ranging 0–1000 cm^{-1}, which is important in the generation of tropospheric cooling. The band located at 1594.78 cm^{-1} (6.25 μm) is the vibrational–rotational band of water that has been used for remote sensing of its concentration from satellites. Water vapor also exhibits less selective absorption in the region 800–1200 cm^{-1}, the so-called thermal infrared window. Carbon dioxide has a linear symmetrical configuration, with the carbon atom in the middle and oxygen atoms on each side. The perpendicular vibration of oxygen atoms coupled with rotational transitions produces the 15-μm CO_2 band, known for causing greenhouse warming. Ozone has an asymmetric top configuration similar to water and exhibits absorption in the 9.6-μm region. Methane and nitrous oxide also show strong absorption bands in the 7–8-μm region. Other gases with absorption bands in the thermal infrared region are sulfur dioxide, ammonia, and chlorofluorocarbons (CFC). The above gases are referred to as greenhouse gases because of their ability to absorb or trap the radiation emitted by the Earth and the atmosphere. *See* CARBON DIOXIDE; GREENHOUSE EFFECT; METHANE; WATER.

Clouds are global in nature and can interact with terrestrial radiation. Clouds are composed of waterdroplets or ice crystals and can both reflect and transmit the radiation emitted from the surface and the

atmosphere while emitting infrared radiation according to the temperature structure within them. If the cloud as a whole is a blackbody, it will behave just like the Earth's surface. In this case, radiation from below and above the cloud would not be able to penetrate the cloud. The emitted radiance at the cloud top or bottom is governed by Planck's law. Most clouds that are composed of waterdroplets are black clouds because of high number concentrations and small droplet sizes. However, clouds that are composed of ice crystals with sizes ranging from a few micrometers to 1000 μm, such as cirrus clouds, are generally nonblack. Determination of the radiative properties of cirrus clouds in conjunction with remote sensing and climate studies is a subject of contemporary research.

Radiation measurement. **Figure 1** displays the principal absorbing gases that are identified in an observed infrared spectrum. This spectrum was obtained from a scanning high-resolution interferometer sounder, an instrument that measures the emitted thermal infrared radiation of 3.3–18 μm from the NASA high-flying ER-2 aircraft at about 20 km (12 mi) with a spectral resolution on the order of 0.01 cm^{-1}. The spectrum is presented in terms of the brightness temperature (K) as a function of wavenumber (cm^{-1}) and wavelength (μm). Identified in Fig. 1 are the 15-μm CO_2 band, 9.6-μm O_3 band, 6.25-μm H_2O vibrational–rotational band, CH_4 and N_2O in the 7–8-μm region, two CFC bands, a part of the H_2O rotational band, and the thermal infrared window of 800–1200 cm^{-1} discussed previously. In addition to these bands in the thermal infrared, the 4.3-μm CO_2 and the spectral bands for CO, N_2O, and CH_4 at wavelengths less than 5 μm are shown. Note that because the temperature structure of the Earth–atmosphere system mentioned above, its emission (terrestrial radiation) is primarily confined to wavelengths longer than about 5 μm, while the radiation from the Sun is largely in the shorter wavelength region from about 0.2 to 5 μm, referred to as solar or shortwave radiation.

K. N. Liou

Meteorological satellites. Terrestrial radiation originating from the Earth–atmosphere–ocean system, as well as solar radiation reflected and scattered back to space, is measured on a daily basis by meteorological satellites. Instruments on meteorological satellites measure visible, ultraviolet, infrared, and microwave radiation. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; ELECTROMAGNETIC RADIATION; METEOROLOGICAL SATELLITES; REFLECTION OF ELECTROMAGNETIC RADIATION; SCATTERING OF ELECTROMAGNETIC RADIATION.

Each spectral region provides meteorologists and other Earth system scientists with information about atmospheric ozone, water vapor, temperature, aerosols, clouds, precipitation, lightning, and many other parameters. Measuring atmospheric radiation allows the detection of sea and land temperature, snow and ice cover, and winds at the surface of the

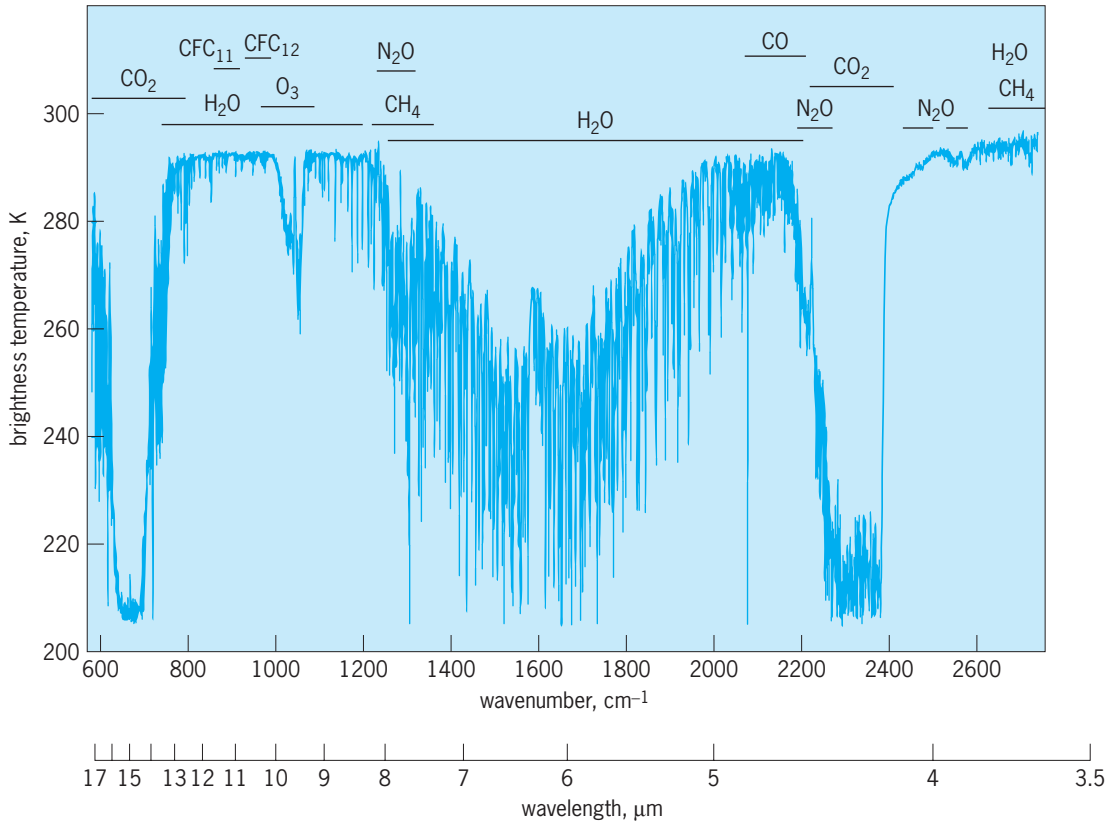


Fig. 1. Observed infrared spectrum illustrating the absorption gases and their spectral location. This spectrum was obtained from the scanning high-resolution interferometer sounder (S-HIS), which measured the emitted thermal radiation between 3.3 and 18 μm, onboard the NASA ER-2 aircraft over the Gulf of Mexico southeast of Louisiana on April 1, 2001. (*Data taken from K. N. Liou, 2002; originally provided by A. Huang and D. Tobin*)

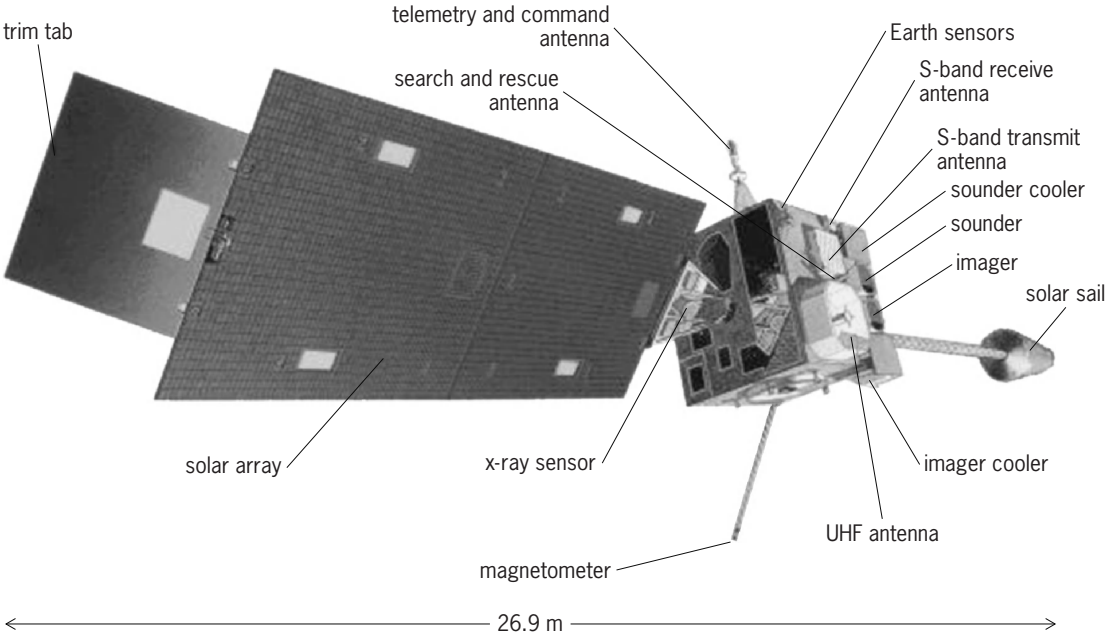


Fig. 2. Sketch of a *Geostationary Operational Environmental Satellite* (*GOES*) of the late 1990s.

ocean. By tracking the movement of clouds and other atmospheric features, such as aerosols and water vapor, it is possible to obtain estimates of winds above the surface. *See* SATELLITE METEOROLOGY.

Scientists use physically based algorithms to process the satellite-observed spectral radiation into either images of atmospheric and surface features or desired estimates of weather parameters. For example, the physical relationship between the absorption and reemission of infrared radiation is used to estimate atmospheric temperature profiles over vast reaches of ocean areas. *See* REMOTE SENSING.

Orbits of meteorological satellites are generally selected as one of two types. Geostationary (geosynchronous) orbits are in the equatorial plane and have the advantage of obtaining continuous, time-lapse observations of a portion of Earth as large as the Americas (**Fig. 2**). Other, near-polar orbits are Sun-synchronous and are designed to measure all areas of Earth at nearly the same local times each day.

Since meteorological satellites were first flown by NASA in the 1960s, a system of research and operational weather forecast-related satellites has been developed in the United States by the National Oceanic and Atmospheric Administration, and by the weather services of Japan, India, Russia, and the European Union. Today, worldwide weather satellite coverage is providing input to computer forecast models and to National Weather Service and television weather forecasters on an hourly basis. *See* WEATHER FORECASTING AND PREDICTION.

The scientific discoveries in weather and climate credited to observations from satellites include (1) measurement of the energy output from the Sun and the radiation budget of the Earth—warmer and darker than was believed before satellite observations; (2) the organization of tropical weather systems and waves; (3) the existence of mesoscale convective clusters—major precipitation-producing weather systems over midlatitude continents; (4) development and growth of severe thunderstorms on "gust fronts" outflowing from other storms; and (5) variation of the warm core inner structure of tropical storms and associated spiral bands.

Thomas H. Vonder Haar

Bibliography. S. Q. Kidder and T. H. Vonder Haar, *Satellite Meteorology: An Introduction*, Academic Press, 1995; K. N. Liou, *An Introduction to Atmospheric Radiation*, 2d ed., Academic Press, 2002.

Territoriality

Behavior patterns in which an animal actively defends a space or some other resource. One major advantage of territoriality is that it gives the territory holder exclusive access to the defended resource, which is generally associated with feeding, breeding, or shelter from predators or climatic forces.

Feeding territories involve defense of an area where food is found, or defense of the food itself. Breeding, or reproductive territories, may involve defense of a breeding site where access to mates is increased, or defense of offspring, such as in a nest. Shelter territories usually involve a hole or other structure where an animal can rest or hide safely, a place where a cold-blooded animal basks in the sun to increase its body temperature, a refuge site from extreme temperatures or humidities, or a physical barrier to wind or water currents. Feeding and breeding territories can be mobile, such as when an animal defends a newly obtained food source or a temporarily receptive mate. Stationary territories often serve multiple functions and include access to food, a place to rear young, and a refuge site from predators and the elements.

Territoriality can be understood in terms of the benefits and costs accrued to territory holders.

Benefits include time saved by foraging in a known area, energy acquired through feeding on territorial resources, reduction in time spent on the lookout for predators, or increase in number of mates attracted and offspring raised. Costs usually involve time and energy expended in patrolling and defending the territorial site, and increased risk of being captured by a predator when engaged in territorial defense.

Because territories usually include resources that are in limited supply, active defense is often necessary. Such defense frequently involves a graded series of behaviors called displays that include threatening gestures such as vocalizations, spreading of wings or gill covers, lifting and presentation of claws, head bobbing, tail and body beating, and finally, direct attack. Direct confrontation can usually be avoided by advertising the location of a territory in a way that allows potential intruders to recognize the boundaries and avoid interactions with the defender. Such advertising may involve odors that are spread with metabolic by-products, such as urine or feces in dogs, cats, or beavers, or produced specifically as territory markers, as in ants. Longer-lasting territorial marks can involve visual signals such as scrapes and rubs, as in deer and bear. *See* CHEMICAL ECOLOGY.

Individual size of combatants often determines the outcome of behavioral interactions for an unused territory. Once such a territory is established, however, a territory holder frequently can defend and repel individual intruders, even when the intruder is larger. This advantage can be overcome by invading as part of a group, with the many members of the group serving as "multiple lures." The defender can repel only a small number of intruders, while other members feed on the territorial resources.

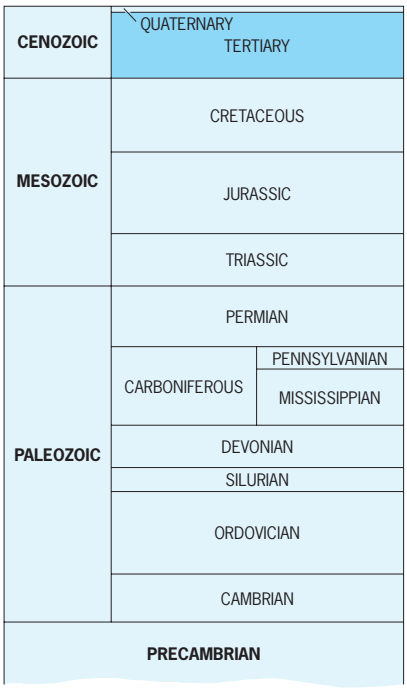
In reproductive territories, the advantage accrued to a successful territory holder can be measured by counting the number of matings that different territory holders obtain. Frequently the largest, strongest, or most experienced male will occupy the best territory and mate with the most females. Such territories may often contain a place for egg deposition, such as in many birds, frogs, and fishes. In other organisms, such as antelope, grouse, and bower birds, many males may compete for adjoining display sites that do not contain obvious resources. Here the central location in the display ground, or lek, is often the preferred one, and its occupant obtains more matings than other males. Females then go off somewhere else to bear young or lay eggs. *See* ETHOLOGY; POPULATION ECOLOGY; REPRODUCTIVE BEHAVIOR.

Gene S. Helfman

Bibliography. J. Alcock, *Animal Behavior: An Evolutionary Approach*, 7th ed., 2001; J. W. Grier and T. Burk, *Biology of Animal Behavior*, 2d ed., 1992; J. R. Krebs and N. B. Davies, *An Introduction to Behavioral Ecology*, 3d ed., 1993.

Tertiary

The older major subdivision (period) of the Cenozoic Era, extending from the Cretaceous (top of the Mesozoic Era) to the beginning of the Quaternary



(younger Cenozoic Period). The term Tertiary corresponds to all the rocks and fossils formed during this period. Although the International Commission on Stratigraphy uses the terms Paleogene and Neogene (pre-Quaternary part) instead, Tertiary is still widely used in the geologic literature. Typical sedimentary rocks include widespread limestones, sandstones, mudstones, marls, and conglomerates deposited in both marine and terrestrial environments; igneous rocks include extrusive and intrusive volcanics as well as rocks formed deep in the Earth's crust (plutonic). *See* CRETACEOUS; FOSSIL; ROCK.

The Tertiary Period is characterized by a rapid expansion and diversification of marine and terrestrial life. In the marine realm, a major radiation of oceanic microplankton occurred following the terminal Cretaceous extinction events. This had its counterpart on land in the rapid diversification of multituberculates, marsupials, and insectivores—holdovers from the Mesozoic—and primates, rodents, and carnivores, among others, in the ecologic space vacated by the demise of the dinosaurs and other terrestrial forms. Shrubs and grasses and other flowering plants diversified in the middle Tertiary, as did marine mammals such as cetaceans (whales), which returned to the sea in the Eocene Epoch. The pinnipeds (walruses, sea lions, and seals) are derived from land carnivores, or fissipeds, and originated in the Neogene temperate waters of the North Atlantic and North Pacific. Indeed, the great diversification on land and in the sea of birds and, particularly, mammals has led to the informal designation of the Tertiary as the Age of Mammals (**Fig. 1**) in textbooks on historical geology.

Geography. The modern configuration of continents and oceans developed during the Cenozoic Era as a result of the continuing process known as plate tectonics. Mountain-building events (orogenies) and uplifts of large segments of the Earth's crust

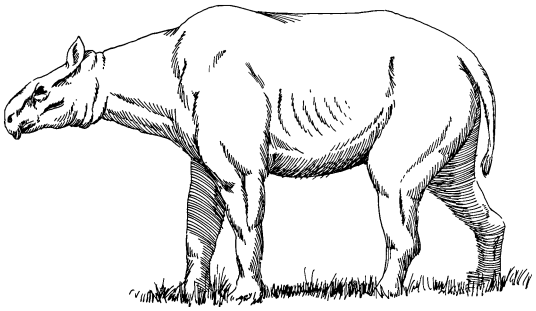


Fig. 1. ***Baluchitherium*****, the largest land mammal known, from the Tertiary (Oligocene Epoch) of Asia. It was 18 ft (5.4 m) high at the shoulders. (*After R. A. Stirton*, *Time*, *Life and Man*, *Wiley*, *1959*)**

(epeirogenies) alternated with fluctuating transgressions and regressions of the seas over land. This resulted in a complex alternation of marine and terrestrial sediments and their contained records of the passage of life (fossils). Some modern inland seas (for example, Lake Baikal and the Caspian Sea) are remnants of once more extensive widespread epeiric (shallow) seaways of the early Tertiary.

The middle to late Tertiary Alpine-Himalayan orogeny and the late Tertiary Cascadian orogeny led to the east-west and north-south mountain ranges, respectively, which are located in Eurasia and western North America. *See* CORDILLERAN BELT; MOUNTAIN SYSTEMS; OROGENY; PLATE TECTONICS.

Rocks. Tertiary sedimentary rocks occur as a relatively thin veneer of marine rocks on the margins of continents around the world. In the petroliferous province of the Gulf of Mexico, Tertiary rocks attain thicknesses in excess of 30,000–40,000 (9000–12,000 m); whereas in the more tectonically active borderlands around the Pacific Ocean, such as the Santa Barbara–Ventura Basin of California, and the flanks of the uplifted Himalayan-Alpine chain of Eurasia, thicknesses in excess of 50,000 ft (15,000 m) have been recorded. Terrestrial (nonmarine) strata are generally thinner, are more patchy in distribution, and occur predominantly in the internal basins of the continents (for example, the Basin and Range Province of North America and the Tarim Depression of Asia). Major Tertiary volcanic provinces include those of the Deccan region of India, the basaltic plateaus of Greenland and Iceland, and the Columbia Plateau of the northwest United States.

Stratigraphy and history. Although the ancient Greeks recognized the shells of mollusks far inland from the Aegean Sea as fossil marine organisms, as did Leonardo da Vinci some 2000 years later, it was not until the era of enlightenment in the eighteenth century that the first attempt was made to place the Earth's rock record into a historical context. The term Tertiary is derived from Giovanni Arduino, who in 1759 formulated a threefold subdivision of the Earth's rock record in Primary, Secondary, and Tertiary. While the first two terms have long since disappeared from geologic hagiography, the term Tertiary persists in modern scientific literature. In its more modern sense the term Tertiary is defined by its usage in 1810 by the French Scientists Alexandre

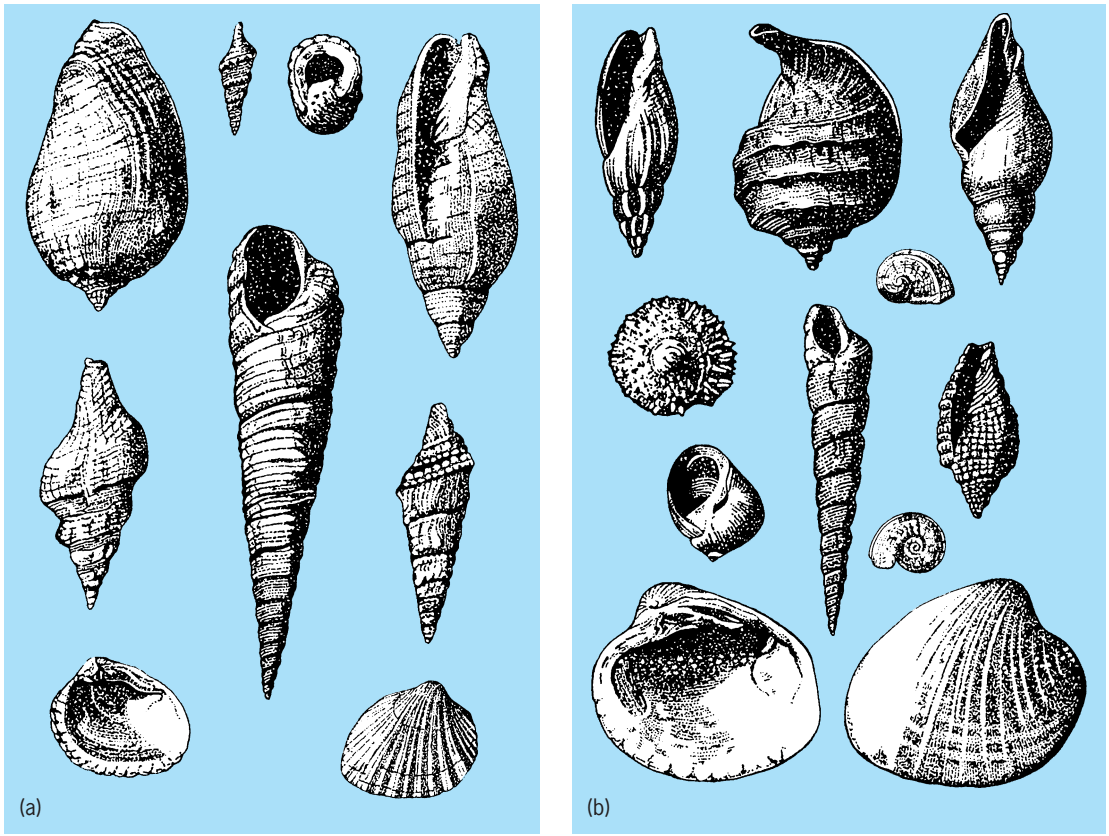


Fig. 2. Mollusk fossils used by Lyell to zone the Tertiary. (*a*) Miocene. (*b*) Eocene. (*After C. Lyell*, *Principles of Geology*, *1833*)

Brogniart and Georges Cuvier for all the rock formations in the Paris Basin that lay above the Cretaceous chalk sequence. Although many subdivisions of the Tertiary exist that developed in the succeeding two centuries, only five major time-rock units generally are recognized. In 1833, Charles Lyell made the first systematic hierarchical subdivision of the Tertiary Period based upon the observations of his Parisian colleague M. Deshayes and other contemporary European conchologists that the percentages of living species in the fossil record increased as the Tertiary stratigraphic record ascended (**Fig. 2**). Lyell's Tertiary subdivisions include, in ascending order, the Eocene, Miocene, Older Pliocene, and Newer Pliocene. The last term was subsequently (1839) changed to Pleistocene. Heinrich Ernst von Beyrich later defined the term Oligocene for rocks exposed in the North German Basin and the Rhine Basin that had been previously allocated to a part of either the Eocene or Miocene by Lyell. The paleobotanist W. P. Schimper added the term Paleocene in 1874 based on the oldest Tertiary terrestrial strata exposed in the east Paris Basin. *See* CENOZOIC; EOCENE; HOLOCENE; MIOCENE; OLIGOCENE; PALEOBOTANY; PALEOCENE; PALEONTOLOGY; PLEISTOCENE; PLIOCENE; STRATIGRAPHY. W. A. Berggren

Bibliography. R. H. Dott, Jr., and D. R. Prothero, *Evolution of the Earth*, 7th ed., 2003; B. M. Funnell and W. R. Riedel, *The Micropaleontology of Oceans*, 1971; S. J. Gould, *Time's Arrow, Time's Cycle*, 1987; H. L. Levin, *The Earth Through Time*, 7th ed., 2003.

Testis

The organ of sperm production. In addition, the testis (testicle) is an organ of endocrine secretion in which male hormones (androgens) are elaborated. In the higher vertebrates (reptiles, birds, and mammals), the testes are paired and either ovoid or elongated in shape. In mammals, the testes are usually ovoid or round. In many species (for example, humans) they are suspended in a pouch (scrotum) outside the main body cavity; in other species they are found in such a pouch only at the reproductive season; in still others the testicles are permanently located in the abdomen (for example, in whales and bats).

Histology

Within a firm and thick capsule of connective tissue, the tunica albuginea, the testis contains a varying number of thin (0.005–0.017 in. or 0.12–0.45 mm, depending upon the species) but very long seminiferous tubules which are the sites of sperm formation. Essentially, these tubules are simple loops which open with both their limbs into a network of fine, slitlike canals, the rete testis. From this the sperm drains through a few narrow ducts, the ductuli efferentes, into the epididymis, where sperm mature and are stored.

The seminiferous tubules constitute most of the testis, and in different species vary greatly in complexity. Often they are extremely coiled and winding. Sometimes they branch and interconnect as in man. Their total length in man is about 700 ft (210 m), in the bull about 3 mi (4.8 km). Each tubule is surrounded by a layer of thin cells (**Fig. 1**) which is contractile and enables the tubules to wriggle slowly. The spaces between tubules are filled with connective tissue, blood vessels, an extensive network of very thin-walled lymph vessels, and secretory cells, the interstitial cells or cells of Leydig, which secrete male hormone. The interstitial cells vary in different species and different periods in life. In the human fetus they are extremely abundant and functional; later they regress and change into inconspicuous connective tissue cells. At puberty they reappear and enter the main productive phase of their life.

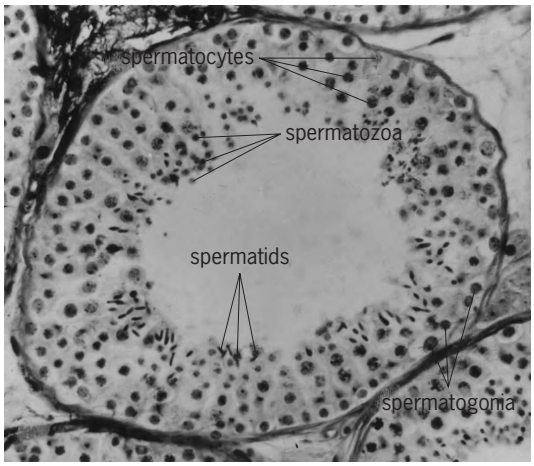


Fig. 1. Human seminiferous tubules in cross section. A multitude of stages of germ cells is present.

Sperm Formation

The sperm cells, spermatozoa, develop in the wall of the seminiferous tubules, either periodically, as in most vertebrates, or continually, as in humans. Most of the cells in the tubules are potential spermatozoa (spermatogenic or germ cells). Nursing cells (Sertoli cells) are interspersed at regular intervals between them (Fig. 1). The Sertoli cells support and surround the developing spermatogenic cells and provide a specialized environment, which is absolutely necessary for normal sperm development. *See* SPERM CELL.

Four cell types representing characteristic stages in the development of sperm cells have been given specific names. The youngest ones are called spermatogonia or, at their earliest stages, stem cells. They lie at the periphery of the tubule and show no resemblance to mature spermatozoa. Once a reproductive phase has begun, the spermatogonia undergo periodic divisions. A few of them stop dividing after one or two mitoses and revert to stem cells. These are the progenitors of future generations of spermatozoa. Most spermatogonia, however, go through additional divisions and transform into another cell type, the spermatocyte, in which the nucleus is slowly preparing for the maturation divisions (meiosis). The small cells resulting from these divisions are called

spermatids. They are haploid; that is, their chromosome number is reduced by one-half. Usually four spermatids arise from one spermatocyte. The subsequent transformation of the spermatid into a spermatozoon is called spermiogenesis. This is an extremely complex process which involves the development of a distinctive head structure containing the highly condensed nucleus, head cap structures such as the acrosome that help the sperm to attach and enter the egg, and a tail which allows the sperm to move in a swimming motion. During condensation of the sperm nucleus, there is considerable reorganization of the chromosomes, with the appearance of many unique nuclear proteins, such as the testis-specific histone Hlt, and the gradual replacement of histones by another group of proteins that are unique to the sperm nucleus, the protamines. These specific nuclear proteins facilitate very dense packing of the nuclear material in the sperm head, although at the expense of the genes ability to remain active. During spermiogenesis, groups of spermatids become attached to each nursing cell and often indent its membrane deeply. In the final stage the spermatozoa retract toward the lumen and are released; they are still hardly motile and usually they are not fertile until they have spent some time within the epididymis.

The duration of spermatogenesis is known with fair accuracy for several species. Usually it is about 5–7 weeks, in human males 75 days, from the first spermatogonial division to the release of the spermatozoa from the testis. About one-fourth of this time is spent in the spermatogonial stage of cell multiplication, a little more than one-third in the long preparation for, and the rapid completion of, the maturation divisions, and about one-third in spermatogenesis. Spermatogenesis in the testis is the result of a balance between proliferation and differentiation, and cell degeneration or apoptosis. Apoptosis of the spermatogenic cells is largely hormonally controlled, and specifically directed apoptosis occurs in conditions of testicular damage due to environmental insults such as heat, radiation, or chemical toxicants. Recovery of spermatogenesis is possible provided the stem cells are not depleted by these processes.

In nearly all mammals, spermatogenesis occurs in rigid patterns which are essentially similar. Large groups of cells filling sections of tubules up to several centimeters in length develop synchronously. Generations follow each other at definite intervals so that in any section of a tubule, characteristic associations of cells are present. In addition, along the tubule, groups of cells which are in synchronous development border on other groups in immediately preceding or following stages. This constitutes the spermatogenic wave. In humans and apes, the spermatogenic wave is less well defined, with a spiral arrangement of several short asynchronous waves of development running along the tubule. As a result, each cross section of a tubule reveals a multitude of stages of germ cells coexisting in random array (Fig. 1).

In lower vertebrates, many patterns and arrangements of spermatogenesis are found, sometimes in gonadal sacs which open directly into the abdominal cavity from which they are released through abdominal pores (cyclostomes), sometimes in tubules or compartments which have only temporary connections with outleading ducts (selachians). In all lower vertebrates a multitude of spermatozoa develops in the presence of definite nursing cells. In most cases, large groups of cells, often whole compartments (in insects and amphibians), develop synchronously. *See* SPERMATOGENESIS.

Mark P. Hedger; Edward C. Roosen-Runge

Physiology

The functions of the testis are dependent on the secretion of gonadotropic hormones, the release of which from the pituitary gland is regulated by the central nervous system.

In some lower forms, for example, many crustaceans, the hormone-producing "androgenic gland" is anatomically distinct from the gametogenic organ or testis proper. In mammals, male-hormone production resides in the Leydig cells, located in the intertubular tissue of the testes. Destruction of the seminiferous epithelium by x-irradiation, heat, or cadmium poisoning does not abolish the endocrine function of the testis.

Formation and release of testosterone. The principal androgenic hormone released by the testis into the bloodstream is testosterone. The testis is able to utilize simple molecules, such as acetic acid, to form cholesterol and to convert this sterol to Δ^5-pregnenolone. The major pathways leading to the formation of testosterone from pregnenolone are shown in **Fig. 2**. One pathway involves progesterone and Δ^4-androstene-3,17-dione as intermediates; androstenedione often accompanies testosterone in the venous blood of the testis. Alternatively, removal of the side chain at carbon-17 may precede oxidation of the hydroxyl group at carbon-3, so that additional Δ^5-3β-hydroxysteroid intermediates, such as dehydroepiandrosterone and Δ^5-androstenediol, are formed; the latter is itself a fairly potent androgen. In humans, production of testosterone occurs mainly through the Δ^5 pathway (pregnenolone to androstenediol), with the Δ^4 pathway through progesterone contributing relatively little to androgen production. This appears to be due to the relatively poor affinity of the human steroidogenic enzyme for the Δ^4 intermediates. In some species, however, the Δ^4 pathway is more important. Both testosterone and androstenedione may be further metabolized into estrogens in the testis. This conversion takes place in the tubular tissue, although interstitial cells also contribute. The production of estrogens in the male varies quite widely among species, from relatively low in humans to very high, for example in stallions and boars. It is now apparent, however, that estrogens are important in the development and proper function of the ducts which drain the testis (the rete testis and ductuli efferentes), even in species with relatively low levels of estrogens. The testis of the stallion is one of the richest natural sources of estrogen. Administration of the luteinizing hormone (LH) of the pituitary to human males leads to increased urinary excretion of estrogens.

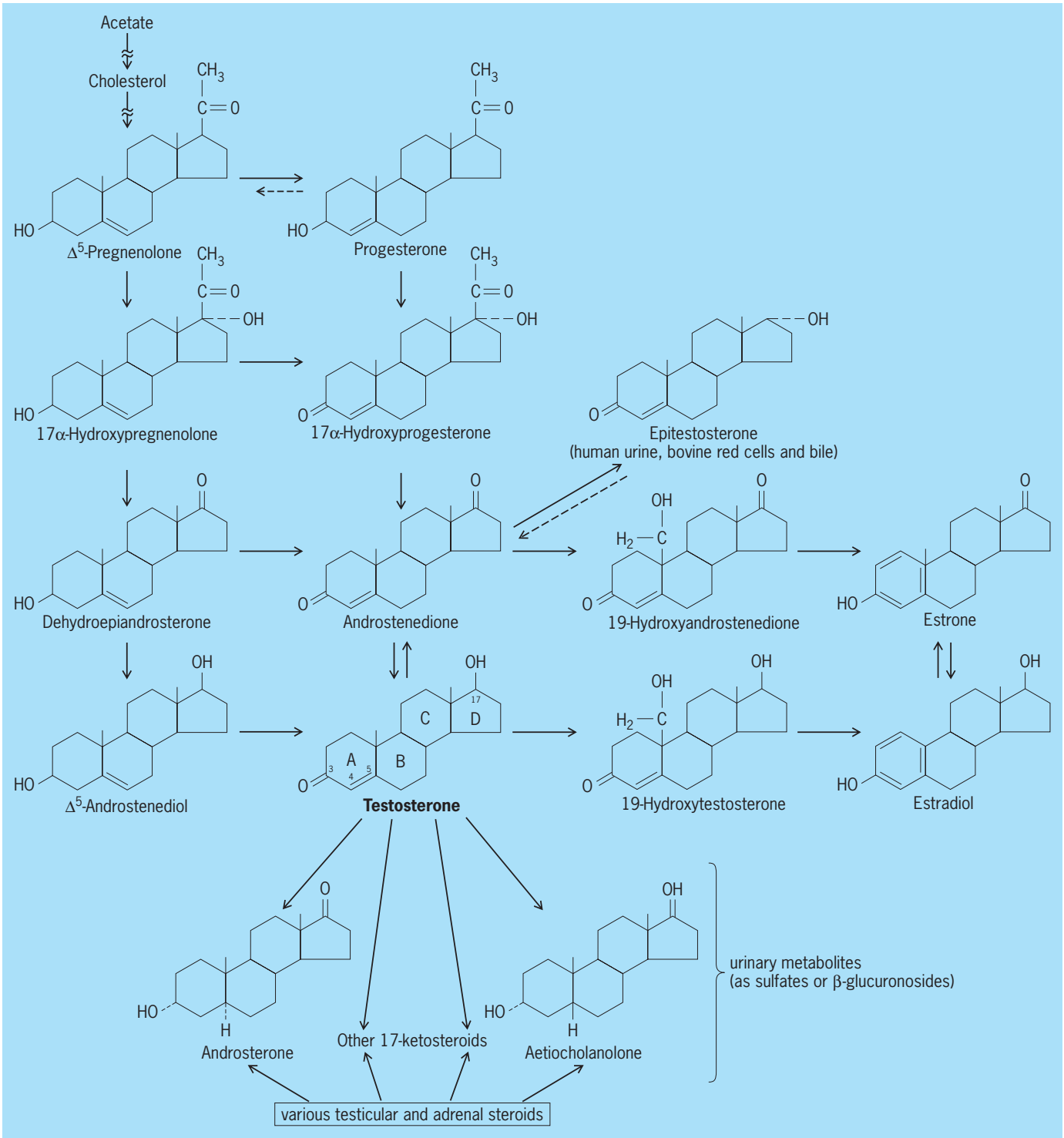


Fig. 2. Major pathways for the biosynthesis of testosterone in the testis and subsequent metabolism of the hormone.

Testosterone synthesis is normally limited by the rate of pituitary gonadotrophin secretion: administration of the luteinizing hormone or of chorionic gonadotrophin results in increased testosterone synthesis and release within minutes. These hormones also stimulate growth and multiplication of Leydig cells. Hypophysectomy leads to cessation of androgen formation.

The bulk of the hormone produced enters the spermatic venous blood, where it becomes associated with albumin and with a specific testosterone-binding globulin on its way to remote target organs. Some of the hormone, however, is released directly into the testicular interstitial fluid and lymph and diffuses across the basement membrane into the tubules; in the tubular fluid it is carried to the epididymis. This local diffusion of androgen assists sperm maturation and storage in this organ.

Metabolism and excretion of testosterone. Testosterone is rapidly inactivated in the liver; some metabolism also occurs in target organs, in red blood cells, and in the kidneys. In man, the final products

are chiefly 17-ketosteroids with a saturated A-ring, such as androsterone and aetiocholanolone (Fig. 2); these are excreted as sulfates or glucuronosides in the urine. Similar products are formed from adrenal steroids and from testicular steroids other than testosterone, so that urinary 17-ketosteroid excretion does not accurately reflect male hormone secretion. In the bull, 90% of the testosterone metabolites, such as epitestosterone (Fig. 2), are excreted via the bile into the feces. In some animal types, both urinary and biliary excretion occur, with partial resorption of steroid metabolites from the gut (enterohepatic circulation).

Development of function and puberty. The enzymes required for the synthesis of testosterone first appear in the testis of the early fetus, but the Leydig cells regress before or shortly after birth and resume their development only toward the onset of puberty. Androgens are present in the blood in small amounts even in infancy, but the testicular secretion increases markedly and changes its character shortly before puberty: The secretion of the prepubertal testis of the bull, rat, guinea pig, and possibly of boys differs from that of the adult in that androstenedione rather than testosterone is the predominant steroid formed; androstenedione is a very feeble androgen but, like testosterone, it stimulates body growth. The testis of the immature rat is further distinguished by the presence of reducing enzymes which convert androstenedione and testosterone to compounds of greater hormonal activity; these enzymes disappear from the testis with the onset of puberty.

The brain plays an important part in initiating puberty. The hypothalamic centers that stimulate pituitary gonadotrophin release are subject to negative feedback exerted by the steroid hormones present in the blood. In the immature animal, this feedback inhibition becomes operative at much lower levels of circulating testicular hormones than in the adult. Thus an upward resetting of the hypothalamic control center, or "gonadostat," appears to be a crucial event in the initiation of puberty. Tumors involving the posterior hypothalamus or the pineal gland are often associated with precocious puberty. Precocious development of genitalia and secondary sex characters, unaccompanied by spermatogenesis, may be due to Leydig cell or adrenal tumors. Much variation in the onset of puberty, however, is normal.

In most mammals, including humans, testicular development is dependent upon the presence of a Y chromosome. The testis-determining region of the Y chromosome has been identified, and is called SRY (for sex-determining region on the Y chromosome). The product of this gene is a protein, which binds to and activates several other genes that are necessary for testicular development. Absence of the *SRY* gene usually excludes the appearance of a testis or development of male characteristics. However, other genes that also play a pivotal role in gonadal development are still being discovered, including *DAX*, which induces ovarian development. This means that creation of the testis involves the sequential activation or inactivation of a number of genes within the developing gonad.

Biological actions of testosterone and development of the testis. At the ambisexual stage of embryonic development, the testis promotes the growth of the paired Wolffian ducts and their differentiation into the epididymis, vasa deferentia, and seminal vesicles; the fetal testis also causes masculinization of the urogenital sinus, fusion of the labioscrotal folds in the midline, and development of the genital tubercle into a phallus. All these actions can be mimicked by androgenic steroids in the organism or in organ cultures, and prevented by administration of antiandrogenic drugs. The Müllerian ducts are suppressed by the fetal testes through the release of a substance, Müllerian-inhibiting hormone or anti-Müllerian hormone (MIH or AMH), which spreads by local diffusion. This hormone is a single-chain protein, produced by the fetal Sertoli cells. Recent research also indicates that MIH may be involved in testicular descent and spermatogenic development in early postnatal life. Castration of male fetuses in mammals leads to persistence of these ducts, which then form a uterus, and to the development of external genitalia of female type. Similar abnormal development is observed in man in cases of gonadal dysgenesis due to a genetic defect, such as the absence of the Y chromosome (45/XO constitution) in Turner's syndrome. In some lower forms, however, the Y chromosome is not essential for the expression of male characters (*Drosophila*) or even for male fertility (mouse). Humans with more than one X chromosome (for example, 47/XXY) have male external genitalia, but the penis may be imperfectly formed (hypospadic) and the testes are small and produce no sperm (Klinefelter syndrome); Leydig cells are abundant, but may not produce normal amounts of androgen. Many forms of male infertility have been recognized to have a genetic cause due to more precise defects of genes involved in testicular function. Deletions of several specific regions of the Y chromosome are associated with spermatogenic failure, and disorders of the androgen receptor gene (located on the X chromosome in humans) can lead to problems ranging from lack of testicular development to arrested spermatogenesis. These disorders can be accompanied by other complications, such as neural degeneration.

In the newborn rat, the hormonal secretion of the testis has been shown to modify permanently the function of brain centers that govern the pattern of gonadotrophin release by the pituitary gland in adult life. As a result, a steady tonic secretion of luteinizing hormone is established, instead of the cyclic surges characteristic of the mature female. The male hormone also acts on other brain centers to impose male-type mating behavior and aggressiveness. However, in man early rearing and environmental factors are important in shaping sexual behavior patterns.

Testosterone promotes the retention of dietary nitrogen and its incorporation into muscle protein. This effect is responsible for the greater muscular

development of the male and contributes to the prepubertal growth spurt of boys. Therapeutically, the anabolic action of the hormone is utilized for treating patients suffering from wasting diseases, with testosterone derivatives of lesser virilizing activity used. Low doses of testosterone enhance linear growth, but higher doses accelerate epiphyseal closure in long bones. Toward puberty, increased secretion of testosterone stimulates the growth of the penis, scrotum, and male accessory glands responsible for the formation of the seminal plasma, for example, the prostate and seminal vesicles. The hormone brings about the appearance of secondary sex characters, such as the male-type distribution of hair and body fat and lowered pitch of voice in man, the growth of the comb and wattles in birds, the clasping pads of amphibians, or the dorsal spine of certain fishes. Substances that produce such effects are referred to as androgens. The thermoregulatory function of the scrotum is testosterone-dependent. Testosterone stimulates the synthesis of ribonucleic acid and protein, and of specific seminal constituents such as fructose and citric acid, in the male accessory glands. Castration leads to atrophy of these glands. Some forms of prostatic cancer regress when deprived of androgen.

Testosterone is essential for efficient spermatogenesis, but high doses of testosterone suppress spermatogenesis by inhibiting pituitary gonadotropin secretion. Protein hormones from the testis also regulate gonadotropin secretion. Follicle-stimulating hormone (FSH), which is important for establishment of the spermatogenic process at puberty and regulates Sertoli cell activity throughout life, is controlled by a protein hormone called inhibin, produced by the Sertoli cell.

Unlike the ovary, the testis remains functional throughout life, with ongoing spermatogenic development. However, the efficiency of spermatogenesis falls away, and androgen levels begin to fall due to a declining Leydig cell activity. These events can lead to reduced fertility, and androgen insufficiency problems in later life in some men, which may require medical treatment.

Control of spermatogenesis. Hypophysectomy abolishes spermatogenesis. Full restoration of spermatogenesis usually requires the replacement of both gonadotropins (LH and FSH), although large doses of testosterone and other steroids, for example, dehydroepiandrosterone and pregnenolone, temporarily maintain spermatogenesis in hypophysectomized rats. Only FSH of human or primate source is effective in humans. In addition to hormonal control, it is now recognized that many protein growth factors, or cytokines, are involved in integrating the control of testicular function. Cytokines such as interleukin-1, which is responsible for many of the manifestations of inflammatory diseases, is produced by the Sertoli cell and is involved in regulation of spermatogonial development. Other cytokines are involved in testicular development, and integration of the function of the separate cells of the testis. Disturbances in cytokine production during inflammation and disease often leads to disruption of testicular function and fertility.

Sympathetic nerves supply vasomotor fibers to the testis and motor fibers to the smooth muscles of the epididymis. Administration of the catecholamine hormones of the adrenal medulla or of serotonin causes intense vasoconstriction of the testicular vessels and damage to spermatogenesis.

Sperm transport from the testis. A fluid forms in the seminiferous tubules, about 2.4 $in.^3$ (40 ml) per testis per day in the ram, which is continually transferred through the rete testis to the epididymis, carrying in it a dilute suspension of immature spermatozoa. This fluid is produced by the Sertoli cells. Almost 99% of the testicular fluid is resorbed in the head of the epididymis. As it leaves the rete, this medium is practically devoid of glucose, the chief energy source of the testis, but contains some lactate, inositol, and strikingly high concentrations of glutamic acid. The role of these substances is not well understood. Glutamine seems to favor the growth and differentiation of testes in tissue and organ culture, and polyglutamic acid, present in high concentration in the reproductive tract of the cock, has a motility-inhibiting, yet life-extending, action on sperm.

Sperm transport is probably assisted by ciliary movement in the efferent ducts and by contractile cells surrounding the tubules. The posterior hypophyseal hormone oxytocin induces contractions of the tubules and epididymis.

During sexual inactivity spermatozoa are voided continually in the urine (marsupials, ram) or by spontaneous seminal discharges, for example, in rats subjected to mechanical restraints.

Blood–testis barrier. The tubular basement membrane and the adjacent layer of Sertoli cells present a barrier to the penetration of certain substances which readily pass across most capillary membranes, for example, inulin, *p*-aminohyppuric acid, and the larger plasma proteins; others, such as Rb^+, are admitted only slowly. This blood-testis barrier is important for maintaining the specialized environment

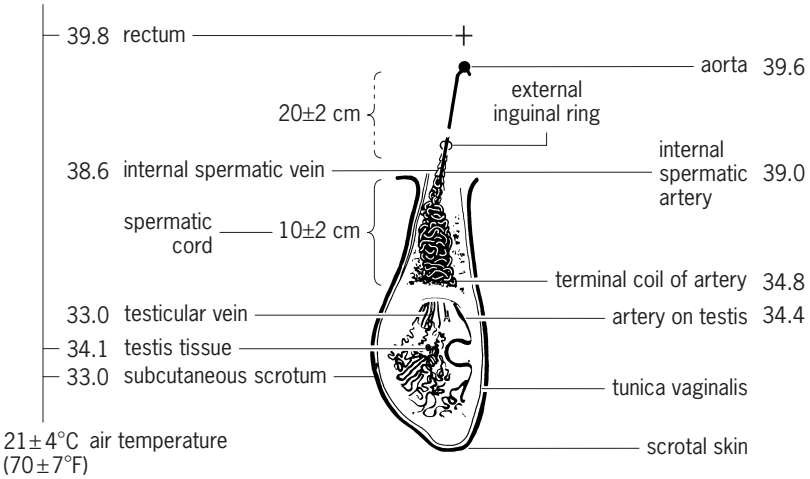


Fig. 3. Relation of testicular temperature to scrotal and deep-body (rectal) temperature, and temperature gradients along the testicular and spermatic vessels in the ram. (*After G. M. H. Waites and G. R. Moule, J. Reprod. Fert., 2:213, 1961*)

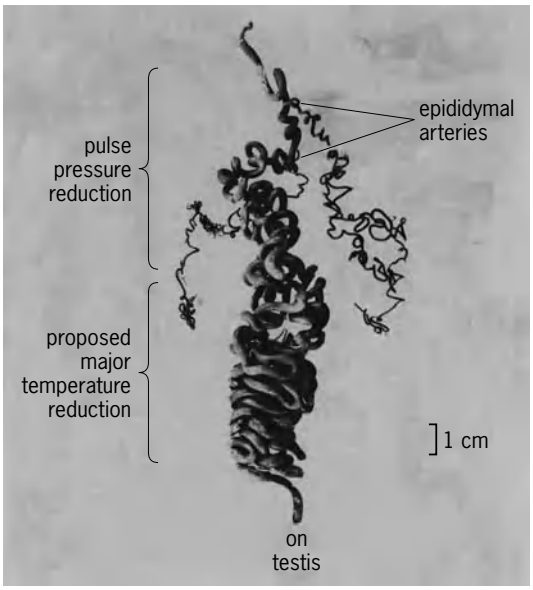


Fig. 4. Latex cast of internal spermatic artery supplying right testis of a ram. Course of the vessel from the external inguinal ring to testis is shown with the arterial coils in their natural disposition. (*From G. M. H. Waites and G. R. Moule, J. Reprod. Fert., 1:223, 1960*)

necessary for spermatogenesis, and also to sequester the developing germ cells from the immune system. Being normally sequestered from the circulation, the spermatogenic cells are not recognized as "self" by the antibody-forming tissues. Thus, when testicular tissue is injected into an animal or when the testis is damaged by inflammation, autoimmune reactions may be induced against the seminiferous tissue and cause its destruction.

Temperature control of testis. In most mammals the testes descend, either permanently or for the duration of the breeding season, into an extraabdominal pouch, the scrotum, where its temperature is maintained 5–13°F (3–7°C) below the animal's deep-body temperature. If this temperature gradient is eliminated by failure of the testis to descend (cryptorchidism) or by severe environmental heat stress, spermatogenesis is impaired and eventually abolished. The reasons for this are not understood. Glucose and oxygen supply to the tubules may become inadequate during scrotal heating. The pachytene stage of the prolonged meiotic prophase of the spermatocytes seems to be particularly sensitive to elevated temperature, but dividing spermatocytes and early stages of spermiogenesis are also adversely affected; spermatogonia are somewhat more resistant. The requirement for low testicular temperature is not universal: Birds, whose body temperature is generally higher than that of mammals, have intraabdominal testes. The testes remain in the abdomen in Monotremata (such as the duckbilled platypus), some Insectivora (such as Madagascar "hedgehogs"), Edentata (such as armadillos), Cetacea (such as whales), Proboscidae (African elephants), Hyracoidea (such as rock hyrax), and Sirenia (such as sea cows).

Testicular descent is accelerated by the luteinizing hormone and testosterone, but may be mechanically obstructed in the presence of teratomatous growth or tumor.

The scrotal skin is thin, devoid of subcutaneous fat, and richly supplied with small blood vessels. It contains heat-sensitive nerve endings of the superficial perineal nerves whose firing rate increases as scrotal temperature rises (above 95°F or 35°C in the ram). Heat loss is effected by increasing the scrotal surface through relaxation of the tunica dartos muscle; by increased scrotal blood flow, including the opening up of arteriovenous shunts; by more frequent synchronous discharges of the large scrotal apocrine sweat glands, triggered by efferent adrenergic fibers in the external spermatic nerves; and by reflex thermal polypnea (panting). The scrotal cooling effect is transmitted to the deeper parts of the gland by the precooling of the arterial inflow through countercurrent heat exchange with the returning venous blood (**Fig. 3**): In most mammals the spermatic artery is extremely long and coiled (**Fig. 4**), thin-walled and closely applied to the veins of the pampiniform plexus. This arrangement also serves to attenuate the arterial pulse (**Fig. 5**) and to utilize its energy to aid venous return. In marsupials, the spermatic artery branches into a leash of more than 100 vessels surrounded by parallel veins. In species with intra-abdominal testes, the spermatic vessels are

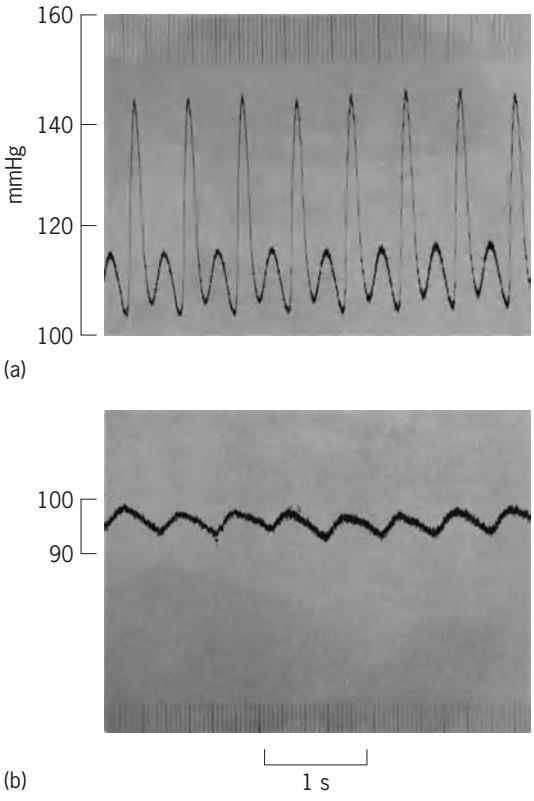


Fig. 5. Record of pressure changes in internal spermatic artery of a dog. (*a*) Above vascular cone. (*b*) On testis. Note attenuation of arterial pulse during its passage through coiled part of artery. 1 mmHg = 133 pascals. (*From G. M. H. Waites and B. P. Setchell, The Gonads, Appleton-Century-Crofts, 1969*)

generally short and straight. *See* REPRODUCTIVE SYSTEM. Mark P. Hedger; Hans R. Lindner

Bibliography. D. M. de Kretser (ed.), *Cellular and Molecular Mechanisms in Male Reproduction*, 1993; C. Desjardins and L. L. Ewing (eds.), *Cell and Molecular Biology of the Testis*, 1993; A. Gorbman et al., *Comparative Endocrinology*, 2d ed., 1987; E. Knobil et al. (eds.), *The Physiology of Reproduction*, 1988; K. McElreavey (ed.), *The Genetic Basis of Male Infertility*, 2000; J. D. Wilson, *Williams Textbook of Endocrinology*, 9th ed., 1998.

Tetanus

An infectious disease, also known as lockjaw, caused by the exotoxin of *Clostridium tetani*. The organism may be isolated from fertile soil and the intestinal tract or fecal material of humans and other animals. Infection commonly follows dirt contamination of deep wounds or other injured tissues such as war wounds or crash injuries, burns, or septic abortion in which tissue necrosis is present. Any deep puncture wound that is contaminated by soil or feces is suspect. Tetanus neonatorum is frequent in developing countries if the cut surface of the umbilical cord is not properly treated. Use of understerilized syringe needles by drug users is also recognized as a cause of tetanus. *See* TOXIN.

The causal organism is a strictly anaerobic, slender bacillus with a spherical terminal spore swelling the vegetative cell. Identification of the organism, however, should not rest on microscopic evidence but should be confirmed by neutralization tests with known tetanus antitoxin of the toxin produced by the pure culture. This species produces two toxic substances: a hemolysin known as tetanolysin, which is probably not involved in the disease, and the potent lethal toxin known as tetanospasmin, which has a strong affinity for the cells of the central nervous system. The neurotoxin has been isolated in crystalline form. This organism does not ferment carbohydrates but depends on the fermentation of amino acids for energy. *See* TOXIN-ANTITOXIN REACTION.

The diagnosis of tetanus is on clinical symptoms, and the incubation period is usually 5–10 days. The disease is characterized by convulsive tonic contraction of voluntary muscles. Therapy of suspected or diagnosed cases includes antibiotics (penicillin, tetracyclines) combined with injections of human antitoxin; equine antitoxin is to be avoided since the individual may be hypersensitive to horse serum. Prevention of tetanus rests on the proper, prompt surgical care of contaminated wounds and prophylactic use of antitoxin if the individual has not been protected by active immunization with toxoid. Babies aged 1–3 months should be actively immunized with DPT (diphtheria, pertussis, tetanus) toxoid; this immunization should be repeated at 10-year intervals. *See* IMMUNOLOGY. Leland S. McClung

Bibliography. S. M. Finegold, *Anaerobic Infections in Humans*, 1989; L. S. McClung, *The Anaerobic Bacteria: Their Activities in Nature and Disease*, 7 vols., 1982; G. L. Mandell, R. G. Douglas, and J. E. Bennett (eds.), *Principles and Practice of Infectious Diseases*, 5th ed., 2000; L. L. Simpson (ed.), *Anaerobic Bacteriology: Clinical and Laboratory Practice*, 3d ed., 1977; A. J. Zehnder (ed.), *Botulinum Necrotoxin and Tetanus Toxin*, 1989.

Tetraodontiformes

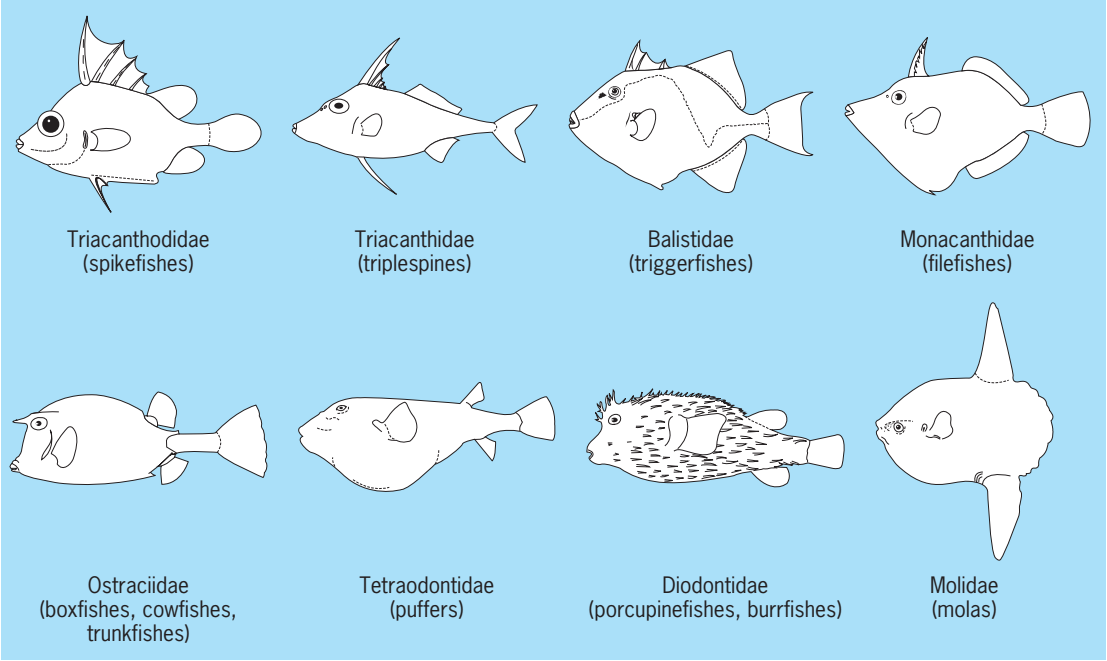
An order of the class Actinopterygii, also known as the Plectognathi. This most highly derived order of actinopterygian fishes is noted for its reduction in skeletal elements. Lost are the parietal, nasal, and infraorbital bones of the skull. Usually there are no pleural ribs; vertebrae are as few as 16 and no more than 30; and the maxillae are united with the premaxillae. The gill opening is a short vertical slit in front of the pectoral fin. Scales are usually modified to form spines, shields, or plates; a lateral line may be present or absent; and a swim bladder is present, except in the family Molidae. The 357 extant species are in three suborders, nine families (see **illustration**), and about 101 genera. *See* ACTINOPTERYGII; OSTEICHTHYES; SWIM BLADDER.

Suborder Triacanthodoidei. The single family, Triacanthodidae (spikefishes), is characterized by a deep and moderately compressed body, which is covered with a thick skin and tiny scales; each scale bears spicules, resulting in the skin having a shagreen-like appearance (that is, rough-looking skin covered with small close-set tubercles); the caudal fin is rounded to truncate; there are usually six dorsal fin spines and 12–18 dorsal fin rays; 11–16 anal rays; and a caudal fin with 12 principal rays. Each pelvic fin consists of one stout spine and one or two soft rays; the pelvic spines can be locked into place. There are 20 vertebrae. Some species have an extremely long snout; for example, in *Halinochirurgus akacki* the snout is about equal in length to the entire body posterior to the eye. Species of spikefishes range in total length from about 5.3 to 26 cm (2 to 10 in.). They are benthic in deep waters of the tropical and subtropical Atlantic and Indo-Pacific oceans.

Suborder Balistoidei. Balistoidei is also known as Sclerodermi. In the following four families, the frontal bones extend far anterior to the articulation between the lateral ethmoid and ethmoid bones.

Triacanthidae (triplespines). This family, formerly placed in the suborder Triacanthodoidei, is superficially similar to the Triacanthodidae but can be distinguished from them by their deeply forked caudal fin vs. round or truncate caudal fin; dorsal fin rays 19–26 vs. 12–18; anal fin rays 13–22 vs. 11–16; and maximum length about 28 cm (11 in.). The family comprises four genera and seven species, limited to shallow waters of the Indo-Pacific Ocean.

Balistidae (triggerfishes). This family is characterized by a deep and moderately compressed body covered with a very thick, tough skin bearing rectilinear scale plates. The scales above the pectoral fin base are usually enlarged and slightly separated, forming a



Representations of eight families (Triodontidae is not pictured) of the Tetraodontiformes. (*Courtesy of J. S. Nelson, Fishes of the World, 4th ed., Wiley, 2006*)

flexible tympanum. There is a small and usually terminal mouth, and a row of eight strong outer teeth in both the upper and lower jaws. There are three dorsal fin spines, the first of which can be locked in an upright position by the second; pelvic fins and spines are rudimentary or absent. There are 18 vertebrae. Eleven genera and about 40 species occur in the Atlantic, Indian, and Pacific oceans, either pelagic open water or benthic around rocky or coral reefs.

Monacanthidae (filefishes). Filefishes are characterized by a deep and compressed body covered with tiny scales, which are prickly or velvety to touch. There are usually two dorsal spines; the second is usually much smaller or absent. The upper jaw usually has three teeth in the outer series and two in the inner on each premaxillary, developed for nibbling. There are 19–31 vertebrae. Most species feed on a wide variety of benthic invertebrates, but some specialize on corals or zooplankton. Filefishes lay demersal (living at or near the bottom of the sea) eggs in a site prepared and guarded by the male or both parents. Some of the subtropical species release eggs in open water. Filefishes are generally small, and some reach a length of only 3.6 cm (1.4 in.); however, the largest member of the family, *Aluterus scriptus*, attains a length of 110 cm (43.3 in.). The family comprises approximately 102 marine species, which occur in the Atlantic, Indian, and Pacific oceans; about half of them occur in Australian waters.

Ostraciidae (boxfishes, cowfishes, trunkfishes). The Ostraciidae are characterized by the entire body (except the caudal peduncle) being encased in an immovable bony carapace consisting of fused hexagonal plates; no pelvic skeleton; no spinous dorsal; dorsal and anal fins each with 9–13 rays; nonprotrusible upper jaw; usually 18 vertebrae; and maximum length of 11–55 cm (4.3–21.6 in.), depending on the species. Some species secrete ostracitoxin, poisonous to other fishes and to some extent other ostracians. The habitat is usually seagrass and coral, where they feed on worms and other small sessile invertebrates. Some species use strong jets of water from the mouth to dislodge prey from the substrate (an interesting adaptation for a fish with a small nonprotrusible mouth and few moving body parts). Movement is slow and accomplished primarily by a sculling motion of the slender caudal peduncle. The family comprises two subfamilies, Aracaninae (a ventral ridge more or less developed and 11 principal caudal rays) and Ostraciinae (no ventral ridge and 10 principal caudal rays). Some authors recognize the subfamilies as separate families. There are about 33 species, some with beautiful bright colors, in the tropical and temperate waters of the Atlantic, Indian, and Pacific oceans.

Suborder Tetraodontoidei. These fishes do not have true teeth; instead the upper and lower jaws have sharp cutting edges that form a beak similar to that seen in the parrotfishes; there may be two, three, or four such "teeth," depending on the presence or absence of sutures; and the posttemporal bones are absent, as are usually the urohyal, pelvis, and pelvic fins.

Triodontidae (threetooth puffers). The single species of this family has three fused teeth (a median suture in the upper jaw, none in the lower); pelvis present; dorsal and anal fins each usually with 11 rays; caudal fin with 12 principal rays and deeply forked; ribs and epipleurals present; large belly flap; and maximum total length about 54 cm (21 in.). It is a reef associate of the Indo-West Pacific to depths of 50–300 m (164–984 ft).

Tetraodontidae (puffers). Members of this family are identified by a robust body, naked or with short prickles on sides and belly; some species with small fleshy appendages (lappets) on the sides; jaw teeth fused but separated by a median suture in each jaw, giving rise to four fused teeth; dorsal and anal fins far posteriorly, each consisting of 7–18 dorsal soft rays; ribs and epipleurals lacking; caudal fin with 10 principal rays and moderately forked to rounded; and maximum size of 3.5–120 cm (1.4–47.2 in.) total length, depending on the species. Puffers are capable of greatly inflating themselves with water or air when agitated. Some puffers contain the potential lethal poison tetrodotoxin, especially in the viscera and in the gonads of some during spawning season. They are chiefly marine, usually inhabiting shallow inshore waters of tropical and temperate waters of the Atlantic, Indian, and Pacific oceans; however, many species inhabit brackish water, but only 12 of about 162 species of puffers are limited to freshwater. *See* TOXIN.

Diodontidae (porcupinefishes, burrfishes). Members of this family are also capable of inflating the body with water or air, but differ from the puffers in having strong sharp spines and two fused teeth in the jaws, which form a powerful parrotlike beak. The spines of burrfishes are permanently erect, whereas those of the porcupinefishes are erect when the body is inflated. The family is further identified by opposite premaxillaries and dentaries being fused at midline. Maximum size is 27–91 cm (10.6–35.8 in.) total length. The family is represented by 19 species in tropical and temperate waters of the Atlantic, Indian, and Pacific oceans. Adults occupy inshore waters, whereas eggs and young are pelagic. Adults feed primarily on hard-shelled invertebrates crushed by the beak.

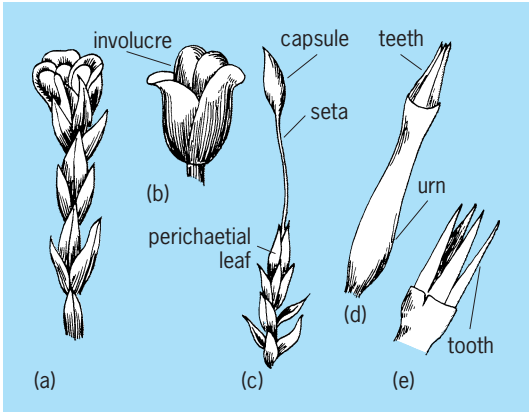
Molidae (molas). Molas are characterized by a moderately compressed body and thick leathery skin; tiny mouth and jaws with two fused beaklike teeth; no caudal peduncle or true caudal fin; posterior end of body reduced to a leathery flap (clavus) or a pseudocaudal fin formed by posteriorly migrated dorsal and anal fin rays; no pelvic fins; no lateral line; no swim bladder; and 16–18 vertebrae. Molas are strong swimmers, using the sculling motion of the powerful dorsal and anal fins. The ocean sunfish (*Mola mola*) is probably the most fecund fish with an estimated 300 million eggs. The young differ markedly from adults in having a spiny globular body. Adults feed largely on jellyfishes and other soft-bodied pelagic invertebrates and grow to 337 cm (11 ft) in total length and 1500 kg (3300 lb) in weight. Distribution is worldwide in tropical to temperate seas.

Herbert Boschung

Bibliography. T. Abe, Balistidae, in W. Fischer and P. J. P. Whitehead (eds.), *FAO Species Identification Sheets for Fishery Purposes: Eastern Indian Ocean (Fishing Area 57) and Western Central Pacific (Fishing Area 71)*, vol. 1, FAO, Rome, 1974; K. Matsuura, Balistidae: Triggerfishes, pp. 3911–3928 in K. E. Carpenter and V. Niem (eds.), *FAO Species Identification Guide for Fishery Purposes: The Living Marine Resources of the Western Central Pacific*, vol. 6, FAO, Rome, 2001; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006; C. R. Robins and G. C. Ray, *A Field Guide to Atlantic Coast Fishes of North America*, Houghton Mifflin, Boston, 1986; R. Santini and C. Tyler, A phylogeny of the families of fossil and extant tetraodontiform fishes (Acanthomorpha, Tetraodontiformes), Upper Cretaceous to Recent, *Zool. J. Linn. Soc.*, 139:565–617, 2003; M. M. Smith and P. C. Heemstra, Balistidae, pp. 876–882 in M. M. Smith and P. C. Heemstra (eds.), *Smith's Sea Fishes*, Springer-Verlag, Berlin, 1986; E. Tortonese, Molidae, pp. 1348–1350 in P. J. P. Whitehead et al. (eds.), *Fishes of the North-eastern Atlantic and the Mediterranean*, vol. 3, UNESCO, Paris, 1986; J. C. Tyler, Balistidae, Diodontidae, Monacanthidae, Ostraciidae, in W. Fischer (ed.), *FAO Species Identification Sheets for Fishery Purposes: West Atlantic (Fishing Area 31)*, vols. 1–5, FAO, Rome, 1978; J. C. Tyler, *Osteology, Phylogeny, and Higher Classification of the Fishes of the Order Plectognathi (Tetraodontiformes)*, NOAA Tech. Rep. NMFS Circ. 434, October 1980.

Tetraphididae

A subclass of the mosses (class Bryopsida) consisting of two families and three genera, especially characterized by growth from protonematal flaps, three-ranked leaves, and peristomes of four teeth made up of whole cells (rather than thickened parts of cells). The Tetraphididae include small acrocarpous mosses with peristome teeth in fours (see **illus.**). The plants grow from buds produced on leaflike protonematal flaps. They are erect and simple or merely forked, with oblong-ovate leaves in three rows. The leaves usually have a single costa that ends near the apex. The cells are short and smooth. Brood bodies are generally present in terminal clusters with or without a cuplike whorl of differentiated leaves. The setae are elongate, and the capsules are cylindric to



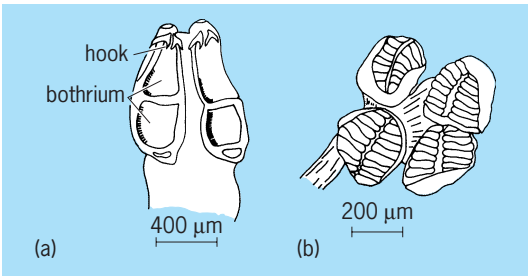
***Tetraphis pellucida*. (*a*) Gemmiferous branch; (*b*) involucre with gemmae (*after W. H. Welch, Mosses of Indiana, Ind. Dept. Conserv., 1957*). (c) Sporophyte; (*d*) capsule with peristome (*after H. S. Conrad, How To Know the Mosses, Jaques, 1944*). (e) Enlarged peristome (*after W. H. Welch, Mosses of Indiana, Ind. Dept. Conserv., 1957*).**

ovoid-cylindric and erect, with a distinct suboral rim of dark cells. The operculum is conic or subulate-pointed, and the peristome teeth are wedge-shaped, consisting of elongate cells thickened all around and derived from the entire tissue internal to the operculum. The calyptrae are mitrate and plicate. Chromosome numbers are 7 and 8. *See* BRYOPHYTA; BRYOPSIDA.

Howard Crum

Tetraphyllidea

An order of tapeworms of the subclass Cestoda. All species are intestinal parasites of elasmobranch fishes and are small in size, usually less than 2 in. (5 cm) in length. An outstanding feature of the order is the variation in the structure of the holdfast organ or scolex (see **illus.**). All species are segmented, and



Tetraphyllidea. (*a*) *Acanthobothrium* sp., scolex. (*b*) *Rhinebothrium* sp., scolex.

segments are usually shed from the body while sexually immature; these develop to sexual maturity as independent units in the host's intestine. Segment anatomy is very similar to that of Proteocephaloides. A complete life cycle is not known, but larval forms have been found in a variety of invertebrates and bony fishes. *See* EUCESTODA; PROTEOCEPHALOIDEA.

Clark P. Read

Tetrapoda

The superclass of the subphylum Vertebrata whose members typically possess limbs in contrast to the other superclass, the Pisces (fishes), whose members have fins. *See* PISCES (ZOOLOGY).

The animals making up the Tetrapoda typically live part or all of their lives on land, whereas the members of the Pisces live in water. The classes of the Tetrapoda are Amphibia (frogs and toads, salamanders, and caecilians), Reptilia (snakes and lizards, turtles, and crocodiles and their kin), Aves (birds), and Mammalia (mammals). The term Tetrapoda comes from Greek words meaning "four feet," but there are tetrapods that have only two limbs or none at all, such as some amphibians and reptiles. These forms have, however, evolved from four-footed ancestors.

The division of the vertebrates into the superclasses Tetrapoda and Pisces is in some respects merely a classification of convenience, and the use of characters other than limbs and fins can result in a different separation. For example, the eggs of reptiles and birds possess an embryonic membrane called the amnion that permits development of the embryo in relatively dry situations. The same membrane also surrounds a developing mammal, and these animals may be classified as the Amniota. Amphibians and other lower vertebrates do not have the amnion; their eggs survive only in water or under very moist conditions. These animals may be classified as the Anamnia. *See* AMNIOTA; AMPHIBIA; ANAMNIA; AVES; MAMMALIA; REPTILIA.

Richard G. Zweifel

Teuthoidea

An order of the class Cephalopoda (subclass Coleoidea) commonly known as squids. They are characterized by 10 appendages (eight arms and two longer tentacles) around the mouth; an elongate, tapered, usually streamlined body; an internal, rod- or bladelike chitinous shell (gladius); and fins on the body. The two tentacles are strongly elastic, contractile, but not retractile into pockets as in cuttlefishes (Sepioidea). Two rows of suckers (infrequently four or six rows) occur on the arms on muscular stalks, with sucker rings that are chitinous, smooth, toothed, or modified as clawlike hooks. The muscular tentacles have terminal clubs with two rows, usually four, ranging up to many rows of suckers (and/or hooks in some families). Adults of the family Octopoteuthidae and genera *Gonatopsis* and *Lepidoteuthis* characteristically lose their tentacles. *See* SEPIOIDEA.

The Teuthoidea are divided into two suborders. The Myopsida, the nearshore, shallow-water squids, have a transparent skin (cornea) covering the two eyes, with a minute pore anteriorly, arms and tentacular clubs with suckers only, never hooks, and a single gonoduct in females, not paired. The Oegopsida, the oceanic squids, have no cornea over the eyes and no anterior pore, arms and tentacular clubs with suckers (and/or hooks in many families), and paired gonoducts in females (some exceptions).

Squids inhabit a wide variety of marine habitats, depending on the species, from very shallow grass flats, mangrove roots, lagoons, bays, and along coasts (myopsids) to the open ocean from the surface of the sea (*Ommastrephes*) to nearly 9600 ft (3000 m) in the deep sea (other oegopsids such as *Bathyteuthis*, *Neoteuthis*, and *Grimalditeuthis*). Most species require full oceanic salinities, 32–37‰ (parts per thousand), but some, like *Lollinguncula brevis*, can tolerate salinities as low as 8.5‰ in the warm-water estuaries and bays of the Atlantic coasts of North and South America and the Caribbean. While some species of squids live near the bottom where they feed, spawn, or rest, principally the myopsids, most species live near the surface or in the middepths away from the bottom. Even the deepest-living species of oegopsids live far above the abyssal ocean bottom.

Teuthoids are the masters of animal jet propulsion; they suck water into the mantle cavity, then forcibly eject it through the funnel, driving the squid

tail-first through the water. The propulsion system is controlled from the well-developed brain by a complex of nerves concentrated in a cluster in the mantle, chief of which is the giant axon, the largest single nerve fiber in the animal kingdom. The highly developed brain and nervous system also support large, acute eyes that rival those of fishes, the ability to rapidly change color and color patterns through thousands of pigment cells (chromatophores), and the capacity to produce light (bioluminescence) from specialized organs (photophores).

The sexes are separate, and the male transfers sperm to the female, often following dramatic and colorful courtship displays, in cylindrical packets (spermatophores), frequently via a specially modified arm, the hectocotylus. The fertilized eggs are laid in gelatinous fingerlike masses attached to the bottom (myopsids) or in gelatinous sausages or balls, or as individual eggs, in the open sea (oegopsids). Embryos undergo direct development and upon hatching are planktonic young that look much like adults; no dramatic larval stages occur in squids, as they do in most other marine mollusks and invertebrates. Adult teuthoids can be from about 1 in. (25 mm) total length (myopsid, *Pickfordiateuthis pulchella*) to 60 ft (18 m) total length, the giant squid (oegopsid, *Architeuthis* species). Squids are preyed upon by fishes, toothed whales, seals, and sea birds, and they account for around 10^6 metric tons captured annually in fisheries around the world for nutritious human food. Scientists study the biology and behavior of squids, as well as their giant axons for neurophysiology, pharmacology, and biophysics. *See* CEPHALOPODA; COLEOIDEA; SQUID. Clyde F. E. Roper

Bibliography. K. N. Nesis, *Cephalopods of the World*, transl. from Russian, 1987; C. F. E. Roper, M. J. Sweeney, and C. E. Nauen, *Cephalopods of the World*, FAO Fisher. Synop. 125, vol. 3, 1984; E. R. Truman and M. R. Clarke (eds.), *The Mollusca*, vol. 11: *Form and Function*, 1988.

Textile

A material made mainly of natural or synthetic fibers. Modern textile products may be prepared from a number of combinations of fibers, yarns, films, sheets, foams, furs, or leather. They are found in apparel, household and commercial furnishings, vehicles, and industrial products. Materials made solely from plastic sheet or film, leather, fur, or metal are not usually considered to be textiles. *See* LEATHER AND FUR PROCESSING; MANUFACTURED FIBER; NATURAL FIBER; PLASTICS PROCESSING.

The term fabric may be defined as a thin, flexible material made of any combination of cloth, fiber, or polymer (film, sheet, or foams); cloth as a thin, flexible material made from yarns; yarn as a continuous strand of fibers; and fiber as a fine, rodlike object in which the length is greater than 100 times the diameter. The bulk of textile products are made from cloth, and this discussion centers on the most common forms of cloth and their manufacture.

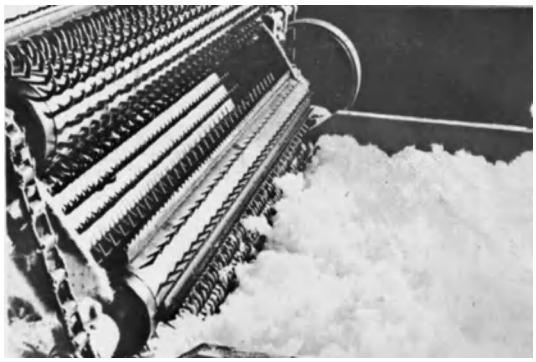


Fig. 1. Cotton going through opening machine where fibers are loosened and straightened. (*Pepperell Mfg*.)

The natural progression from raw material to finished product requires the cultivation or manufacture of fibers; the twisting of fibers into yarns (spinning); the interlacing (weaving) or interlooping (knitting) of yarns into cloth; and the finishing of cloth prior to sale.

Spinning processes. The ease with which a fiber can be spun into yarn is dependent upon its flexibility, strength, surface friction, and length. Exceedingly stiff fibers or weak fibers break during spinning. Fibers which are very smooth and slick or fibers which are very short do not hold together. To varying degrees, the common natural fibers (wool, cotton, and linen) have the proper combinations of the above properties. The synthetic fibers are textured prior to use in order to improve their spinning properties by simulating the convolutions of the natural fibers. Natural and synthetic filament fibers, because of their great length, need not be twisted to make useful yarns.

The properties of a yarn are influenced by the kind and quality of fiber, the amount of processing necessary to produce the required fineness, and the degree of twist. The purpose of the yarn determines the amount and kind of processing. The yarn number (yarn count) is an indication of the size of a yarn—the higher the number, the finer the yarn. The degree of twist is measured in turns per inch (tpi) and is varied from three to six times the square root of the yarn number for optimum performance.

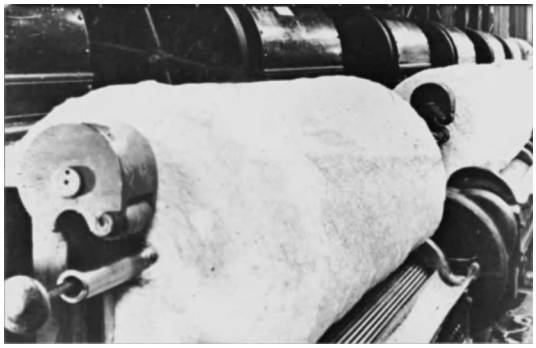


Fig. 2. Cotton lap from the picker room where the dust, leaves, twigs, and other foreign matter have been removed. (*Pepperell Mfg*.)

Fig. 3. Sliver leaving carding machine, where cotton is further cleaned and disentangled. (*Pepperell Mfg.*)

Fig. 4. After carding, the slivers are doubled to increase the density of the future cotton yarn. (*Pepperell Mfg.*)

The conversion of staple fiber into yarn requires the following steps: picking (sorting, cleaning, and blending), carding and combing (separating and aligning), drawing (reblending), drafting (reblended fibers are drawn out into a long strand), and spinning (drafted fibers are further attenuated and twisted into yarn).

Picking. Raw fibers are delivered to the spinster from a number of suppliers. They must be sorted, trash must be removed, and fibers must be blended in order to assure quality and uniformity. **Figure 1** shows tufts of raw cotton being fed into an opener where the spiked teeth will pull the tufts apart and separate the fiber from its impurities. The product of this step is the picker lap (**Fig. 2**).

Carding. The picker lap is still too tangled and coarse to be suitable for spinning. Fibers are aligned in the carding process, in which the lap is unrolled and drawn on a revolving cylinder covered with very fine wire brushes. A moving belt covered with wire brushes lies atop the cylinder. The motion of the belt and cylinder pulls the fibers, disentangles them, and arranges them in parallel in the form of a thin web. The web is drawn through a funnel-shaped tube that molds it into a round ropelike mass, about the thickness of a broomstick, called sliver (**Fig. 3**). Card sliver produces yarns serviceable for inexpensive fabrics. After carding, several slivers are combined to form a relatively narrow lap of compacted staple fibers. The compactness of these fibers permits this stock to be drawn out to a sliver of smaller diameter without falling apart (**Fig. 4**).

Combing. When the fiber is intended for fine yarns, the sliver is put through an additional alignment. In this operation, fine-toothed combs continue straightening the fibers until they are arranged with such a high degree of parallelism that the short fibers are combed out and separated from the long fibers. The comb sliver is made of the longest fibers and produces a smoother and more even yarn. The combined process, therefore, is identified with better-quality goods.

Drawing and drafting. The combining of several slivers in the drawing process eliminates irregularities which would affect yarn quality. The drawing frame has several pairs of rollers, each advanced set of which revolves at a progressively higher speed. This action pulls the fibers lengthwise over each other and produces a longer, thinner sliver. After several stages of drawing, the condensed sliver is given a slight twist and wound onto bobbins. The product is called roving.

Spinning. The roving is placed on the spinning frame (**Fig. 5**) and passed through another set of drafting rolls. The highly attenuated yarn is fed onto a high-speed spindle by a guide (traveler) which rotates on a ring surrounding the spindle. The traveler rotates at a speed slightly slower than that of the spindle. This difference in speed causes twist to be inserted in the yarn. At the same time, the traveler is raised and lowered along the length of the spindle to form the yarn into a neat package.

It should be noted that modern spinning plants are equipped with integrated machines which perform many of these operations within a single unit; for example, opening, blending, and carding are now combined in one operation. In addition, the new open-end spinning systems can convert card sliver into yarn without the necessity of separate drawing and drafting steps.

Weaving. The process of weaving allows a set of yarns running in the machine direction (warp) to be interlaced with another set of yarns running across

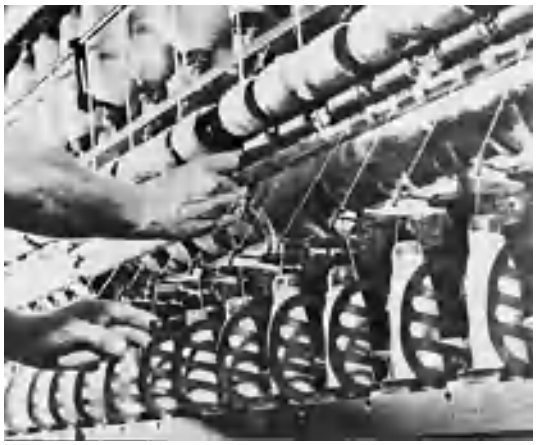


Fig. 5. On spinning frame, roving passes from the top through a series of rollers that draw out the cotton into thread. The thread is twisted as it is being wound onto the bobbin. (*Pepperell Mfg.*)

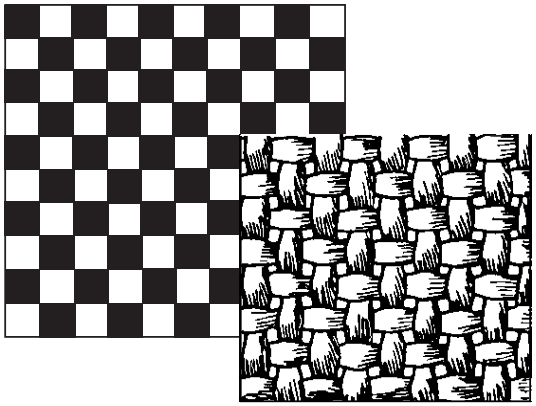


Fig. 6. Construction design for plain weave; filling yarns pass under and over alternate warp yarns, as shown at right. When fabric is closely constructed, there is no distinct pattern. (*After M. D. Potter and B. P. Corbman, Fiber to Fabric, 3d ed., McGraw-Hill, 1959*)

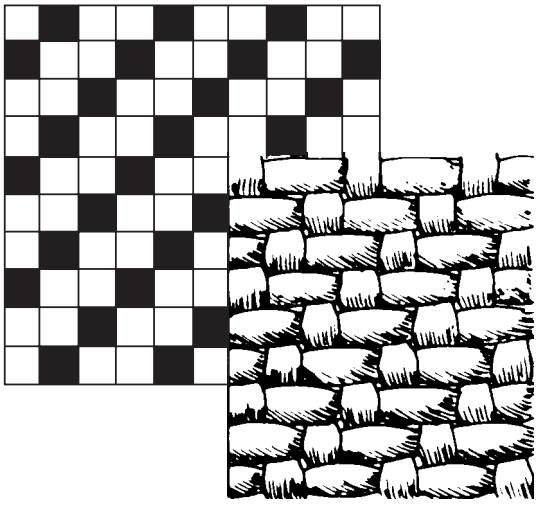


Fig. 7. Three-shaft twill. Two warp yarns are interlaced with one filling yarn. (*After M. D. Potter and B. P. Corbman, Fiber to Fabric, 3d ed., McGraw-Hill, 1959*)

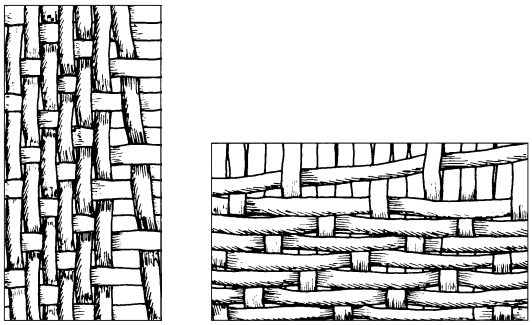
the machine (filling or weft). The weaving process involves four functions: shedding (raising the warp yarns by means of the appropriate harnesses); picking (inserting the weft yarn); battening (pushing the weft into the cloth with a reed); and taking up and letting off (winding the woven cloth onto the cloth beam and releasing more warp yarn from the warp beam).

Mechanical systems. From the earliest handlooms to the most innovative weaving machines, weaving systems have six basic components: a warp beam, heddles, harnesses, a means for inserting the filling yarns (shuttle, rapier, or fluid jet), a reed, and a takeup roll. The warp beam is a large metal spool capable of holding about 4500 separate yarns, each about 500 yards (450 m) long; it holds the warp yarns and feeds them into the cloth. Heddles are thin, flat metal rods pierced by a central slot, and are used for separating and guiding the warp yarns. Harnesses are frames which can be used to raise and lower groups of heddles. The reed is a large, comblike device that packs the filling yarns into the cloth so that they do not slip. The takeup roll looks much like a warp beam, and is used to hold the woven cloth. The warp yarns are raised preparatory to inserting the weft yarn.

In a process called beaming, the warp yarns are taken from packages mounted on a large rack (creel), passed through a comb which separates and guides them, passed through a bath of starch or other viscous material (sizing) which applies a protective lubricant, and wound onto the warp beam. The positioning of the yarns in the heddles is determined by the particular weave pattern. The simplest weaves require only two harnesses, while more complex patterns may take up to 40. The warp yarns are fixed to the takeup roll after passing through the reed. The filling yarns are wound on small spindles (quills) which can be mounted in the shuttle. Most looms are equipped with automatic devices which change quills when they have delivered their yarn.

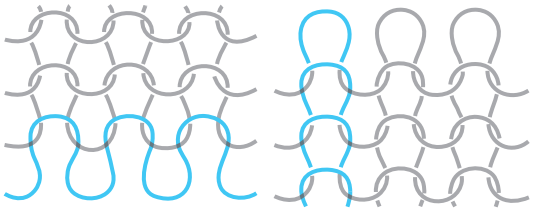
Classification of weaves. The manner in which groups of yarns are interlaced determines the pattern of the weave and, to a great extent, the properties of the cloth. The three basic weaves are the plain, twill, and satin (**Figs. 6–8**) and their respective variations. Important constructions are also obtained from more elaborate weaves such as dobby, swivel, lappet, gauze, double cloth, and pile. Jacquard looms can produce cloth of extremely intricate patterns.

Knitting. Knit cloth is produced by interlocking one or more yarns through a series of loops. The lengthwise columns of loops are known as the wales, and the crosswise rows of loops are called courses. Filling (weft) knits (**Fig. 9**) are those in which the



(a) (b)

Fig. 8. Long floats typical of satin and sateen weaves: (*a*) five-shaft satin construction with floats in warp direction; (*b*) eight-shaft sateen construction with floats in the filling direction. (*After M. D. Potter and B. P. Corbman, Fiber to Fabric, 3d ed., McGraw-Hill, 1959*)



(a) (b)

Fig. 9. Interlocking yarns of (*a*) course and (*b*) wale in a jersey knit cloth. (*After B. P. Corbman, Fiber to Fabric, 5th ed., McGraw-Hill, 1975*)

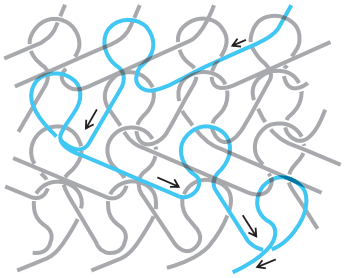


Fig. 10. Single-warp (one-bar) tricot knit. (*After B. P. Corbman, Fiber to Fabric, 5th ed., McGraw-Hill, 1975*)

courses are composed of continuous yarns, while in warp knits (**Fig. 10**) the wale yarns are continuous.

Weft. This type of knitting is accomplished on linear (flat knits) or cylindrical (circular knits) machines in which latch needles are raised and lowered by the action of cams (**Fig. 11**). As the needle rises, the hooked end loops over the yarn. As the needle descends, the yarn is trapped within the crook by the previously formed loop. At the bottom of its motion, the earlier loop slips off the needle, while the new loop is held within the closed needle. At the next traverse of the cam, the process is repeated.

Warp. This type of knitting differs from weft knitting in that each needle is supplied with its own yarn, taken from a warp beam similar to that used in weaving. A yarn guide lays the yarn across the ascending needle, and returns to position as the needle descends. The yarn, which is fixed at one end by the previous loop and at the other by the yarn guide, is trapped within the crook of the spring beard needle as the needle descends. A depresser bar closes the hook, allowing the old loop to slip off while the new loop is held.

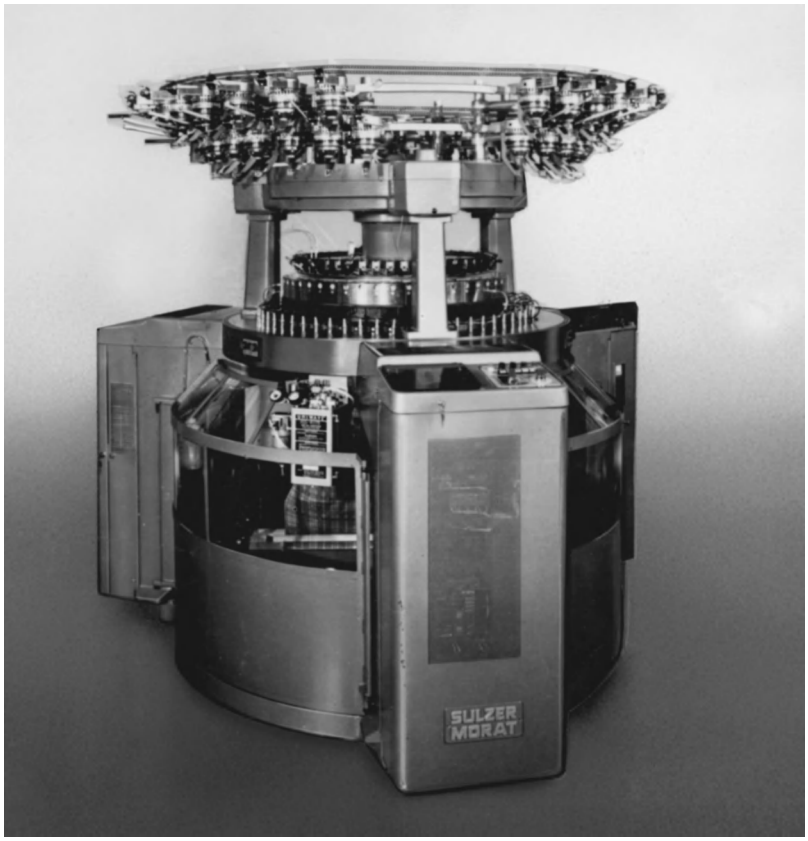


Fig. 11. Circular Jacquard knitting machine. (*Sulzer Co.*)

Classification of knits. In weft knitting, four stitches may be formed: knit, in which the loop is drawn from the back to the front of the cloth; purl, in which the loop is drawn from the front to the back of the cloth; miss, in which no loop is formed; and tuck, in which two courses on one wale are looped over a third. From these stitches all of the patterns of knit and double knit cloths can be made. The basic knit cloths are the jersey, made from only knit or only purl stitches; the rib, made from alternating wales of knit and purl stitches; and the purl, made from alternating courses of knit and purl stitches.

Simple warp knitting machines make an open loop, a closed loop, or no loop. Variations in the arrangement of these elements provide the different types of patterned fabrics. The simplest of the warp knits is the tricot, in which parallel wales of closed loops are interlocked in a zigzag pattern. More intricate forms of warp knitting may be used to produce fibers which range from simple, coarse netting to intricate, sophisticated lace.

Finishing. Newly constructed knit or woven fabric is not suitable for sale. It must pass through various finishing processes to make it suitable for its intended purpose. Finishing enhances the appearance of fabric and also adds to its serviceability. Finishes can be solely mechanical, solely chemical, or a combination of the two. Those finishes, such as scouring and bleaching, which simply prepare the fabric for further use are known as general finishes. Functional finishes, such as durable press treatments, impart special characteristics to the cloth. For discussions of important finishing operations *see* BLEACHING; DYEING; TEXTILE CHEMISTRY; TEXTILE PRINTING.

Performance. The choice of a fabric is determined by the requirements of appearance, comfort, ease of maintenance, durability, and cost. Appearance is affected by light reflection, pattern, and drape. Comfort is sensitive to the diffusion of air and moisture through the cloth, texture, elasticity, and stretch. Ease of maintenance includes soil and stain resistance and removal, wrinkle resistance and recovery, and dimensional stability. Durability is measured by strength, abrasion resistance, and the ability to withstand attack by sunlight, atmospheric fumes, insects, and microorganisms. These factors are all affected, to different degrees, by the choice of fiber, yarn construction, fabric type, and finish. In general, woven cloths provide greater durability than knit cloths, while knit fabrics tend to be more comfortable.

The density of a woven construction is measured by the yarn count (thread count), which is defined as the number of warp and the number of filling yarns in 1 in.^{2} (6.5 cm^2) of cloth. The density of a knit cloth is measured by the gauge, which is defined as the number of stitches in $1\frac{1}{2}$ in. (3.8 cm). Heavy, compact constructions are more durable than

lightweight, open fabrics. Usually the more intricate the pattern, the more costly the material. Ira Block

Bibliography. B. P. Corbman, *Textiles: Fiber to Fabric*, 6th ed., 1983; M. L. Joseph, *Introductory Textile Science*, 5th ed., 1993; M. L. Joseph and A. Giesecking-Williams, *Illustrated Guide to Textiles*, 4th ed., 1986; D. S. Lyle, *Modern Textiles*, 2d ed., 1982; B. F. Smith and I. Block, *Textiles in Perspective*, 1996; P. G. Tortora, *Understanding Textiles*, 6th ed., 2000.

Textile chemistry

The applied science of textile materials, consisting of the application of the principles of the many basic fields of chemistry to the understanding of textile materials and to their functional and esthetic modification into useful and desirable items. Textile chemistry includes the study of the chemical and physical properties of the various textile fibers, both natural and synthetic. Included also is the application of the principles of surface chemistry to the many cleaning operations and modifications of the textile material. Finally, there is the application of organic chemistry to the synthesis and formulation of the chemical products necessary for achieving these objectives.

Being primarily an applied form of chemistry, textile chemistry draws many of its concepts from the more basic and theoretical fields. At the same time, it is a highly specialized field which requires the blending of theory and knowledge with empirical information. The textile chemist must be able to relate knowledge of the organic structures of both the fibers and the chemicals being used to specific chemical, physical, and esthetic properties. This combination of the theoretical and the practical makes possible the development of the thousands of different textile chemicals necessary for the production of finished articles of commerce. *See* TEXTILE.

Fibers. The study of textile chemistry begins with the knowledge of the textile fibers themselves. These are normally divided into two groups: natural and manufactured, those produced from substances of natural or synthetic origin. The natural fibers include the protein fibers (silk and wool), as well as the cellulose and cellulose-related fibers (cotton, linen, ramie, and jute), plus certain other minor vegetable fibers. The manufactured fibers of natural origin are derived by chemical and physical modification of the naturally occurring fibrous materials, particularly cellulose; the important members of this group are rayon, acetate, and triacetate. The manufactured fibers of synthetic origin, also known as synthetic fibers, are produced by polymerization of synthetic chemicals; the starting point for these fibers is not a naturally occurring polymer. This group includes the polyamides (nylons), polyacrylics, and polyesters. The synthetic fibers share many common properties, but also differ greatly in certain ways because of their varying chemical natures. *See* MANUFACTURED FIBER; NATURAL FIBER; POLYACRYLONITRILE RESINS; POLYAMIDE RESINS; POLYESTER RESINS.

Chemicals. The enormous number of chemicals used in textile processing may be divided broadly into two categories: those intended to remain on the fiber, and those intended to wet or clean the fiber or otherwise function in some related operation. The former includes primarily dyes and finishes, which will be covered in the discussion of the appropriate operations. The latter group consists mainly of surface-active agents, commonly known as surfactants. These chemicals are characterized by the general structure R—Y, in which R, the major portion of the molecule, is oil-soluble, and Y is extremely water-soluble. The balance between these two sections determines the properties and uses of the surfactant. Typically, R is a long hydrocarbon chain. However, Y may be charged positively or negatively, or be uncharged. Some common Y groups are:

Anionic: carboxylate, $—COO^-$
sulfate, $—O—SO_3^-$
sulfonate, $—C—SO_3^-$
phosphate, $—PO_4^-$
Cationic: quaternary ammonium, $—N^+(CH_3)_3$
Nonionic: ethyoxylate, $(—CH_2—CH_2—O)_n$

Thus, probably the oldest surfactant is soap, sodium stearate, $CH_3(CH_2)_{15}CH_2—COO^-Na^+$. Choice of a particular combination of structures will depend on the specific use intended for the chemical, and will be illustrated in discussing the textile processes. *See* DETERGENT; QUATERNARY AMMONIUM SALTS; SOAP; SURFACTANT.

Processes. All textile fibers go through a variety of physical and chemical processes designed to produce useful articles of commerce. Although the order may vary in certain instances, there is a general sequence of physical and chemical operations, beginning with the fiber itself and ending with the finished fabric ready for use by its ultimate consumer. The common order of operations is: aligning the fiber and spinning the yarn; weaving or knitting the yarn into the fabric; preparation; fabric coloring or printing; and fabric finishing.

The major physical operations are, of course, spinning the fibers into yarns, followed by weaving or knitting these yarns into fabrics. Although a certain percentage of textile materials reaches the ultimate consumer in the form of yarn, unquestionably the major usage is in woven, knitted, or nonwoven fabrics. Spinning, weaving, and knitting are almost completely physical operations, although chemical products such as lubricants and sizings composed of starch or other polymers are used to protect the fibers and yarns from damage. These materials must eventually be removed, and thus knowledge of their composition is important to the textile chemist. In addition to these physical operations, textile fibers, either in that form or in the form of yarn or fabric, are usually subjected to a variety of chemical operations; the study of these operations constitutes a major portion of the field of textile chemistry.

Preparation. Preparation is a term applied to a group of essentially wet chemical processes having

as their object the removal of all foreign matter from the fabric. This results in a clean, absorbent substrate, ready for the subsequent coloring and finishing operations.

Since the object of preparation is the removal of foreign matter from the fiber, and since the nature of these impurities is dependent on the fiber itself, the operations constituting preparation depend primarily on the fibers being handled. Synthetic fibers contain little or no natural impurities, so that the only materials that normally must be removed are the oils and lubricants or water-soluble sizes needed to facilitate earlier processing. This is generally accomplished by washing with water and a mild detergent capable of emulsifying the oils and waxes. A typical material for this use would be a nonionic with a relatively low degree of ethoxylation. On the other hand, natural fibers contain relatively high amounts of natural impurities, and in addition frequently are sized with materials presenting difficulties in removal. In the case of cotton, prolonged hot treatment with alkali, usually sodium hydroxide, and strong detergent is necessary to break down and remove the naturally occurring impurities. Sizing, which is usually based on starch, must be broken down by treatment with enzymes or peroxide, substances which attack the starch molecule and convert it to more soluble materials. After these treatments, strenuous washing is necessary to remove the residue from the cotton fiber. Sulfates and sulfonates are frequently used in these operations, but the phosphate surfactants are generally preferred, because of their superior functioning in hot alkaline solutions, and their ability to disperse and suspend solid soil. *See* STARCH; SULFONATION AND SULFATION.

Special scours are necessary for cleaning such materials as wool and silk: these are actually processed in relatively small volume in the United States. The protein fibers are very sensitive to alkali and strong detergents; they are usually washed with mild soap or sulfated alcohols. Agents having a protective action on the fiber, such as highly sulfated castor oil, are frequently included. Wool may be carbonized, that is, treated with strong sulfuric acid to destroy vegetable matter adhering to the fiber.

After other impurities are removed from the fiber, it is usually desirable to remove any coloring material. This process, known as bleaching, is necessary to some degree with certain synthetics; it is essential with the natural fibers if white or light-colored fabrics are to be prepared. Several chemicals destroy colored impurities. Sodium hypochlorite and sodium chlorite were used in the past, and are still used to some extent. However, by far the major bleaching agent in use is hydrogen peroxide, which is efficient in color removal, while still being considered relatively controllable and safe for use. *See* BLEACHING; HYDROGEN PEROXIDE.

The preparation of blends of fibers frequently presents some difficulty, but in general it may be stated that the blend is prepared in a manner similar to that component which in itself is most difficult to clean. Thus, a polyester-cotton blend would be prepared in a manner relatively similar to a 100% cotton fabric.

Mercerization. Mercerization is a special process applied only to cotton. The fabric or yarn is treated with a strong sodium hydroxide solution, usually about 20% strength, while being held under tension. This process causes chemical and physical changes within the fiber itself, resulting in a substantial increase in luster and smoothness of the fabric, plus important improvements in dye affinity, stabilization, tensile strength, and chemical reactivity. It is necessary in this process that the caustic solution penetrate quickly and uniformly, and equally important that after reaction has taken place, the caustic be removed rapidly and thoroughly. Special surface-active agents have been developed to offer maximum effectiveness in highly concentrated sodium hydroxide solutions. Originally these were based on cresylic acid, the mixed isomer of cresol, but increasingly strict pollution controls have led to the development of noncresylic penetrants. These are typically sulfate esters of relatively short-chain alcohols, such as hexyl or octyl: $CH_3(CH_2)_nCH_2—O—SO_3Na$ ($n = 4$ or 6).

Coloring. Although many textiles reach the consumer in their natural color or as a bleached white, most textiles are colored in one way or another. Coloring may be accomplished either by dyeing or printing, and the coloring materials may be either dyes or pigments.

Dyeing. In its simplest form, dyeing consists of immersing the fabric or yarn in a solution of a dyestuff. This dyestuff in one way or another is attracted to the fiber, and travels from the solution into the interior of the fiber. By contrast, a pigment is soluble neither in the solution nor in the fiber itself; it is simply deposited on the surface of the fabric, where it is held, or glued, by an insoluble binder. There are some dyeing procedures which appear to omit some of the normal steps, and there are also certain procedures which partly involve working with the dye in pigment form. However, the important distinction is that a dye exists in solution or ultrafine dispersion in the solvent, usually water, and is deposited or dissolved within the interior of the fiber; a pigment is never soluble during any stage of the process, and is deposited and bound to the surface of the fiber only. *See* DYEING.

Printing. Dyeing essentially consists of immersing the entire fabric in the solution, so that the whole fabric becomes colored. Printing may be considered localized dyeing. In printing, a thickened solution of dyestuff or pigment is used. This thickened solution, or paste, is applied to specific areas of the fabric by means such as engraved rollers or partially porous screens. Application of steam or heat then causes the dyestuff to migrate from the dried paste into the interior of the fiber, but only in those specific areas where it has been originally applied. Printing makes possible extremely complicated, multicolored designs, without resort to difficult and unusual weaving or knitting procedures. *See* TEXTILE PRINTING.

Dyestuffs. The choice of dyestuff depends on the physical and chemical nature of the fiber. Fibers

such as wool and nylon contain positively charged NH groups, and thus the dyes used contain negative groups such as the sulfonate group (SO_3). On the other hand, acrylic fibers have negative sulfonate groups present in the form of copolymers, and are thus processed with dyes carrying a positive charge. Polyester offers no such ionic sites, and thus dyes used for polyester contain no charge, but are in the form of ultrafine particles, which form solid solutions in the fiber. Cotton contains only hydroxyl groups as dye sites, and thus the dyes used may be fiber-reactive, forming covalent bonds, or direct, forming hydrogen bonds with these groups. Other cotton dyes, such as vats and naphthols, are instead converted to insoluble derivatives in the fiber. *See* DYE.

Other dyeing chemicals. Many textile chemicals are used in the dyeing process, besides the actual dyestuffs themselves. It is necessary that the dyestuff penetrate as fully as possible into the individual fibers, and be as uniformly distributed as possible. Certain chemicals aid the penetration of the dyestuff; others retard the penetration, and are used to permit the color to build up at a gradual rate, thus ensuring uniformity. Another type of textile chemical used in dyeing is the migrating agent, which aids the individual dyestuff molecules in moving from one site to another within the fiber, thus promoting uniformity in a different way.

Penetrants may be surfactants of appropriate type or, in the case of polyester or certain other fibers, may be swelling agents. Retarders and migrating agents may be of two types. Certain chemicals have structures similar to, but simpler than, the dyestuffs. These smaller molecules occupy the available dye sites before the dyestuff molecule, but eventually are replaced by it.

The other method of retardation is the use of either a material of reverse charge or a nonionic, or both, to form an unstable complex with the dyestuff molecule.

Finally, certain dyestuffs are after treated with chemicals designed to increase their resistance to being washed out of the fiber. The textile chemist must therefore have a knowledge of the fiber, the dyestuff, and the dyeing auxiliaries. Many of the same factors apply also to printing. In addition, a knowledge of the different types of thickening agents available is also essential. These agents may be derived from natural sources, such as starch or gum, or may be a synthetic material such as a polyacrylate. Proper choice of thickener is absolutely essential for effective printing.

Finishing. Finishing includes a group of mechanical and chemical operations which give the fabric its ultimate feel and performance characteristics. The fabric may be compressed, to minimize shrinkage, or the surface may be polished or roughened; these are all essentially mechanical operations. On the other hand, many desirable characteristics may be imparted to the fabric through the application of various chemical agents at this point.

Chemical operations. Softeners are used to give a desirable hand or feel to the fabric. These chemicals are generally long fatty chains, with solubilizing groups which may be cationic, nonionic, or occasionally anionic in character. They are similar to many of the surfactants mentioned earlier, but are constructed so as to contain a higher proportion of fatty material in the molecule. Thus poly(ethylene glycol) monostearate is a typical nonionic softener. Similarly, dimethyl tallow ammonium chloride is widely used as a cationic softener.

Other softeners are emulsions formed of oil, wax, or polyethylene, using nonionic or cationic surfactants as emulsifiers. By proper selection of softener, it is possible to make a fabric feel slick, or limp, or lofty. Conversely, certain types of polymeric material such as poly(vinyl acetate) or polymerized urea-formaldehyde resins are used to impart a stiff or crisp hand to a fabric. *See* POLYVINYL RESINS; UREA-FORMALDEHYDE RESINS.

It is in finishing that the so-called proof finishes are applied, including fire-retardant and water-repellent finishes. A fire-retardant finish is a chemical or mixture containing a high proportion of phosophorus, nitrogen, chlorine, antimony, or bromine. For example, the combination of tetrakishydroxymethyl phosphonium chloride (THPC) with trimethylol melamine (TMM) is effective in producing fire-retardant fabrics. A truly waterproof fabric may be made by coating with rubber or vinyl, but water-repellent fabrics are produced by treating with hydrophobic materials such as waxes, silicones, or metallic soaps.

In the case of cellulosic fabrics, various cross-linking agents are applied to react with the cellulose chains and impart resistance to shrinking and creasing. The most widely used cross-linking agents are *N*,*N*-dimethylol carbamate and dimethylol dihydroxyl ethylene urea. These materials form covalent links between adjacent cellulose chains.

Other types of highly specialized treatments, such as antistatic, antibacterial, or soil-repellent finishes, may be applied to fit the fabric to a particular use.

Mechanical operations. Finishes are usually applied by an operation known as padding, which consists of drawing fabric through a water solution of the chemical being applied, followed by passage through a set of squeeze rolls which removes the excess liquid. A wetting agent, or penetrant, is frequently included in this padding solution to aid penetration into the fabrics. The amount of chemical deposited within the fabric is thus determined first by the concentration of the chemical in the solution, and second by the amount of pressure applied by the squeeze rolls, and thus the amount of solution allowed to remain in the goods. Typically, a fabric may still retain 50–100% of its own weight in solution, and a considerable amount of water must remain to be evaporated.

Evaporation of unneeded water is an energy-intensive, and thus expensive, procedure, and attempts have been made to devise means of reducing the amount of water to be evaporated. In one approach, the concentration of chemical in the solution is greatly increased, and this solution is

sprayed on the fabric. Unfortunately, serious difficulties have been encountered in maintaining uniform distribution. Another procedure offering considerable promise is foam finishing; in this technique a concentrated solution containing the chemical to be applied, as well as a foaming agent, is converted into a foam of specific density and wetness by mechanical beating and injection of air. In effect, air is substituted for the major part of the water as a diluent for the chemicals. *See* FOAM.

Techniques involving use of organic solvents in place of water have been demonstrated, as well as other more novel finishing techniques. It is safe to assume that one area of textile chemical research to become increasingly important will be development of low-energy processing techniques.

Garment processing. In traditional processing, the operations of preparing, dyeing, and finishing are applied to the fabric; this fabric is later cut and sewed into garments. In garment processing, the fabric may be converted into garments at any time; the remaining operations are performed on the garments in large laundry washing machines.

Dyeing by this method is not nearly as uniform as a classical yarn or fabric dyeing. However, casual styles have been developed and accepted for these unevenly dyed garments. The main advantage is the reduction of inventories; garments are colored in response to actual sales.

A variety of this form of processing is the creation of partially worn-out garments. As a full load of garments is tumbled in a washing machine, chemicals such as potassium permanganate or sodium hypochlorite are used to destroy color irregularly. Furthermore, rocks, pumice stones, or cellulose-attacking enzymes are introduced to weaken and fray the garment, creating an overall worn and used appearance. In fact, the garment actually is somewhat weakened so that the result is poorer quality.

David H. Abrahams

Bibliography. S. V. Kulkarni, *Textile Dyeing Operations: Chemistry, Equipment, Procedures, and Environmental Aspects*, 1986; M. Lewin, *Handbook of Fiber Chemistry*, 2d ed., 1998; E. R. Trotman, *Dyeing and Chemical Technology of Textile Fibers*, 6th ed., 1991; T. L. Vigo, *Textile Processing and Properties*, 1994.

Textile microbiology

That branch of industrial microbiology concerned with textile materials. Most of the microorganisms on textiles—the fungi, actinomycetes, and bacteria—originate from air, soil, and water. Some of the microorganisms are harmful to either the fibers or the consumer. They may decompose the cellulose or protein in the fiber or affect the consumer's health. Since the minimum moisture content for microorganism development is 7%, dry storage is an effective prevention measure. Some of the microorganisms are useful, for example in the retting process, in which fibers are liberated from the stalks of such fiber plants as flax, hemp, and jute.

Cotton. Microbial attack of cotton can occur at any time from the opening of the boll. The fiber or fabric may be degraded, resulting in a decrease in length of the fiber and strength of the cloth, uneven dyeability, darkening, or formation of colored spots. A large variety of fungi can be active in this process (**Fig. 1**). Representatives of the genera *Chaetomium* and *Myrothecium* have the highest cellulose-decomposing activity; representatives of *Alternaria*, *Cladosporium*, *Fusarium*, and *Diplodia* are active in field cotton; *Aspergillus*, *Penicillium*, and *Stachybotrys* are active in stored cotton and fabrics (mildewing). *See* ASCOMYCOTA.

The number of contaminating bacteria varies from about 1 million to 100 million per gram of raw cotton. Among these are the cellulose-decomposing bacteria (**Fig. 2**). They all belong to the Myxobacteriales and have been found on field cotton where they probably play an important role in degradation. *Aerobacter cloacae* has also been found on raw cotton. This organism is of interest because it causes a respiratory disease.

Cotton is examined for microbial damage by a series of tests. In the Congo red test, microscopic examination is made of the cotton fibers after a dye, Congo red, has been applied. The undamaged cotton fibers appear pink, and the damaged cotton fibers appear red and may be cracked. In the swelling test,

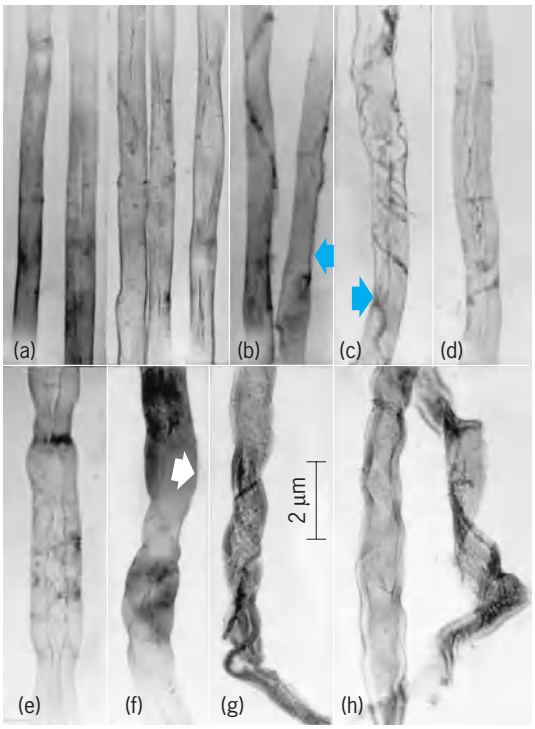


Fig. 1. Stages of deterioration of cotton fiber under influence of fungi. (*a*) Normal fibers. (*b*, *c*) Fungal hyphae growing on the outside of the fiber (shown by arrows). (*d*) Fungal hyphae growing in the lumen of the fiber. (*e*, *f*) Fibers showing excessive and irregular swelling. (*e*) Cavitomic stage. (*f*) Mildewed fiber with cuticle damaged and loosened (shown by arrow). (*g*, *h*) Final stages of deterioration. Note fungal hyphae in spiral around fiber *g*. (*From A. N. J. Heyn, Textile Ind., vol. 120, 1956*)

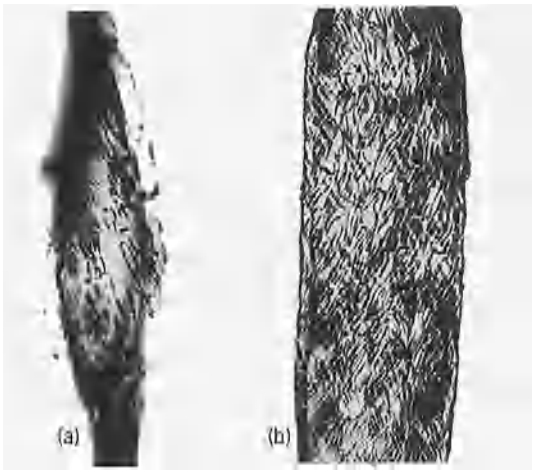


Fig. 2. Two growth stages of cellulose bacteria (*Sporocytophaga myxococcoides*) on cotton fiber. (*a*) Initial stage. (*b*) Final stage. (*From A. N. J. Heyn, Textile Res. J., 27:591–602, 1957*)

cotton attacked by microbes has a higher degree of swelling than normal cotton. Other testing methods determine the pH of an aqueous extract of the cotton and the fluorescence of the cotton fibers in ultraviolet light.

Wool. The number of bacteria and molds on raw wool has been reported as 1.2 million per gram and it may increase to 65 million in wet, scoured wool. *Achromobacter liquefaciens*, a nonsporeformer, and various sporeforming bacteria, such as *Bacillus subtilis*, cause the greatest damage to wool. In the degradation process the fiber scales are broken down and hydrolysis of the intercellular substances of the cortex takes place. If woolen fabrics are heated to 140°F (60°C), putrefaction caused by growth of *B. subtilis* and *Agarbacterium mesentericum* may develop and produce an uneven dyeability.

Fungi may develop on stored wool under humid conditions and when degradation products of fibrous proteins are present. Microbial damage of wool can be tested, for example, by the brilliant blue fluorescence of the degreased sample under ultraviolet light, and by the absence of partial decomposition of the scales.

Retting. This is a microbial process used for liberating fiber bundles from the stalks of fiber plants. In principle, retting consists of a breaking down of pectic substances between the cell walls (middle lamellae) of the individual cells of the tissue surrounding the bundles. As a result, the bundles become separated from the surrounding tissue and can then easily be extracted mechanically.

In water retting, the stalks are immersed in cold or warm, slowly renewed water, for from 4 days to several weeks. The water may be from rivers or constructed containers. The active organism is *Clostridium felsineum* and related types, which break down the pectin to a mixture of organic acids (chiefly acetic and butyric), alcohols (butanol, ethanol, and methanol), carbon dioxide (CO_2), and hydrogen (H_2). Its optimum activity at 91.4°F (33°C) is the basis for the warm-water retting process. Related species may be active in cold-water retting. In dew retting the stems are spread out in moist meadows; here the pectin decomposition is accomplished by molds and aerobic bacteria with the formation of CO_2 and H_2O. A retting process has been found to cause the liberation of fibers from the fruit husk of the coconut during soaking. *See* INDUSTRIAL MICROBIOLOGY; TEXTILE CHEMISTRY.

A. N. J. Heyn

Textile printing

The localized application of color on fabrics. In printing textiles, a thick paste of dye or pigment is applied to the fabric by appropriate mechanical means to form a design. The color is then fixed or transferred from the paste to the fiber itself, maintaining the sharpness and integrity of the design. In a multicolor design, each color must be applied separately and in proper position relative to all other colors. Printing is one of the most complex of all textile operations, because of the number of variables and the need for a high degree of precision, particularly since there is no way to correct a bad print. *See* DYE; DYEING.

Methods. A design may be applied in three major ways: raising the design in relief on a flat surface (block printing); cutting the design below a flat surface (intaglio or engraved printing); and cutting the design through a flat metal or paper sheet (stencil or screen printing). All three methods have been used for hand printing and reciprocating printing machines. In addition, these methods have been converted into rotary action by replacing blocks or plates with cylinders.

Block printing still exists to a very limited degree, and cylinders with a raised design are still occasionally used in wallpaper printing. However, this method is rarely found in textile applications, except in printing carpets.

The engraved plate has been replaced by the engraved cylinder, or roller. The design is etched into copper rollers, which are then generally chromium-plated. This method is still in wide use, particularly when large quantities of a relatively simple design are required.

The stencil was first developed into the flatbed screen, in which a silk gauze is stretched over a frame. Parts of the screen are blocked out, where no color is to penetrate, and paste is forced through the open parts of the screen, either by hand or mechanically, by a moving squeegee. The silk gauze was later replaced by a fine metal mesh that permitted this screen to be rolled into a cylinder, so that the mechanical action could be rotary instead of reciprocating. This resulted in rotary screen printing, actually a variation of flatbed printing.

Another method of printing utilizes individual computer-controlled nozzles for each color. The nozzles are used to paint a design on the fabric.

Paste and thickener. Each printing method requires a paste with special characteristics, frequently referred to as flow characteristics. For instance, a roller

printing operation requires a paste that tends to flow from the engraved portions of the roll onto the fabric; screen printing operations require a paste that flows only under the direct pressure of the squeegee. For this reason, the choice of thickener is dependent not only on the type of dye-stuff, but on the type of printing machine on which the printing is to be done, and frequently also on the type of fixation to be utilized.

Most natural thickening agents are based on combinations of starch and gums such as guar, locust, or alginate, chemically modified so as to be soluble in cold water. The synthetic thickening agents used are generally extremely high-molecular-weight polymers capable of developing a very high viscosity at a relatively low concentration.

In a few instances, thickening is accomplished by forming a thick emulsion, either oil in water, or water in oil. This last method is falling into disfavor because of the relatively large amount of solvent evaporated into the air during drying.

Sequence of operations. The first step is the preparation of print paste, which is made by dissolving the dyestuff and combining it with a solution of the appropriate thickening agent. Humectants and other materials necessary to aid in dye-stuff fixation are added at this point, as well as other auxiliaries, such as defoamers, which may be necessary to aid in printing. The fabric is then printed by any of the standard methods. The goods are then dried in order to retain a sharp printed mark and to facilitate handling between printing and subsequent processing operations.

The next operation, steaming, may be likened to a dyeing operation. Before steaming, the bulk of the dyestuff is held in a dried film of thickening agent. During the steaming operation, the printed areas absorb moisture and form a very concentrated dyebath, from which dyeing of the fiber takes place. The thickening agent prevents the dyestuff from spreading outside the area originally printed, because the printed areas act as a concentrated dyebath that exists more in the form of a gel than a solution and restricts any tendency to bleed. Excessive moisture can cause bleeding, and insufficient moisture can prevent proper dyestuff fixation. Steaming is generally done with atmospheric steam, although certain types of dyestuffs, such as disperse dyes, can be fixed by using superheated steam or even dry heat. In a few instances, acetic or formic acid is added to the steam to provide the acid atmosphere necessary to fix certain classes of dyes.

Flash aging is a special fixation technique used for vat dyes. The dyes are printed in the insoluble oxidized state by using a thickener which is very insoluble in alkali. The dried print is run through a bath containing alkali and reducing agent, and then directly into a steamer, where reduction and color transfer take place.

Printed goods are generally washed thoroughly to remove thickening agent, chemicals, and unfixed dyestuff. This washing must be carefully done to prevent staining of the uncolored portions of the fabric. Drying of the washed goods is the final operation of printing.

Special types of printing. There are several special printing processes.

Pigment printing. This process is actually one of the most widely used processes in textile printing. Instead of a dyestuff which can actually penetrate the fiber, an insoluble pigment is used as a coloring agent. A resinous binder added to the print paste serves to hold the pigment in place on the finished goods. The print is fixed by a simple dry heat application, and normally no washing is necessary after drying. This is generally the lowest-cost printing process, but the color is frequently less permanent, and the fabric generally tends to have a harsh feel.

Discharge printing. In this process, a color-destroying agent such as sodium hydrosulfite is printed onto dyed fabric, resulting in a white design appearing on a colored background.

Resist printing. This process is somewhat the reverse of discharge printing. The goods are printed with a material which will resist dyestuff. After printing, the goods are immersed in a dyebath, leaving the printed areas undyed. *See* TEXTILE CHEMISTRY.

David H. Abrahams

Bibliography. J. Fish, *Designing And Printing Textiles*, 2005; L. W. C. Miles, *Textile Printing*, 2d rev. ed., 2003; T. L. Vigo, *Textile Processing and Properties*, 1994; D. R. Waring and G. Hallas (eds.), *The Chemistry and Application of Dyes*, 1989; G. Wasike Namwamba, *Digital Textile Printing*, 2005.

Thaliacea

A small class of pelagic Tunicata especially abundant in warmer seas. This class of animals contains three orders: the Salpida, Doliolida, and Pyrosomida. Oral and atrial apertures occur at opposite ends of the body. Members of the orders Salpida and Doliolida are transparent forms, partly or wholly ringed by muscular bands (**Fig. 1**). The contractions of these

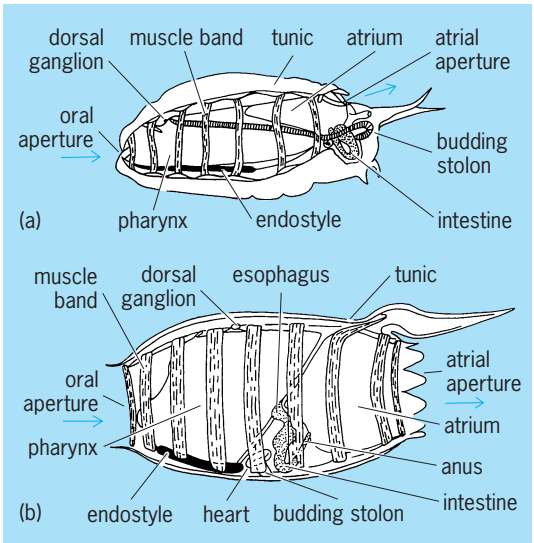


Fig. 1. Representatives of two orders of the class Thaliacea. (*a*) Salp, *Thalia democratica* (Salpida), solitary asexual form. (*b*) Doliolid, *Doliolum* (Doliolida), solitary asexual form. Arrows show direction of current.

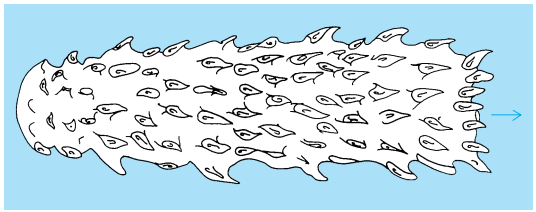


Fig. 2. Colony of *Pyrosoma atlanticum* (Pyrosomida).

bands produce currents used in propulsion, feeding, and respiration. The life cycle involves an alternation between solitary, asexual oozooids, which reproduce by budding from a complex stolon, and colonial, sexually reproducing blastozooids. The order Pyrosomida includes species which form tubular swimming colonies (**Fig. 2**) and which are often highly luminescent. *Salpa*, *Doliolum*, and *Pyrosoma* are familiar genera. *See* BIOLUMINESCENCE; TUNICATA (UROCHORDATA).

Donald P. Abbott

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; H. Thompson, *Pelagic Tunicates of Australia*, Commonwealth Council for Scientific and Industrial Research, Australia, 1948.

Thallium

A chemical element, TI, atomic number 81, relative atomic weight of 204.37. The valence electron notation corresponding to its ground state term is $6s^2 6p^1$, which accounts for the maximum oxidation state of III in its compounds. Compounds of oxidation state I and apparent oxidation state II are also known. *See* PERIODIC TABLE.

1												13	14	15	16	17	18
1 H	2											5 B	6 C	7 N	8 O	9 F	2 He
3 Li	4 Be											13 Al	14 Si	15 P	16 S	17 Cl	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12						18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Thallium occurs in the Earth's crust to the extent of 0.00006%, mainly as a minor constituent in iron, copper, sulfide, and selenide ores. Minerals of thallium are considered rare. Thallium compounds are extremely toxic to humans and other forms of life.

The insolubility of thallium(I) chloride, bromide, and iodide permits their preparation by direct precipitation from aqueous solution; the fluoride, on the other hand, is water-soluble. Thallium(I) chloride resembles silver chloride in its photosensitivity.

Thallium(I) oxide is a black powder which reacts with water to give a solution from which yellow thallium hydroxide can be crystallized. The hydroxide is a strong base and will take up carbon dioxide from the atmosphere.

Thallium also forms organometallic compounds of the following general classes, R_3TI, R_2TIX, and $RTIX_2$, where R may be an alkyl or aryl group and X a halogen. *See* ORGANOMETALLIC COMPOUND.

Edwin M. Larsen

Bibliography. P. W. Atkins et al., *Inorganic Chemistry*, 4th ed., 2006; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004.

Thallobionta

One of the two commonly recognized subkingdoms of plants. In contrast to the more closely knit subkingdom Embryobionta, the Thallobionta (also called Thallophyta) are diverse in pigmentation, food reserves, cell-wall structure, and flagellar structure. They still form a natural group, however, in that they are all probably derived from ancestors which would be referred to the Thallobionta, without the intervention of any ancestors which would have to be referred to other groups. The Thallobionta are here considered to include seven divisions, the Rhodophyta, Chlorophyta, Euglenophyta, Pyrrophyta, Chrysophyta, Phaeophyta, and Fungi. All these groups have both modern and fossil representatives.

The Thallobionta are united more by the absence of certain specialized tissues or organs than by positive resemblances. They do not have the multicellular sex organs commonly found in most divisions of Embryobionta. Gametes, when produced at all, are borne in unicellular gametangia which may or may not be differentiated into oogonia (which bear eggs) and antheridia (which bear sperms). Many of the Thallobionta are unicellular, and those which are multicellular seldom have much differentiation of tissues. None of them has tissues comparable to the xylem (a water-conducting tissue) found in most divisions of the Embryobionta, and only some of the brown algae (division Phaeophyta) have tissues comparable to the phloem (a food-conducting tissue) found in most divisions of the Embryobionta.

A large proportion of the Thallobionta are aquatic, and those which grow on dry land seldom reach appreciable size. The most obvious exceptions to this generality are provided by some of the fungi, such as mushrooms and puffballs, and even here the visible fruiting body is merely the aerial reproductive structure of a plant which has a much more diffuse mycelium (mass of filamentous vegetative strands) permeating the substrate. The Thallobionta thus consist of all those plants which have not developed the special features that mark the progressive adaptation of the Embryobionta to life on dry land. *See* CHLOROPHYCOTA; CHRYSOPHYCEAE; EMBRYOBIONTA; EUGLENOPHYCEAE; FUNGI; PHAEOPHYCEAE; PLANT KINGDOM; RHODOPHYCEAE.

Arthur Cronquist

Bibliography. A. Cronquist, *Basic Botany*, 1973.

Theales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Dilleniidae of the class Magnoliopsida (dicotyledons). The order consists of 18 families and nearly 3500 species. The largest families are the Clusiaceae, sometimes called Guttiferae (about 1200 species), Theaceae (about 600 species), Dipterocarpaceae (about 600 species), and Ochnaceae (about 400 species). They are mostly woody plants, less often herbaceous, with simple or occasionally compound leaves. The perianth and stamens of the flowers are attached directly to the receptacle; the calyx is arranged in a tight spiral; and the petals are usually separate from each other (see **illus.**). The stamens are numerous and initiated

Franklin tree (*Franklinia alatamaha*) flowers, a characteristic member of the family Theaceae in the order Theales, named in honor of Benjamin Franklin. A rare native of the southeastern United States, it has not been seen in the wild since 1790, but it is frequently planted in the United States and Europe. (*Photograph by A. W. Ambler, from National Audubon Society*)

in centrifugal sequence or, less often, are few and cyclic; the pollen is nearly always binucleate. The gynoecium consists of two to many, more or less united carpels. The tea plant (*Thea sinensis*), *Camellia* (in the Theaceae), and St.-John's-wort (*Hypericum*, in the Clusiaceae) are familiar members of the Theales. *See* DILLENIIDAE; FLOWER; MAGNOLIOPHYTA; PLANT KINGDOM; TEA. Arthur Cronquist; T. M. Barkley

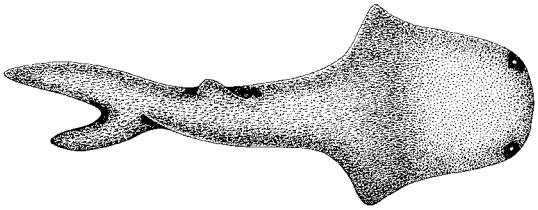
Thecanephria

An order of Pogonophora, a group of elongate, tentaculated, tube-dwelling, sedentary, nonparasitic marine worms lacking a digestive system. In this order the "coelomic" space in the anterior tentacular region is horseshoe-shaped, and the excretory (osmoregulatory) portion of its ducts come close together medially near the median, adneural blood vessel. The spermatophores are generally broad and leaf-shaped. The tube is often rigid anteriorly. The species in this order are usually large (0.04–0.12 in. or 1–3 mm in diameter) and multitentaculate.

The order includes four families: Polybrachiidae (with seven genera), Sclerolinidae (one genus), Lamellisabellidae (two genera), and Spirobrachiidae (one genus). *See* POGONOPHORA. Edward B. Cutler

Thelodontida

Sometimes called Coelolepida, this is an extinct group of Agnatha or jawless vertebrates known from the Lower Silurian to Middle Devonian of Europe, Asia, Australia, and North America. Since they had no hard skeleton, their structure and relationships are poorly known. Well-preserved specimens of *Logania* (see **illus.**) suggest a relationship to Heterostraci



***Logania scotica*, dorsal, showing the eyes and dorsal fin. The tail, in profile, shows the caudal fin. (*After R. H. Traquair, Geological survey in the Upper Silurian rocks of Scotland, Trans. Roy. Soc. Edinburgh, vol. 40, 1905*)**

in their widely spaced, lateral eyes, nearly terminal mouth, pectoral flaps at the posterior end of the head, and hypocercal tail with a downwardly directed main lobe. *Phlebolepis* resembles Anaspida in general form but lacks the dorsal nostril of that group. *See* JAWLESS VERTEBRATES; HETEROSTRACI; OSTRACODERM; PTERASPIDOMORPHA. Robert H. Denison

Theorem

A proposition arrived at by the methods of logical deduction from a set of basic postulates or axioms accepted as primitive and therefore not subject to deductive proof. So long as a theorem is part of a purely formal system, it is not meaningful to speak about the "truth" of a theorem but only about its "correctness." It becomes true when it, or its consequences, can be shown to be in accord with observable facts. The classic example of a system of theorems is afforded by Euclid's system of geometry, which is now recognized to have only a purely formal character, although it was formerly considered to be meaningful to ask whether Euclid's geometry was "true." *See* LOGIC. Percy W. Bridgman; Henry Margenau

Theoretical ecology

The use of verbal models, analytical models, or simulation models to explain patterns, suggest experiments, or make predictions in ecology. Because ecological systems are idiosyncratic, extremely complex, and variable, ecological theory faces special

challenges. Unlike physics or genetics, which use fundamental laws of gravity or of inheritance, ecology has no widely accepted first-principle laws. Instead, different theories must be invoked for different questions, and the theoretical approaches are enormously varied. Indeed, a central problem in ecological theory is determining what type of model to use and what to leave out of a model. The traditional approaches have relied on analytical models based on differential or difference equations; but recently the use of computer simulation has greatly increased with advances in computational power and ease of use. *See* ECOLOGICAL MODELING; ECOLOGY; ECOLOGY, APPLIED; SIMULATION.

Levels of ecological theory. The nature of ecological theory varies depending on the level of ecological organization on which the theory focuses. The primary levels of ecological organization are (1) physiological and biomechanical, (2) evolutionary (especially applied to behavior), (3) population, and (4) community.

Physiological and biomechanical theory. At the physiological and biomechanical level, the goals of ecological theory are to understand why particular structures are present and how they work. The approaches of fluid dynamics and even civil engineering have been applied to understanding the structures of organisms, ranging from structures that allow marine organisms to feed, to physical constraints on the stems of plants. A simple example of this type of analysis is the demonstration that there are lower limits to the body size of a mammal. The ratio of surface area to volume increases as an animal gets smaller. If the animal is too small, it cannot maintain a constant temperature. This argument can be made quantitative, and the predictions of the smallest body size agree well with the size distribution of existing mammal species.

Behavioral and evolutionary theory. At the behavioral level, the goals of ecological theory are to explain and predict the different choices that individual organisms make. Underlying much of this theory is an assumption of optimality: the theories assume that evolution produces an optimal behavior, and they attempt to determine the characteristics of the optimal behavior so it can be compared with observed behavior. One area with well-developed theory is foraging behavior (where and how animals choose to feed). One basic formulation assumes that resources (such as seeds for birds or flowers for pollinators) are present in different patches, and it seeks to understand how an animal should divide its searching time among patches. In the simplest case, an animal should leave any patch when the rate of return (rate of food consumption) in the patch is less than the mean rate of return over all patches. This prediction has been upheld by observations of many bird species. This theory is limited, however, because it ignores any interactions among the foraging individuals. If the fitness consequences of a behavioral action depend on which conspecifics (individuals or populations of a given species) an animal interacts with, a more sophisticated approach is needed. *See* BEHAVIORAL ECOLOGY.

In the 1970s, John Maynard Smith introduced the use of game theory to understand the evolution of behaviors that are apparently not optimal for an individual but may instead be better for a group—say, ritualized fighting. For example, in some African ungulates, one animal will back down when a pair engages in a fight. At first sight, this appears to contradict the principle of optimality at the level of the individual—the animal that gives up (a "dove") would seem to clearly have some chance of winning the contest. Conversely, an individual that fights until it dies or wins (a "hawk") is clearly worse off than a dove on the occasions that it encounters another hawk and dies in battle. Thus it can be shown that in a population of all hawks, doves are favored because they do not suffer from having to fight and win. Conversely, in a population of all doves, hawks are favored since they win every contest. Thus, evolution should lead to a population in which some individuals are hawks and some are doves. Another possibility is that individuals are sometimes hawks and doves, as the model does not distinguish who plays each strategy, but focuses only on the fraction of time the strategy is used. Note that the optimal situation for the population as a whole (which would be all doves and thus no battle casualties) is not the solution that is predicted to be reached by evolution (which should always include some hawks, since hawks are favored when they are rare). *See* OPTIMIZATION.

Determining the precise outcome in this example involves the use of concepts from game theory (originally developed by J. von Neumann and J. Nash to look at questions from economics) to find what has been called the evolutionary stable strategy. In this example, the evolutionary stable strategy is the proportion of hawks and doves at which the fitnesses of both are equal and any increase in the frequency of hawks reduces the fitness of hawks and any increase in the frequency of doves reduces the fitness of doves. Expressing the result this way clearly shows that frequency dependence drives the outcome. This approach is powerful and has been used to understand a wide range of ecologically important behaviors, from seed dispersal to the production of alarm calls. *See* GAME THEORY.

Population theory for single species. The population level has the longest history of ecological theory and perhaps the broadest application. The simplest models of single-species populations ignore differences among individuals and assume that the birth rates and death rates are proportional to the number of individuals in the population. If this is the case, the rate of growth is exponential (like compound interest), a result that goes back at least as far as Malthus's work in the 1700s. As Malthus recognized, this result produces a dilemma: exponential growth cannot continue unabated. Thus, one of the central goals of population ecology theory is to determine the forces and ecological factors that prevent exponential growth and to understand the consequences for the dynamics of ecological populations. *See* ECOLOGICAL METHODS; POPULATION ECOLOGY.

Density dependence. For a single species, intraspecific effects on the rate of growth are called density dependence. One primary density-dependent factor is food supply limitation; as the density of a species goes up, it increases competition for limited resources, causing the birth rate or survival rate to go down. Other density-dependent factors include the number of territories or nesting sites, incidence of cannibalism, and availability of space, light, or nutrients for plants. If density dependence acts without delay, the populations will tend toward an equilibrium, explaining the persistence of species at a stable level. This model is usually expressed in mathematical terms with a differential equation called the logistic equation (1),

$$dN/dt = rN(1 - N/K) \tag{1}$$

where N is the size of the population in numbers of biomass at time t, r is the intrinsic rate of increase of the population, and K is the carrying capacity. The population approaches the stable equilibrium, $N = K$. *See* MATHEMATICAL ECOLOGY.

This simple theoretical approach has many extensions. A simple modification of the logistic model theorizes that for many species, such as small mammals like mice or voles and forest insects, density dependence acts in a delayed fashion and can produce cycles. Delayed density dependence would arise if a population consumes its food supply, which would then reduce the food available for the next generation. Cannibalism would have the same dynamical effect.

If the population is assumed to have discrete, nonoverlapping generations, the resulting dynamics can be much more complex, as was first elucidated by Robert May. One model of this form is called the Ricker equation (2), where N_t is the population size

$$N_{t+1} = N_t e^{r[t - (N_t/k)]} \tag{2}$$

in generation t, r is the intrinsic rate of increase for the population, and K is the carrying capacity. The exact form of the equation is not important, but the property of overcompensation is—the total number of individuals the following year goes down at high population levels, not just the number per capita (the number next year divided by the number this year). In this case, not only can cycles result, but also a pattern of population change called chaos.

Chaos. The study of chaos has been a recent focus of theoretical ecology. A system is defined as chaotic if it exhibits sensitive density dependence on initial conditions for all possible initial conditions, and differences in initial conditions grow exponentially. Theoretical models have clearly shown that chaos is a plausible possibility for natural populations. Several empirical studies have attempted to determine if chaos is in fact present by using statistical techniques to look at time series from a variety of natural populations. The results of these empirical studies are still being debated. *See* CHAOS.

Maximum sustainable yield. Modifications and extensions of the theoretical approaches like the logistic model have also been used to guide the management of renewable natural resources. Here, the most basic concept is that of the maximum sustainable yield, which is the greatest level of harvest at which a population can continue to persist. In simple models, this is found by maximizing the growth rate of the populations as a function of the population level. The growth rate of the population at this level is then the maximum sustainable yield. In the logistic model, the maximum sustainable yield occurs when the population is at half the carrying capacity. Not surprisingly, allowing for random environmental fluctuations in population levels greatly changes the determination of maximum sustainable yield.

Stochasticity. Inclusion of random population fluctuations is important in many theories of population dynamics, especially in applied areas. Ecologists distinguish demographic stochasticity (fluctuations due to the randomness of births and deaths in small populations) from environmental stochasticity (fluctuations due to environmental randomness such as weather). Once stochasticity is included, all populations will eventually go extinct if there is a ceiling to the population level, so the focus changes from looking for a population equilibrium to calculating times to extinction or probability of extinction over a fixed time. Using the concept of minimum viable population size, conservation biologists have emphasized the role of stochasticity in predicting the probability that a population will persist over a fixed time. *See* STOCHASTIC PROCESS.

Metapopulation. Another important theoretical approach to modeling the dynamics of a single species, tracing back to the work of Richard Levins, focuses on the concept of a metapopulation. Ecologists have observed that many species have a population structure in which there are relatively small subpopulations, each of which can go extinct, and new subpopulations are founded by colonizers from other subpopulations. The simplest model describing this situation would focus only on the proportion of habitat or patches that are occupied by the species of interest. The rate of change for this fraction of occupied patches is given by the balance between the per-patch extinction rate and the colonization rate. This model has the same mathematical form as the logistic equation and thus predicts (if the colonization rate is high enough relative to the extinction rate) that the fraction of occupied habitat approaches an equilibrium. The metapopulation approach has been used to understand the effects of human-caused habitat loss and fragmentation on ecological species and communities.

Population dynamics of interacting species. Interactions between two species have been very fruitful areas of theoretical ecology, as ecologists attempt to understand communities by first considering the simplest communities—those with two species. In the study of these interactions, the concept of stability is paramount. The basis behind many of the investigations is that natural communities correspond to stable equilibria of the models, so the conditions for stable equilibria to exist should be determined. An equilibrium is stable if solutions that start near

the equilibrium approach the equilibrium. A simple analogy would be a marble in a bowl: even though you roll the marble around in the bowl, it will always approach the stable equilibrium, which is the marble at the bottom of the bowl. Although this reliance on stable equilibria has been called into question in recent years, it still is a fruitful way to proceed in many cases.

The interactions between two species are often classified as competition (two species that share common resources), predation (one species eating another), or mutualism (two species that help each other). Additional categories are host-parasitoid interactions (a specialized predatory-prey interaction among insects in which the parasitoid lays its eggs in the larvae of the host) and disease or parasitic interactions.

Competition. Competition theory, beginning with the classic work of A. J. Lotka, V. Volterra, and G. F. Gause in the 1920s and 1930s, is perhaps the most influential ecological theory. The central theme of this work is that coexistence between competitors requires differences between the two species. More recent work suggests that coexistence of plants requires that they have different limiting nutrients. *See* PLANT-ANIMAL INTERACTIONS.

Predator-prey. The study of interaction between predator and prey also goes back to the work of Lotka, Volterra, and Gause. Here, a central finding is that this kind of interaction can produce cycles in population numbers because there is essentially a form of delayed density dependence—increases in prey numbers lead to increases in predator numbers, which reduces the number of prey, which leads to a reduction in the number of predators, which causes prey populations to rebound (and so on).

Diseases. Diseases are another kind of interaction between species. Theory describing the dynamics of diseases was developed in the 1920s, when biologists and mathematicians attempted to predict and describe how diseases such as malaria spread. A disease will increase in prevalence only if its reproductive number (the number of new infections caused by a single individual) is greater than one. This number depends on the probability of transmission of the disease and the size of the host population. Using this concept, the fraction of a population that needs to be immunized to cause a disease to die out (as with polio or smallpox) can be determined. Because good records are available for childhood diseases such as measles, disease dynamics has played a central role in recent attempts to make quantitative matches between ecological theory and natural systems. *See* EPIDEMIOLOGY.

Community theory. The primary goal of ecological theory at the community level is to understand diversity at local and regional scales. Recent work has emphasized that a great deal of diversity in communities may depend on trade-offs. For example, a trade-off between competitive prowess and colonization ability is capable of explaining why so many plants persist in North American prairies. Plant species that are competitively superior can hold onto a piece of land they currently occupy, but when they die, species that are good colonizers (but poor competitors) arrive early and may persist for long enough that they reproduce before being displaced. Thus instead of only one competitively superior species persisting in a prairie, a variety of species can find opportunities for reproduction in prairies because of the trade-off between competitive prowess and dispersal. Similarly, trade-off between competitive prowess and resistance to grazing or physical disturbance may explain the diversity of algae in the rocky intertidal shoreline communities. Some species thrive as colonizers of newly open intertidal space following a storm, whereas other species owe their success to their ability to displace those early colonizers. Diversity arising from such trade-offs typically produces a patchwork of species, with each species having lived in its current position for different periods of time.

Another major concept in community theory is the role of disturbance. Understanding how disturbances (such as fires, hurricanes, or wind storms) impact communities is crucial because humans typically alter disturbance. Humans may think that they are benefiting nature by stopping wildfires, when in fact the opposite is true—the fires could be essential for maintaining diversity. *See* BIODIVERSITY; ECOLOGICAL COMMUNITIES.

A related issue at the community level is to understand the effects of changes at one trophic level on numbers at another trophic level. The central theme here is the outgrowth of a question raised in the 1960s—why is the Earth "green" (with plants and vegetation)? Presumably, consumers of plants should increase in number until they reduce their food supply. One proposed solution to this paradox is that predators hold the number of herbivores low enough that there remains a large supply of green plant material. If this scenario is correct, the addition of nutrients will lead to increased plant growth, which would be reflected by increases in the number of herbivore predators. Conversely, removal of top predators that feed on the predators of herbivores should lead to a decrease in the herbivores and subsequent increase in the density of plants—an explanation for the increase in phytoplankton in overfished lakes. This notion of trophic cascades is a central theme in the study of food webs. *See* FOOD WEB; TROPHIC ECOLOGY.

Alan Hastings

Bibliography. T. Case, *An Illustrated Guide to Theoretical Ecology*, Oxford University Press, New York, 2000; A. Hastings, *Population Ecology: Concepts and Models*, Springer-Verlag, New York, 1997; R. M. May (ed.), *Theoretical Ecology*, Sinauer Associates, Sunderland, MA, 1981; J. Maynard Smith, *Evolution and the Theory of Games*, Cambridge University Press, 1982.

Theoretical physics

The description of natural phenomena in mathematical form. It is impossible to separate theoretical physics from experimental physics, since a complete

understanding of nature can be obtained only by the application of both theory and experiment. *See* PHYSICS.

Purposes. There are two main purposes of theoretical physics: the discovery of the fundamental laws of nature and the derivation of conclusions from these fundamental laws.

Discovery of fundamental laws. Physicists aim to reduce the number of laws to a minimum to have as far as possible a unified theory. When the laws are known, it is possible from any given initial conditions of a physical system to derive the subsequent events in the system. Sometimes, especially in quantum theory, only the probability of various events can be predicted. *See* DETERMINISM.

Conclusions from fundamental laws. The conclusions to be derived from the fundamental laws of nature may be of several different types.

1. Conclusions may be derived in order to test a given theory, particularly a new theory. An example is the derivation of the spectrum of the hydrogen atom from quantum mechanics; the verification of the predictions by accurate measurements is a good test of quantum mechanics. On rather rare occasions experiment has been found to contradict the predictions of an existing theory, and this has then led to the discovery of important new physical laws. An example is the Michelson-Morley experiment on the constancy of the velocity of light, an experiment which led to special relativity theory. *See* ATOMIC STRUCTURE AND SPECTRA; LIGHT; RELATIVITY.

2. Theory may be required for experiments designed to determine physical constants. Most fundamental physical constants cannot be accurately measured directly. Elaborate theories may be required to deduce the constant from indirect experiments. An example is the Millikan oil-drop determination of the electron charge, which requires the knowledge of the motion of small droplets in air as deduced from hydrodynamic theory. *See* FUNDAMENTAL CONSTANTS.

Another example is the precise determination of the value of the magnetic moment of the electron by Hans Dehmelt, to an accuracy of 1 part in 3×10^{10}. By using an elaborate theory, this can be used (1) to establish the correctness of quantum electrodynamics and (2) to determine the value of the fine structure constant α to be given by Eq. (1), where h is Planck's

$$\frac{1}{\alpha} = \frac{hc}{2\pi e^2} = 137.035991 \qquad (1)$$

constant, c is the speed of light, and e is the charge of the electron. *See* QUANTUM ELECTRODYNAMICS.

3. Predictions of physical phenomena may be made in order to gain understanding of the structure of the physical world. In this category fall theories of the structure of the atom leading to an understanding of the periodic system of elements, or of the structure of the nucleus in which various models are tested (for example, shell model or collective model). In the same category fall applications of theoretical physics to other sciences, for example, to chemistry (theory of the chemical bond and of the rate of chemical reactions), astronomy (theory of planetary motion, internal constitution, and energy production of stars), or biology.

4. Engineering applications may be drawn from fundamental laws. All of engineering may be considered an application of physics, and much of it is an application of mathematical physics, such as elasticity theory, aerodynamics, electricity, and magnetism. The generation and propagation of radio waves of all frequencies are examples of application of theoretical physics to direct practice. *See* AERODYNAMICS; ELASTICITY; ELECTRICITY; MAGNETISM; RADIO-WAVE PROPAGATION.

Content. Apart from the classification of the fields of theoretical physics according to purpose, a classification can also be made according to content. Here one may perhaps distinguish three classification principles: type of force, scale of physical phenomena, and type of phenomena.

Type of force. At present four different types of force are known to physics. The best-understood type of force is electricity and magnetism. Here the fundamental laws, Maxwell's equations, are completely known. Corrections due to quantum theory exist but can be calculated. For practical purposes electromagnetic fields can be calculated with confidence and precision, from dc fields to the shortest gamma-ray wavelength. *See* ELECTROMAGNETIC RADIATION; FORCE; MAXWELL'S EQUATIONS.

The second type of force, known for the longest time, is the gravitational force. For practical purposes, Newton's inverse-square law is usually sufficient. The most complete theory of gravitation, however, is Einstein's general theory of relativity, which has great beauty and has now been well established by the propagation of radio waves from satellites on paths going close to the surface of the Sun. *See* GRAVITATION.

The weak force of nuclear physics is responsible for radioactive beta decay. Well over 100 nuclei are known to undergo beta decay. The process is essentially

$$\text{neutron} \rightarrow \text{proton} + \text{electron} + \text{neutrino}$$

The theory was given by Enrico Fermi in 1934, and fits observation. However, quantum theory requires that the theory be renormalizable, and a theory in which one particle turns into three is not renormalizable. It was therefore postulated that the process goes in two steps:

$$\text{neutron} \rightarrow \text{proton} + W^-$$
$$W^- \rightarrow \text{electron} + \text{neutrino}$$

The W^- particle has been discovered experimentally and has a mass about 80 times that of a proton. This large mass means that the W^- is not actually formed in beta decay but is only "virtual" in the sense of quantum theory. *See* INTERMEDIATE VECTOR BOSON; RADIOACTIVITY; RENORMALIZATION; WEAK NUCLEAR INTERACTIONS.

There is also a W^+ particle, and a neutral counterpart, Z^0, of mass about 100 protons. The mass

difference between W and Z is of theoretical interest. Steven Weinberg, Sheldon L. Glashow, and Abdus Salam predicted the existence of the W and Z particles and showed that they are related to the electromagnetic force, even though in most beta decays the weak force has a strength of only about 10^{-20} times the electromagnetic one. *See* ELECTROWEAK INTERACTION; STANDARD MODEL.

The last force is the strong force which holds atomic nuclei together. According to experiment, each nucleon contains three subunits called quarks, having charges of $+2/3\ e$ (up quarks) and $-1/3\ e$ (down quarks), where e is the charge of the proton. The quarks interact by interchanging gluons. Neither quarks nor gluons exist outside the nucleon. In addition to charge, they carry another property, called color. The theory of their interaction is called quantum chromodynamics (QCD). The theory is quite complicated, but its coupling constant has been determined [Eq. (2)] and is much bigger than the fine

$$\frac{2\pi g^2}{hc} = 0.117 \tag{2}$$

structure constant α [Eq. (1)]. *See* COLOR (QUANTUM MECHANICS); GLUONS; QUANTUM CHROMODYNAMICS; QUARKS.

Calculation of nuclear properties from quantum chromodynamics requires elaborate computer programs, and only a beginning has been made. The size of the nucleon has been determined and is in accord with experiment.

For calculation of nuclear properties, a semiempirical approach has been used. The two-nucleon interaction has many components whose magnitude is derived from two-nucleon scattering experiments. An additional interaction between three nucleons is needed to account for the binding energy of the triton, ^{3}H. With this, the binding of ^{4}He can be calculated and agrees with experiment, without invoking an additional force between four nucleons. The structure of nuclei up to ^{16}O has been calculated with this scheme. *See* FUNDAMENTAL INTERACTIONS; NUCLEAR STRUCTURE; STRONG NUCLEAR INTERACTIONS; TRITON.

Scale of physical phenomena. The motion of bodies on the scale of everyday life can be described by the classical mechanics of Isaac Newton. Phenomena in very small dimensions, especially inside atoms or atomic nuclei, must be described by quantum mechanics. The latter theory contains Newton's mechanics as a special case. *See* CLASSICAL MECHANICS.

The description of physical phenomena is also different according to the velocities of the bodies involved. When the velocity is a substantial fraction of that of light, the special theory of relativity must be used to describe the motion. (The special theory of relativity has hardly anything in common with general relativity theory except the name, and is established beyond doubt by an enormous number of experiments.) Newton's classical mechanics again is a special case of the mechanics of special relativity. *See* RELATIVISTIC ELECTRODYNAMICS; RELATIVISTIC MECHANICS.

Special relativity and quantum mechanics are examples of the development of physical theory. Neither of them has made classical mechanics wrong or obsolete, but they have extended classical mechanics into domains which were outside of human experience until 1900. When a physical law is discovered, it can be expected to hold as long as the general conditions are not radically changed from those holding in the experiments from which the law was originally derived; for example, classical mechanics holds for objects of not too small size moving with moderate velocities. Classical mechanics, not quantum theory, is still valid for microbes of 1 micrometer size. Relativity is negligible for the speed of a spacecraft. Going to still smaller sizes, no limit has been found as yet on the validity of quantum mechanics. In a completely new area (for example, where there is very small size or very high speed) it cannot be expected a priori that the same laws will continue to hold; but if the laws do change under the new conditions, this does not invalidate the old laws in the domain for which they were originally formulated, except for minor corrections. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

The most general theory of motion now known is quantum field theory, which combines both quantum mechanics and relativity theory and at the same time embodies the observed fact that particles can be created and annihilated. This theory may thus be called a unified field theory. Attempts have also been made toward other unification, in particular to unify the theories of two types of forces, gravitational and electromagnetic; this is commonly called unification field theory in the literature. In the 1930s and 1940s, these attempts were not successful. In the 1980s, by the addition of quantum field theory, a beginning toward unification may have been made. *See* GRAND UNIFICATION THEORIES; QUANTUM FIELD THEORY; SUPERGRAVITY.

Type of phenomena. The most customary classification of theoretical physics is according to the type of phenomena described. The following are the main fields under this heading:

1. Mechanics is the theory of motion of bodies under given forces. It is normally understood to involve classical mechanics only, and includes particle mechanics and mechanics of rigid bodies. In particle mechanics, celestial mechanics is an important subdivision; this includes planetary motion, the motion of artificial satellites, and the complicated motions resulting when three bodies interact (the classical three-body problem). The field of rigid-body mechanics includes the complicated theory of gyroscopic motion with and without external fields of force. *See* CELESTIAL MECHANICS; GYROSCOPE; MECHANICS; RIGID-BODY DYNAMICS.

2. Continuum mechanics is the theory of motion of bodies, taking into account their internal properties. One branch of this is the theory of elasticity, which is basic for structural engineering design. Another branch is hydro- and aerodynamics. Here a number of problems can be solved approximately by potential theory, but most of modern aerodynamics

requires a more physical approach. Knowledge of physical properties, such as those given by the equation of state of a gas, is essential; these properties can be explained only on a molecular scale. Acoustics is a classical branch of continuum mechanics. A combination of aerodynamics and electrodynamics is required for the modern field of magnetohydrodynamics. *See* ACOUSTICS; CLASSICAL FIELD THEORY; HYDRODYNAMICS; MAGNETOHYDRODYNAMICS; POTENTIALS.

3. Heat presents a problem that can be treated on a phenomenological level by thermodynamics, which is the basis of heat engineering, as well as of the theory of chemical equilibrium. On the molecular level, heat is described by statistical mechanics, which may be considered the physical foundation of thermodynamics. Beyond this, statistical mechanics permits the calculation of the properties of bulk substances (gases, liquids, and solids) in terms of their atomic properties. *See* HEAT; STATISTICAL MECHANICS; THERMODYNAMIC PRINCIPLES.

4. Electrodynamics is well understood. Subdivisions are electrostatics; the theory of stationary currents (the basis of electrical generating machinery); the theory of oscillating electrical circuits (the basis of the technology of ordinary radio); the theory of electromagnetic waves, including their propagation in air as well as in waveguides and similar devices (the basis of radar); and finally the electromagnetic theory of light. *See* ALTERNATING CURRENT; DIRECT CURRENT; ELECTRODYNAMICS; ELECTROMAGNETIC WAVE TRANSMISSION; ELECTROSTATICS.

5. Optics is customarily treated as a special field, although, strictly speaking, it is a branch of electrodynamics. Geometrical optics and the theory of diffraction phenomena are two of the principal topics. Emission and absorption of light can be understood only on the basis of atomic physics. The same is true of dispersion, that is, the behavior of the refractive index as a function of frequency. *See* OPTICS.

6. Atomic physics includes the theory of the structure of the atom; the motion of the electrons in the atom; the periodic system; the energy levels and spectral lines of atoms and molecules; the behavior of atoms and molecules in external fields; and collisions of atoms with each other, with electrons, and with other particles. Atomic physics is the basis of the calculation of properties of matter in bulk and of the emission and absorption of light. Related is the theory of molecular structure, which is the basis of theoretical chemistry. Collisions between molecules explain the rate of chemical reactions. *See* ATOMIC PHYSICS; ATOMIC STRUCTURE AND SPECTRA; MOLECULAR PHYSICS; MOLECULAR STRUCTURE AND SPECTRA.

7. Nuclear and particle physics includes the theory of nuclear forces and of the structure of atomic nuclei. A complete theory would predict all energy levels of any nucleus and thus the electromagnetic radiations which can be emitted by the nucleus. The topic also includes the theory of nuclear reactions, which is the basis of the technology of nuclear reactors. In an effort to understand the origin of nuclear forces, theoretical physicists have investigated the production and properties of mesons and the so-called strange particles. Radioactive decay, and particularly beta decay, is another branch of nuclear physics involving weak rather than strong nuclear forces. High-energy nuclear physics aims at understanding the properties of particles—nucleons as well as unstable particles of various kinds. *See* ELEMENTARY PARTICLE; MATHEMATICAL PHYSICS; NUCLEAR PHYSICS.

Hans A. Bethe

Bibliography. H. Frauenfelder and E. M. Henley, *Subatomic Physics*, 2d ed., Prentice Hall, 1991; L. D. Landau and E. M. Lifshitz, *A Shorter Course of Theoretical Physics*, 2 vols., 1992.

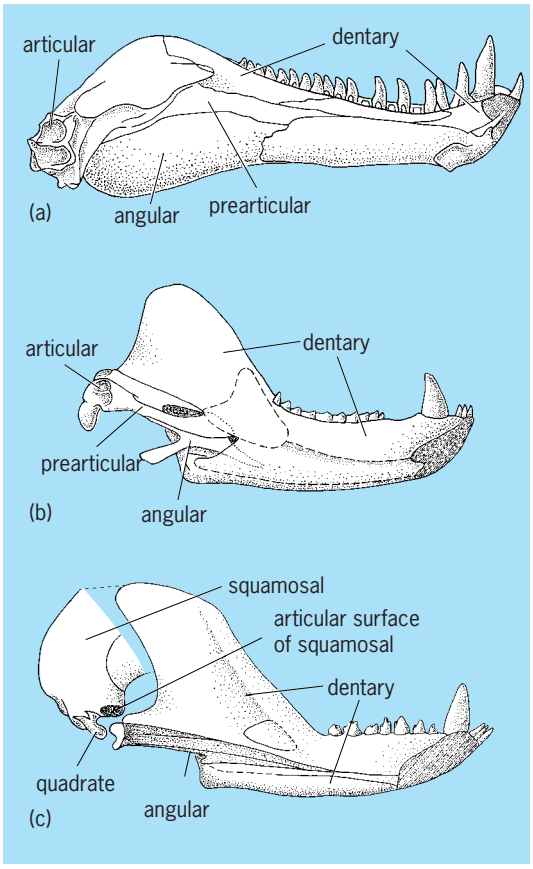
Therapsida

An order of Reptilia, subclass Synapsida, often called advanced mammallike reptiles, that flourished from the middle Permian through the Late Triassic. The group is highly diverse and subdivided into six suborders. Two of these, Eotitanosuchia and Dinocephalia, include relatively primitive mid-Permian carnivores and herbivores. A third, the Dicynodontia, made up of small to large herbivores, was abundant in the late Permian. Dicynodonts were associated with two carnivorous suborders, the Therocephalia and Gorgonopsia, which are morphologically intermediate between Eotitanosuchia and the cynodonts. Although these five developing lines are distinct, the skulls and skeletons in each became increasingly mammal-like.

The trend continued among the highly diverse Cynodontia. This suborder includes a variety of carnivores, omnivores, and herbivores. The most highly derived herbivorous cynodonts were the tritylodonts of the Late Triassic and Early Jurassic. They were very mammallike and, although now they are placed among the therapsids, when first found they were classified as mammals.

During the evolution of some lines of the carnivorous cynodonts, the medioposterior bones of the lower jaw, including the articular which formed part of the jaw joint, became successively reduced, accompanied by increase in the relative size of the tooth-bearing bone, the dentary (see **illus.**). This trend was most fully expressed among highly derived cynodonts of the infraorder Chinquodontoidea, including the family Trithelodontidae (=Ictidosauria). In the most specialized forms the skull–jaw joint is formed partly by the "old reptilian joint" between the articular and quadrate bones and partly by the "mammalian joint" between the dentary and squamosal bones. The quadrate and articular eventually became free of the skull and jaw and, along with the single reptilian ossicle, the stapes, went to form the three ossicles of the middle ear of mammals. Coincident in time with the existence of the chinquodontoids and the tritylodonts was the appearance of the first fossils classified as mammals.

The late Triassic Therapsida existed along with the rapidly expanding Archosauria, including the diversifying dinosaurs. The two groups for the most part



Therapsid lower jaws. (*a*) Sphenacodont pelycosaur (*after A. S. Romer and L. I. Price, Review of the Pelycosauria, Geol. Soc. Amer. Spec. Pap. 28, 1940*). (*b*) Intermediate cynodont therapsid (*after E. F. Allin, Evolution of the mammalian ear, J. Morphol., vol. 148, 1975*). (*c*) Cinquodontid therapsid [showing posterior part of the skull] (*after A. S. Romer, The Chañares (Argentine) Triassic Reptilian Fauna, Museum of Comparative Zoology, Harvard University, Breviora 344, 1970*)

inhabited different areas. Climatic changes during the Triassic rather than direct competition largely accounted for the decline of the therapsids and the rapid expansion of the Archosauria. Only the very mammallike therapsids, the herbivorous trity-lodonts, and the minute derived first mammals survived into the Early Jurassic. *See* ARCHOSAURIA; MAMMALIA; REPTILIA; SYNAPSIDA. Everett C. Olson

Bibliography. M. J. Burton, *Vertebrate Paleontology*, 1990; R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1987.

Theria

One of the four subclasses of the class Mammalia, including all living mammals except the monotremes. The Theria were by far the most successful of the several mammalian stocks that arose from the mammallike reptiles in the Triassic. The subclass is divided into three infraclasses: Pantotheria (no living survivors), Metatheria (marsupials), and Eutheria (placentals). These were not strictly contemporaneous; the Pantotheria arose directly from mammallike reptiles, and the Metatheria and Eutheria in turn arose from pantotherelike forms during the Jurassic or Cretaceous, many millions of years later. Therian mammals are characterized by the distinctive structural history of the molar teeth. The fossil record shows that all the extremely varied therian molar types were derived from a common tribosphenic type in which three main cusps, arranged in a triangle on the upper molar, are opposed to a reversed triangle and basinlike heel on the lower molar. *See* MAMMALIA. D. Dwight Davis; Frederick S. Szalay

Thermal analysis

A group of analytical techniques developed to continuously monitor physical or chemical changes of a sample which occur as the temperature of a sample is increased or decreased. Thermogravimetry, differential thermal analysis, and differential scanning calorimetry are the three principal thermoanalytical methods. Commercial instruments are available for all types of thermal analysis. Most instruments use modern computer technology to control the instrument while collecting and analyzing data. *See* ANALYTICAL CHEMISTRY; COMPUTER.

The occurrence of physical or chemical changes upon heating a sample may be explained from either a kinetic or thermodynamic viewpoint. Kinetically the rate of a process may be increased by raising the temperature as shown by the Arrhenius equation (1),

$$\text{Rate} = A\epsilon^{-E_a/RT} \tag{1}$$

where A, E_a, and R represent the preexponential factor, activation energy, and the gas law constant, respectively. At some point, the rate becomes significant and readily observable. Similarly, an increase in temperature can change the Gibbs free energy [Eq. (2), where ΔG° is the Gibbs free energy, ΔH°

$$\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ \tag{2}$$

is the reaction enthalpy, and ΔS° is the entropy change for the process] to a more favorable (that is, more negative) value. In particular, ΔG° will become more negative if ΔS° is positive and the temperature is increased. In many cases a combination of these factors causes the observed physiochemical process. *See* CHEMICAL THERMODYNAMICS; KINETICS (CLASSICAL MECHANICS).

Instrumentation for thermal analysis requires sensitive and robust temperature sensors and mechanisms for heating samples. Results are best with small (milligram) sample sizes and low heating rates. Sample and reference holders must be chemically inert and allow rapid thermal equilibration. Samples may be studied in air or other atmospheres, inert or reactive, at reduced, ambient, or elevated pressures. Most samples are solids, although an increasing number of liquid samples are being analyzed. References are used only in differential methods. Exact experimental conditions are chosen to enhance the process under study and ensure reproducibility. Selectable experimental parameters are heating

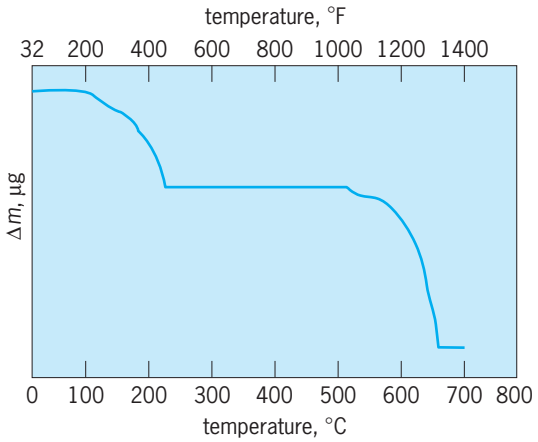


Fig. 1. Typical thermogravimetric curve for calcium nitrate tetrahydrate, $Ca(NO_3)_2 \cdot 4H_2O$, illustrating mass loss.

(cooling) rate, temperature range, atmospheric composition and pressure, and sample preparation. Each of these may affect the thermogram produced.

Thermograms are plots of the measured property, such as mass in thermogravimetry or differential heat flow in differential scanning calorimetry, versus the sample temperature (**Fig. 1**). Under rigorously controlled conditions, thermograms uniquely represent the system under study. Melting, crystallization, decomposition, oxidation, adsorption, absorption, desorption, polymerization reactions, and heat-capacity changes are observable on thermograms. A unique feature of thermograms is that they may also reflect a sample's history or method of preparation such as the cooling rate used to prepare a polymer, or ultraviolet or nuclear radiation exposure and damage.

In the study of reaction kinetics, isothermal analyses are often performed. This procedure changes the sample temperature rapidly from ambient to a desired value; the thermogram is the appropriate physiochemical property plotted as a function of time.

Thermogravimetry. This method involves measuring the changes in mass of a substance, typically a solid, as it is heated. Specially designed thermobalances are required to continuously monitor sample mass during the heating process. Modern balances have a capacity of 1–1500 milligrams and can accurately detect mass changes of 0.1 microgram.

Any type of physiochemical process which involves a change in sample mass may be observed by using thermogravimetry. Mass losses are observed for dehydration, decomposition, desorption, vaporization, sublimation, pyrolysis, and chemical reactions with gaseous products. Mass increases are noted with adsorption, absorption, and chemical reactions of the sample with the atmosphere in the oven, such as the oxidation of metals.

Quantitative gravimetric analyses may be performed due to the precise measure of the mass change obtained. Rates of mass change have been used to evaluate the kinetics of a process and to estimate activation energies. Fine details of these thermograms may also be used to deduce reaction intermediates and reaction mechanisms (Fig. 1).

Primary applications of thermogravimetry are to deduce stabilities of compounds and mixtures of elevated temperatures and to determine appropriate drying temperatures for compounds and mixtures. Evaluation of polymers, food products, and pharmaceuticals is a major application of thermogravimetry.

Differential thermal analysis. This method involves the monitoring of the temperature difference T_D between a sample and inert reference material (such as aluminum oxide) as they are simultaneously heated, or cooled, at a predetermined rate. Multijunction thermocouples and thermistors are the most common temperature sensors used for this purpose; they are arranged in an oven (**Fig. 2**). As enthalpic changes occur, T_D will be positive if the process is exothermic and negative if it is endothermic. Endotherms are plotted as negative deviations from the baseline (**Fig. 3**).

Differential thermal analysis thermograms are affected by instrumental factors such as furnace design, sample-holder material and geometry, thermocouple size and placement, and instrument response and furnace atmosphere. Sample characteristics that affect heat flow such as particle size, thermal conductivity, heat capacity, packing density, and the amount of sample can affect the appearance of thermograms. Under rigorously specified conditions, thermograms will uniquely represent the physicochemical characteristics of the sample.

More physical and chemical processes may be observed using differential thermal analysis as compared to thermogravimetry. Endothermic physical processes include crystalline transitions, fusion, vaporization, sublimation, desorption, and adsorption. Endothermic physical processes include crystalline

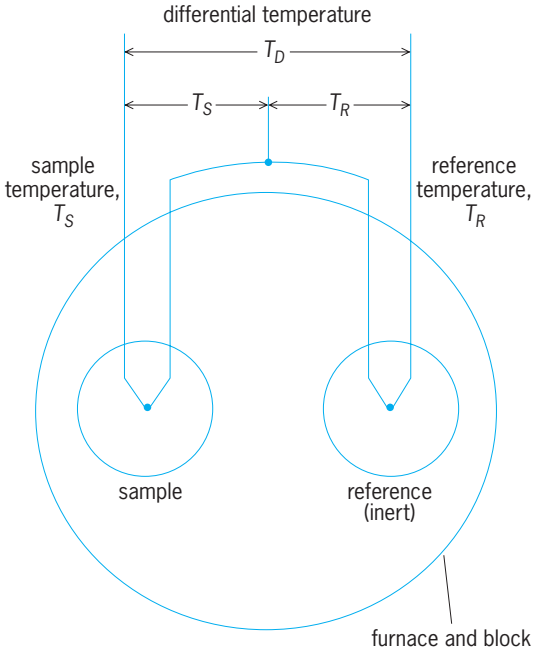


Fig. 2. Diagram of heating and temperature sensing units for differential thermal analysis. The oven heats the sample and the reference, and thermocouples determine ΔT (change in temperature) as well as the oven temperature.

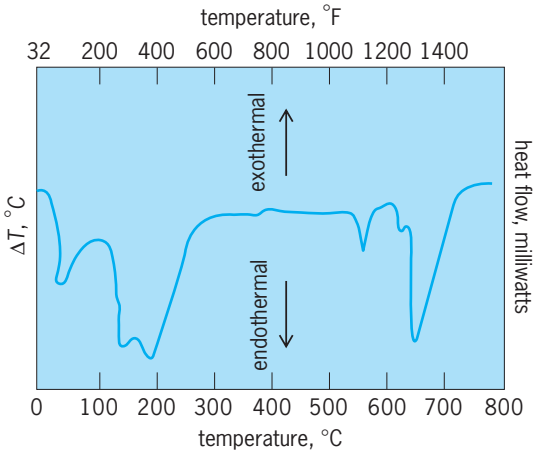


Fig. 3. Typical thermogravimetric curve for calcium nitrate tetrahydrate, $Ca(NO_3)_2 \cdot 4H_2O$, for differential thermal analysis when the *y* axis is ΔT, and for differential scanning calorimetry when the *y* axis has units of heat flow in milliwatts.

transitions, fusion, vaporization, sublimation, desorption, and adsorption. Endothermic chemical processes include dehydration, decomposition, gaseous reduction, redox reactions, and solid-state metathesis. Exothermic processes include adsorption, chemisorption, decomposition, oxidation, redox reactions, and solid-state metathesis reactions. Both solids and liquids can be studied by differential thermal analysis. Hermetically sealed capsules are often used for liquids and some solids. Other samples are studied in open or crimped pans.

Analytical applications of this technique include the identification, characterization, and quantitation of a wide variety of materials, including polymers, pharmaceuticals, metals, clays, minerals, and inorganic and organic compounds. Characteristic thermograms can be used to determine purity, heats of reaction, thermal stability, phase diagrams, catalytic properties, and radiation damage.

Differential scanning calorimetry. In this method, a sample and a reference are individually heated at a predetermined rate by separately controlled resistance heaters (**Fig. 4**). Enthalpic (heat-generating or -absorbing) processes are detected as differences in electrical energy supplied to either the sample or the reference material to maintain this heating rate (Fig. 3). This difference in electrical energy, in milliwatts per second, of the heat flow into or out of the sample is due to the occurrence of a physical or chemical process. Modulated differential scanning calorimetry is a new method that superimposes a sine wave on the heating ramp. A significant increase in sensitivity is often observed with modulated differential scanning calorimetry.

Analytical uses of differential scanning calorimetry are very similar to those of differential thermal analysis. Usually one calibration standard is sufficient to calibrate the entire operating range of the instrument. Differential scanning calorimetry instruments are highly sensitive and may measure heat flows as small as 1 nanowatt. Differential scanning calorimetry is very useful in determining heat capacities of substances over large temperature ranges. Such evaluation has become important in polymer and biochemical studies. Small (approximately 1–10 mg) samples are used in most cases, although some instruments have been developed which use up to 1 ml of a liquid sample. *See* CALORIMETRY.

Quantitative analysis in both differential thermal analysis and differential scanning calorimetry involves relationship (3) between the peak area A,

$$A = k(m)(\Delta H) \tag{3}$$

sample mass m, and enthalpy ΔH of the transition. The proportionality constant k varies with temperature in differential thermal analysis. The best results are obtained when the instrument is calibrated (k is determined) with a standard that has an enthalpic peak at a temperature close to the sample's peak. In differential scanning calorimetry the value of k varies less with temperature but still should be determined as with differential thermal analysis. A number of reliable calibration standards are readily available.

A significant property measurable by both differential thermal analysis and differential scanning calorimetry is the glass transition temperature T_g of polymers. Polymers often soften over a wide temperature range as glass does, rather than melt at a sharply defined temperature as smaller compounds do. The T_g of a polymer is an important physical property. *See* POLYMER.

Other methods. Thermomechanical analysis, thermoluminescence, emanation thermal analysis, evolved-gas analysis, thermomagnetic analysis, differential thermal rheology, differential photocalorimetry, dynamic mechanical thermal analysis, simultaneous thermal analysis (thermogravimetry and

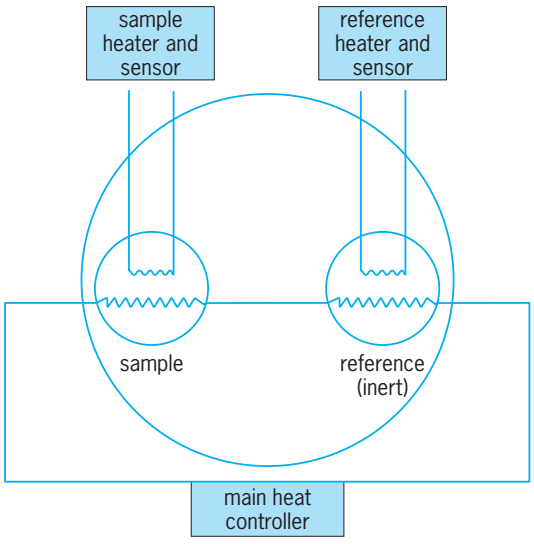


Fig. 4. Diagram of heating and temperature sensing units for differential scanning calorimetry. The main heater heats the sample and the reference equally and at a rate slightly less than the preset heating rate. Individual sample and reference heaters add heat to maintain the preset heating rate. The difference in energy to individual heaters is the heat flow plotted in the thermogram.

differential thermal analysis together), and microthermal analysis are additional thermal analytical methods. Thermomechanical analysis can be used to evaluate the physical stability of structural or electronic components. Thermoluminescence is often used for authenticating ancient objects. Emanation and evolved-gas thermal analysis are used to determine gaseous products, often from destructive oxidation. These methods are often combined with mass spectrometry or gas chromatography/mass spectrometry. Thermomagnetic analysis can be used to determine Curie points of metals. Differential thermal rheology measures the deformation of matter with temperature increases. Differential photo calorimetry monitors heat generated by light-induced chemical processes. Dynamic mechanical thermal analysis instruments can monitor cantilever bending, tensile strength, shear, compression, and creep as a function of temperature. Simultaneous analysis allows differential thermal analysis and thermogravimetry to be performed on the same sample at the same time with a considerable savings in operator time. Microthermal analysis places a heater and temperature sensor at the tip of an atomic force microscope to measure microscopic thermal features. *See* RHEOLOGY; SCANNING ELECTRON MICROSCOPE.

In some instances the first derivative of the thermogram is the best format for qualitative assessment of samples. For instance, the first-derivative plots show small features, such as shoulders on peaks, more clearly. Derviative curves are produced by digital calculations on the computer stored data. Virtually every method discussed above has been adapted to differential methods. Integration of peak areas for quantitative work is now done by special algorithms and computer programs. Computer-driven calculations provide consistent data treatment unattainable with manual analysis of thermograms. *See* CHEMICAL DYNAMICS; PHASE EQUILIBRIUM; THERMAL EXPANSION; THERMOCHEMISTRY; TRANSITION POINT.

Neil D. Jespersen

Bibliography. J. Cazes (ed.), *Ewing's Analytical Instrumentation Handbook*, 3d ed., 2004; P. J. Haines (ed.), *Principles of Thermal Analysis and Calorimetry*, 2002; G. Hohne, W. Hemminger, and H. J. Flemmersheim (eds.), *Differential Scanning Calorimetry: An Introduction*, Springer, Berlin, 2d ed., 2003; B. Wunderlich, *Thermal Analysis of Polymeric Materials*, 2005.

Thermal conduction in solids

Thermal conduction in a solid is generally measured by stating the thermal conductivity K, which is the ratio of the steady-state heat flow (heat transfer per unit area per unit time) along a long rod to the temperature gradient along the rod. Thermal conductivity varies widely among different types of solids, and depends markedly on temperature and on the purity and physical state of the solids, particularly at low temperatures.

From the kinetic theory of gases the thermal conductivity can be written as in Eq. (1), where S is the

$$K = (\text{constant})\, Svl \tag{1}$$

specific heat per unit volume, v is the average particle velocity, and l is the mean free path. In solids, thermal conduction results from conduction by lattice vibrations and from conduction by electrons. In insulating materials, the conduction is by lattice waves; in pure metals, the lattice contribution is negligible and the heat conduction is primarily due to electrons. In many alloys, impure metals, and semiconductors, both conduction mechanisms contribute. *See* CONDUCTION (HEAT); KINETIC THEORY OF MATTER; LATTICE VIBRATIONS; SPECIFIC HEAT OF SOLIDS.

Insulating solids. In insulators, heat is transported by lattice vibrations. The vibrations of the crystal lattice can be resolved into traveling elastic waves (lattice waves). Each lattice wave can be specified by a wave vector $\mathbf{k}$ and by a mode of polarization (either longitudinal or transverse). The energy of the lattice vibrations is taken as the energy of lattice waves of frequency ν, where ν is of the order of k_BT/h; k_B is Boltzmann's constant, T is the absolute temperature, and h is Planck's constant. The energy of each wave is described as consisting of an integral number of phonons, each of energy $h\nu$ and of momentum $hk_B/2\pi$. *See* PHONON.

At equilibrium, the phonons are distributed isotropically in momentum space; when a temperature gradient is established along the crystal, an asymmetry in the phonon distribution is set up. An interchange of energy among the lattice waves, which can result from the anharmonicity of the lattice forces and from lattice imperfections, tends to restore equilibrium; the rate at which equilibrium is restored is associated with a finite mean free path l for the

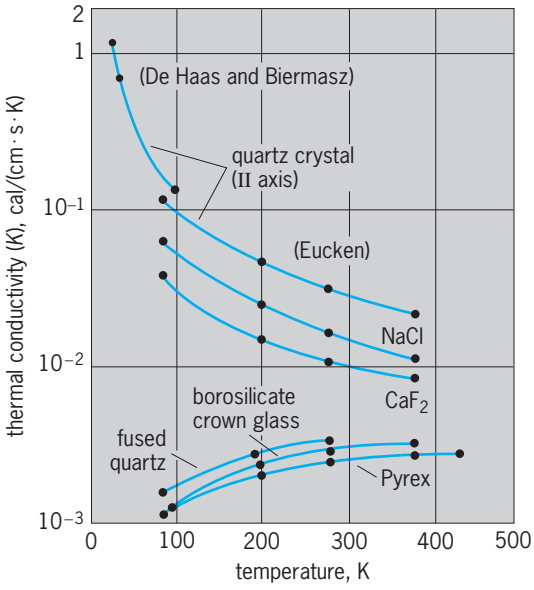


Fig. 1. Thermal conductivity of some insulating and some amorphous materials as a function of temperature. °F = (K × 1.8) − 459.67. (*After C. Kittel, Introduction to Solid State Physics, 3d ed., Wiley, 1966*)

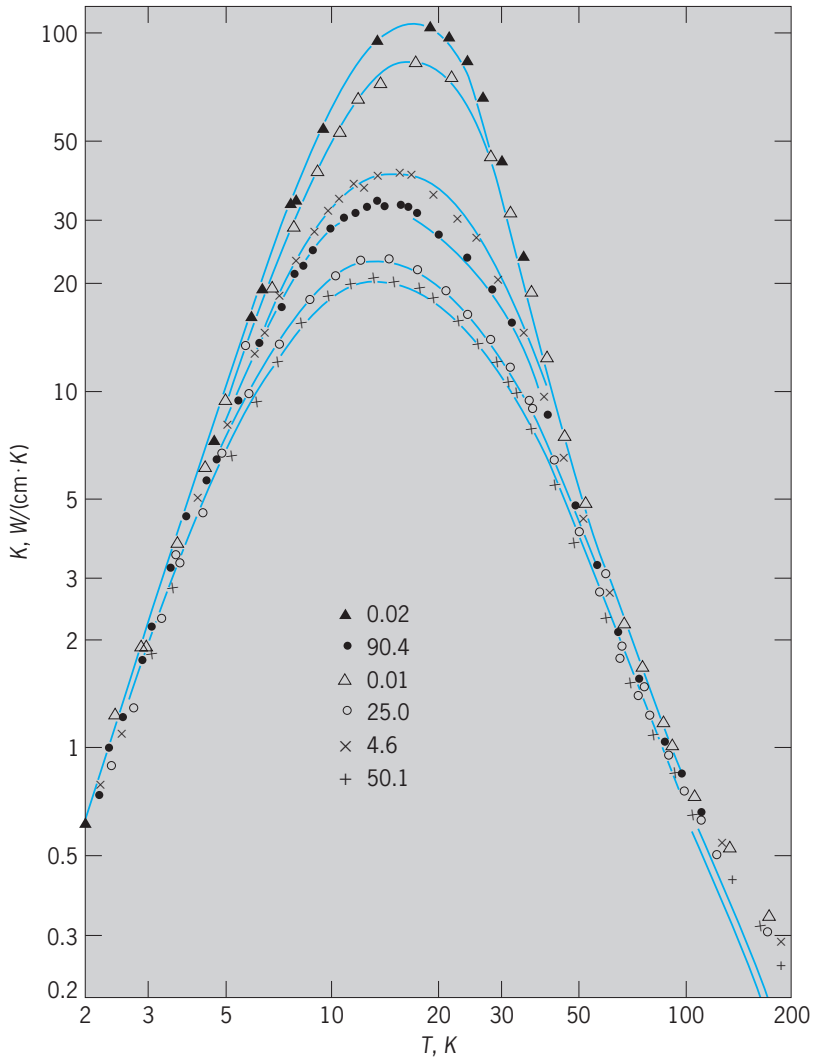


Fig. 2. Thermal conductivity of lithium fluoride as a function of temperature for various percentage concentrations of ^{6}Li. In the peak the highest value is approximately five times that of the lowest peak. $^\circ F = (K \times 1.8) - 459.67$. (*After R. Berman, Heat conductivity of non-metallic crystals at low temperatures, Cryogenics, 5:297–305, 1965*)

phonons. At low temperatures K varies markedly with temperature and is dependent on the variation of l with phonon frequency ν. **Figure 1** shows the conductivity of some insulating materials.

The form of l at low temperatures is closely related to the type of phonon scattering and hence to the type or types of imperfections present in the particular solid. The thermal conductivity at low temperatures is a sensitive function of these imperfections. For example, in **Fig. 2**, the effect of the isotope of lithium ^{6}Li on the conductivity of lithium fluoride crystals is shown; here the differences in conductivity are due principally to the mass difference between ^{6}Li and ^{7}Li.

The introduction of molecular impurities into dielectric crystals produces resonance scattering, as shown in **Fig. 3** for the potassium chloride crystals containing cyanide ions; here the center of the resonance (the dip in the thermal conductivity curves) corresponds to the frequency of the particular phonons carrying heat at that temperature. The thermal conductivity of solid helium has been measured at low temperatures, and is shown in **Fig. 4**; for comparison, the thermal conductivities of very pure lithium fluoride and sapphire are given. At less than 1 K (2°F above absolute zero), the conductivity of solid helium increases very rapidly with temperature and is characterized as Poiseuille flow, in which the phonons are described as flowing as a viscous fluid. *See* FLUID MECHANICS; HYDRODYNAMICS.

Glass and polycrystalline materials. For amorphous materials, glasses, and polycrystalline materials, the mean free path for the phonons is generally small, of the order of the distance between atoms, and is independent of temperature. The thermal conductivity for these materials is much less than that for the insulating materials above; for example, quartz (fused) glass can be compared with crystal quartz in Fig. 1, and sintered alumina (aluminum oxide) in **Fig. 5** can be compared with sapphire (single-crystal aluminum oxide) in Fig. 4.

Metals, alloys, and semiconductors. The thermal conductivity of a metal can be written as the sum of an electronic component K_e and a lattice component

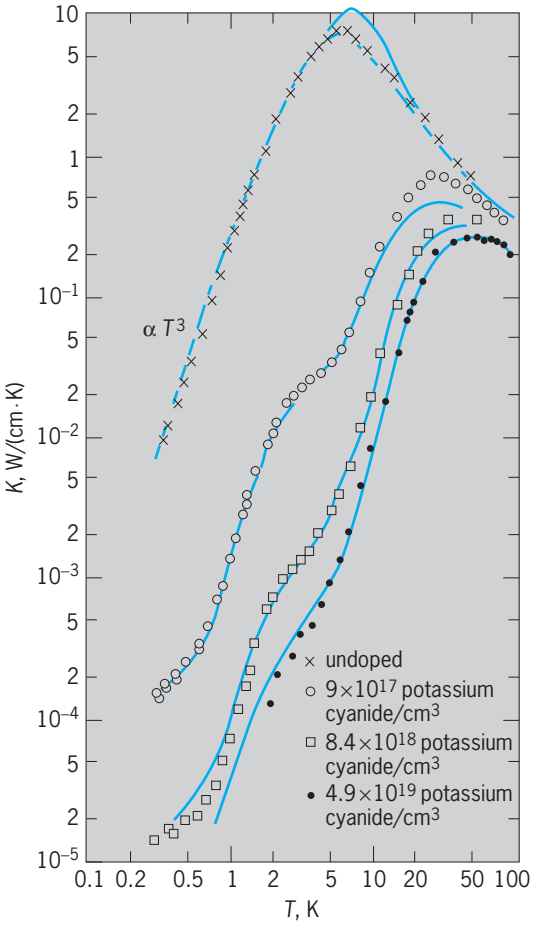


Fig. 3. Thermal conductivity of pure potassium chloride and of potassium chloride containing cyanide ions as a function of temperature. °F = (K × 1. 8) − 459.67. (*After R. Berman, Heat conductivity of non-metallic crystals at low temperatures, Cryogenics, 5:297–305, 1965*)

K_g; K_e and K_g are each limited by various scattering mechanisms. In real crystals, as compared to the ideal crystal with a perfect lattice, the electrons undergo scattering caused by the thermal vibrations of the lattice and by imperfections (such as impurities, point defects, dislocations, and grain boundaries).

In pure metals, the electronic component accounts for nearly all the heat conducted, while the lattice component, in most cases, is negligible. The electronic thermal conductivity is related to the electrical conductivity through the mean free path. In certain temperature regions the value of the mean free path for both thermal and electrical conduction can be assumed to be the same; for these cases, the Wiedemann-Franz law is applicable, as in Eq. (2),

$$\frac{K_e}{\sigma T} = L = \frac{\pi k_B^2}{3e^2} \qquad (2)$$

where L (the Lorentz constant) $= 2.45 \times 10^{-8}$ (watt)(ohm)(K^{-2}) $= 7.5 \times 10^{-9}$ (watt)(ohm)($°F^{-2}$), σ is the electrical conductivity, and e is the electronic charge. Here the free-electron theory of metals is assumed, in which the conduction electrons form a gas which obeys Fermi-Dirac statistics. *See* FREE-ELECTRON THEORY OF METALS.

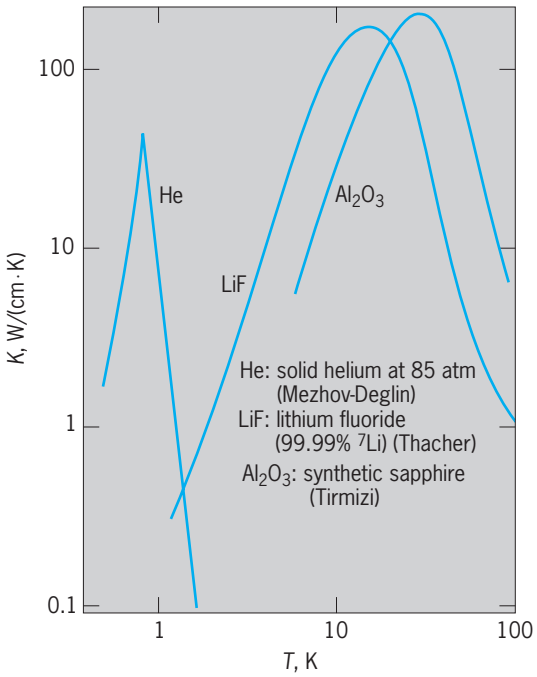


Fig. 4. Thermal conductivity of solid helium as a function of temperature. Thermal conductivities of very pure lithium fluoride and sapphire are given for comparison. °F = (K × 1.8) − 459.67. (*After R. Berman, Heat conductivity of non-metallic crystals at low temperatures, Cryogenics, 5:297–305, 1965*)

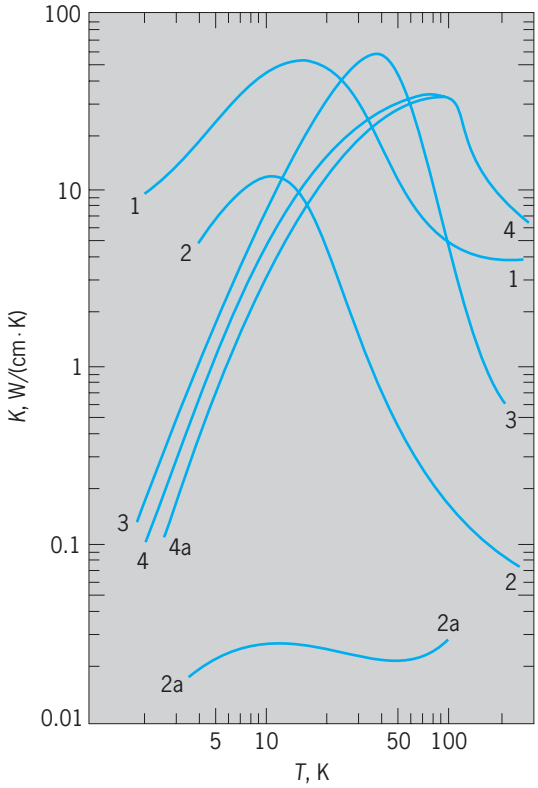


Fig. 5. Thermal conductivity of some insulators as a function of temperature. Curve 1, pure copper (for comparison). Curve 2, quartz crystal. Curve 2*a*, quartz crystal after intense neutron irradiation. Curve 3, aluminum oxide (sintered alumina). Curve 4, diamond. Curve 4*a*, diamond, reduced diameter. °F = (K × 1.8) − 459.67. (*After R. Berman, Heat conductivity of non-metallic crystals at low temperatures, Cryogenics, 5:297–305, 1965*)

The electrical resistivity ρ (the reciprocal of the electrical conductivity σ) can be separated into two parts, one called the residual resistivity ρ_0, which results from the elastic scattering of electrons by imperfections and which is independent of temperature, and the other called the ideal resistivity $\rho_i(T)$, which results from scattering due to lattice vibrations and which is temperature-dependent. Thus the electrical resistivity is written $\rho = 1/\sigma = \rho_0 + \rho_i(T)$. The thermal resistivity for metals W (the reciprocal of the thermal conductivity K_e) can be written analogously as the sum of two terms, one, W_0, due to scattering by imperfections, and the other, $W_i(T)$, due to scattering by lattice vibrations. Thus the thermal conductivity is written $W = 1/K_e = W_0 + W_i(T)$. In **Fig. 6** the thermal conductivities of three samples of gold are given; since the purity of the gold differs, the samples differ in both ρ_0 and W_0. In the low-temperature region, the thermal conductivity is proportional to T. At higher temperatures, the thermal resistivity due to lattice vibrations exceeds W_0, and the curves for the three different samples converge.

In the case of alloys, W_0 is much larger than in pure metals and a lattice thermal conductivity must also be included. The thermal conductivities of several alloys of gold are also shown in Fig. 6. It is possible from such curves to obtain K_g as a function of T by also measuring ρ; ρ_0 can be determined from W_0. In well-annealed alloys, K_g is proportional to T_2 at low temperatures; at higher temperatures, K_g decreases with T.

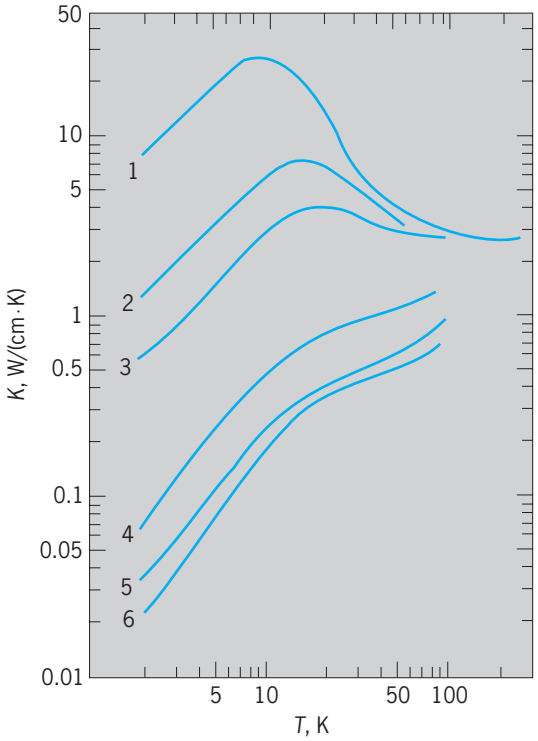


Fig. 6. Thermal conductivity of gold and some alloys as a function of temperature. Curve 1, 99.999% gold, annealed. Curve 2, same, cold-drawn. Curve 3, 99.9% gold, cold-drawn. Curve 4, gold with 0.7% platinum. Curve 5, gold with 1.7% platinum. Curve 6, gold with 0.2% chromium. $^\circ$F = (K × 1.8) − 459.67. (*After F. Seitz and D. Turnbull, eds., Solid State Physics, vol. 7, 1958*)

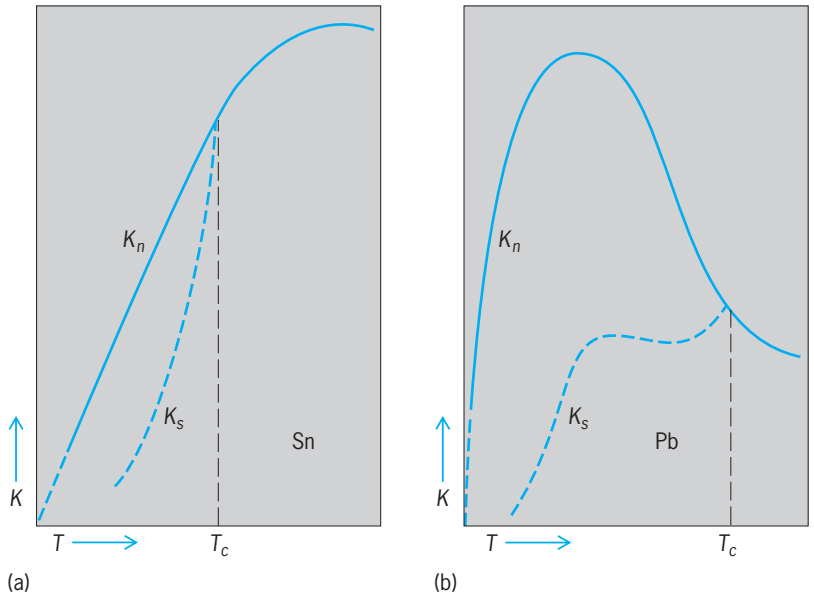


Fig. 7. Temperature dependences of thermal conductivity of (*a*) tin and (*b*) lead in the normal and superconducting states. (*After F. Seitz and D. Turnbull, eds., Solid State Physics, vol. 12, 1961*)

In the case of semiconductors, such as high-purity germanium and silicon, phonons are primarily responsible for the thermal conduction. The form of the thermal conductivity versus temperature curves is similar to that for single crystals of dielectric materials.

Superconductors. In superconductors at temperatures below the critical temperature T_c, the electronic conduction is reduced; at sufficiently low temperatures, the thermal conductivity becomes entirely due to lattice waves and is similar to the form of the thermal conductivity of an insulating material. In **Fig. 7** the temperature dependences of the thermal conductivity of tin and lead are given; here K_n is the thermal conductivity with the material in the normal state, and K_s is the thermal conductivity with the material in the superconducting state. *See* SUPERCONDUCTIVITY. Kathryn A. McCarthy

Bibliography. R. Berman, *Thermal Conduction in Solids*, 1976; C. M. Bhandari and D. M. Rowe, *Thermal Conduction in Semiconductors*, 1988; G. E. Childs, L. J. Erick, and R. L. Powell, *Thermal Conductivity of Solids at Room Temperature and Below: A Review and Compilation of the Literature*, National Bureau of Standards Monogr. 131, 1973; Y. Godovsky and V. P. Privalko, *Thermal and Electrical Conductivity of Polymer Materials*, 1995; P. G. Klemens and T. K. Chu, *Thermal Conductivity*, 1976.

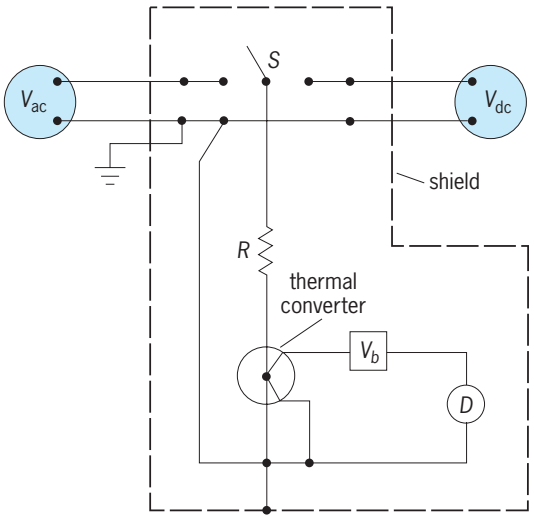
Thermal converters

Devices consisting of a conductor heated by an electric current, with one or more hot junctions of a thermocouple attached to it, so that the output emf

responds to the temperature rise, and hence the current. Thermal converters are used with external resistors for alternating-current (ac) and voltage measurements over wide ranges and generally form the basis for calibration of ac voltmeters and the ac ranges of instruments providing known voltages and currents.

Basic form. In the most common form, the conductor is a thin straight wire less than 0.4 in. (1 cm) long, in an evacuated glass bulb, with a single thermocouple junction fastened to the midpoint by a tiny electrically insulating bead. Thermal inertia keeps the temperature of the heater wire constant at frequencies above a few hertz, so that the constant-output emf is a true measure of the root-mean-square (rms) heating value of the current. The reactance of the short wire is so small that the emf can be independent of frequency up to 10 MHz or more. An emf of 10 mV can be obtained at a rated current less than 5 mA, so that resistors of reasonable power dissipation, in series or in shunt with the heater, can provide voltage ranges up to 1000 V and current ranges up to 20 A. However, the flow of heat energy cannot be controlled precisely, so the temperature, and hence the emf, generally changes with time and other factors. Thus an ordinary thermocouple instrument, consisting of a thermal converter and a millivoltmeter to measure the emf, is accurate only to about 1–3%. *See* THERMOCOUPLE.

AC-DC transfer instrument. To overcome this, a thermal converter is normally used as an ac-dc transfer instrument (ac-dc comparator) to measure an unknown alternating current or voltage by comparison with a known nearly equal dc quantity (see **illus.**). By replacing the millivoltmeter with an adjustable, stable, opposing voltage V_b in series with a microvoltmeter D, very small changes in emf can be detected. The switch S is connected to the unknown ac voltage V_{ac}, and V_b is adjusted for a null (zero) reading of D. Then S is immediately connected to the dc voltage V_{dc}, which is adjusted to give a null again, without changing V_b. Thus $V_{ac} = V_{dc}(1 + d)$, where d is the ac-dc difference of the transfer instrument, which can be as small as a few parts per million (ppm).



Basic circuit for ac-dc transfer measurements of ac voltages.

In many commercial instruments, all of the components are conveniently packaged in the shield, shown with a broken line, and several ranges are available by taps on R. Accuracies of 0.001% are attainable at audio frequencies. For the highest accuracy, the average of the two directions of V_{dc} should be taken as the reference, because of slight imperfections of the thermal converter, and the ac measurement should be repeated, to guard against small drifts. The resistances of the two pairs of leads from V_{ac} and V_{dc} to S should be equal, and the leads should be shielded.

Multijunction thermal converter. A converter with many thermocouples attached uniformly along a bifilar heater can have intrinsic values of d less than 0.1 ppm from thermoelectric effects. Such multijunction thermal converters (MJTCs) are the primary ac-dc standards in several national metrology laboratories. MJTCs can be made from wire heaters and thermocouples or by using microfabrication techniques to produce thin-film heaters and thermocouples on silicon substrates. The heaters and thermocouple hot junctions are formed on a thin, thermally isolating membrane such as SiO_2/Si_3N_4 (not directly on the silicon substrate), and the cold junctions are formed on the silicon substrate to provide a stable temperature reference.

Solid-state converter. This type of converter uses a transistor, sensitive to temperature, in place of a thermocouple. Dual heaters and transistors are connected in a feedback circuits so that an output voltage proportional to the input current is obtained. *See* TRANSISTOR.

MJTCs and solid-state converters are also used in a number of more complex instruments. In some instruments the heater is switched at regular intervals between the unknown ac input and a dc current that is automatically controlled to keep the output constant. Other instruments have dual multijunction thermal converters, with two sets of thermocouples connected to each other in opposition and connected to an amplifier that supplies that second heater. In these feedback circuits the output voltage is equal to the voltage on the first heater. *See* FEEDBACK CIRCUIT.

Calibration. The values of d (which depend on range and frequency) are best determined by comparison with ac-dc transfer instruments of known characteristics. These, in turn, are evaluated in metrology laboratories, such as the National Institute of Standards and Technology in the United States, where ac-dc differences have been studied very carefully, because the ultimate accuracy of almost all alternating-current and voltage measurements depends upon thermal converters. *See* ELECTRICAL MEASUREMENTS. F. L. Hermach; Joseph R. Kinard

Bibliography. F. L. Hermach, AC-DC comparators for audio-frequency current and voltage measurements of high accuracy, *IEEE Trans. Instrum. Meas.*, IM-25:489–494, 1976; J. R. Kinard et al., Development

of thin-film multijunction thermal converters at NIST, *IEEE Trans. Instrum. Meas.*, 46(2):347–351, April 1997; E. S. Williams, *The Practical Uses of AC-DC Transfer Instruments*, NBS Tech. Note 1166, 1982.

Thermal ecology

The examination of the independent and interactive biotic and abiotic components of naturally heated environments. Geothermal habitats are present from sea level to the tops of volcanoes and occur as fumaroles, geysers, and hot springs. Hot springs typically possess source pools with overflow, or thermal, streams (rheotherms) or without such streams (limnotherms). Hot-spring habitats have existed since life began on Earth, permitting the gradual introduction and evolution of species and communities adapted to each other and to high temperatures. Other geothermal habitats do not have distinct communities.

Hot-spring pools and streams, typified by temperatures higher than the mean annual temperature of the air at the same locality and by benthic mats of various colors, are found on all continents except Antarctica. They are located in regions of geologic activity where meteoric water circulates deep enough to become heated. The greatest densities occur in Yellowstone National Park (northwest United States), Iceland, and New Zealand. Source waters range from 40°C (104°F) to boiling (around 100°C or 212°F depending on elevation), and may even be superheated at the point of emergence. Few hot springs have pH 5–6; most are either acidic (pH 2–4) or alkaline (pH 7–9).

Alkaline hot springs. Alkaline springs have been studied in the greatest detail. The main results of these studies are described below.

Physicochemical properties. Heated subterranean water becomes highly charged with dissolved materials due to its increased dissolution capabilities. Hot-spring waters have 10–15 times more dissolved inorganic solutes than nonthermal waters; their nutrient levels are similar to eutrophic lakes. The major solutes of nonacid springs include calcium, chlorine, potassium, magnesium, sodium, silica, and sulfates. Toxic elements, such as arsenic and mercury, may render the water unpotable. *See* EUTROPHICATION.

While source waters remain relatively constant, the abiotic properties do change with distance from source. Thus, unidirectional gradients of heat, hydrogen sulfide, carbon dioxide, nitrates, and pH exist along these streams (**Fig. 1**). The rate of change per unit distance depends on stream characteristics, meteorological conditions, and biotic use or release of solutes. In Fig. 1, region A is inhabited by patches of bacteria attached to the substrate. Region B begins at the 73°C (163°F) isotherm with the first appearance of continuous benthic mat. Above 70°C (158°F) this mat contains only *Synechococcus lividus*; below 70°C (158°F) it is joined by *Chloroflexus aurantiacus*. Region C represents an area of increasing species diversity. At temperatures in the upper 50s

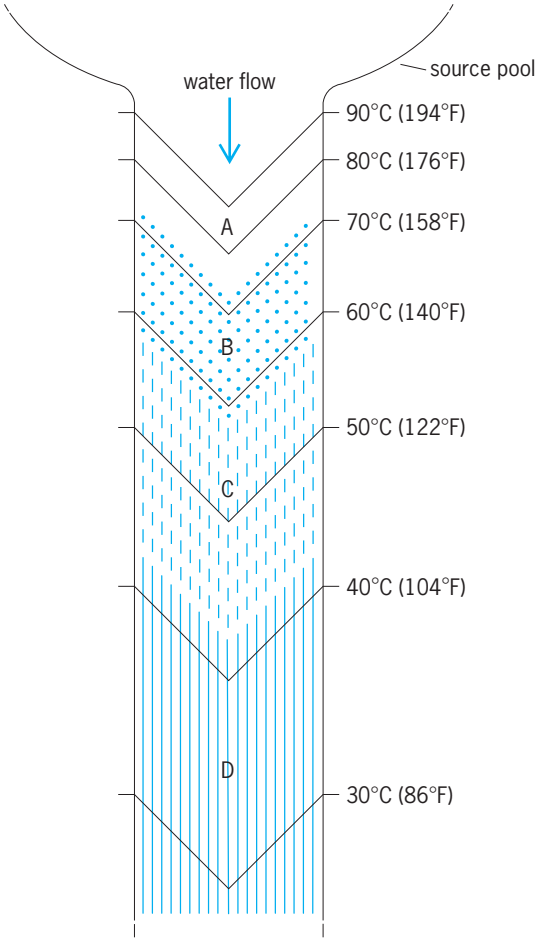


Fig. 1. Distribution of organisms and temperatures along an alkaline thermal spring. Isotherms are shown for 10°C (18°F) intervals. The distinctive V shape results from more rapid heat loss at edges than in main current; heat loss is approximately exponential with distance from source. Four regions (A–D) of differing biological character are illustrated; these are described in the text.

[58°C (136°F) here] other cyanobacteria join the mat, and at lower temperatures eukaryotes are present (see **table**). Region D often is influenced by the grazing of herbivores. While most upstream (B and C) organisms are present here as well, *C. aurantiacus* is absent, and the mat may be dominated by different species (for example, *Calothrix* spp.).

Communities. Cyanobacteria (blue-green algae) and other auto- and heterotrophic bacteria form the benthic mats of alkaline hot springs from 30 to 73°C (86 to 163°F). Mat coloration results from variations of surface pigments due to species distributions, temperature, and light history. Only prokaryotes are found above 60°C (140°F); eukaryotes are not common above 40–45°C (104–113°F; see table).

Source pools, even at boiling temperatures, may contain bacteria. These chemoautotrophs oxidize sulfides, but have not been cultured or identified. Yellow patches of organoheterotrophic bacteria *Thermus aquaticus* are present from 73 to 79°C (163 to 174°F).

As water temperature decreases downstream, overlapping populations of microorganisms appear

Upper temperature limit and physiological activity of representative hot-spring organisms

Organism	Taxonomic category	Physiological activity*	Maximum temperature, °C (°F)
	Alkaline hot springs		
Thermothrix thioparus	Bacterium	CA	80 (176)
Thermus aquaticus	Bacterium	OH	79 (174)
Synechococcus lividus	Cyanobacterium	PA	73 (163)
Chloroflexus aurantiacus	Photosynthetic bacterium	PH	70 (158)
Phormidium laminosum	Cyanobacterium	PA	60 (140)
Mastigocladus laminosus	Cyanobacterium	PA	58 (136)
Oscillatoria terebriformis	Cyanobacterium	PA	54 (129)
Potamocypris sp.	Ostracod	OH	48 (118)
Achnanthes exigua	Alga (diatom)	PA	44 (111)
Oscillatoria princeps	Cyanobacterium	PA	~42 (108)
Paracoenia turbida	Brine fly	OH	36–38 (97–100)
	Acid hot springs		
Sulfolobus acidocaldarius	Bacterium	CA	89 (192)
Bacillus acidocaldarius	Bacterium	OH	70 (158)
Cyanidium caldarium	Alga (red)	PA	56 (133)
Dactylaria gallopava	Fungus (deuteromycete)	OH	52 (126)
Euglena mutabilis	Alga (euglenoid)	PA	35–40 (95–104)
Zygogonium sp.	Alga (green)	PA	~30 (86)

*Autotrophs (A) obtain cell carbon through reduction of CO_2; heterotrophs (H) rely on organic molecules. Cellular energy may be derived from sunlight [photo- (P)], from oxidation of reduced inorganic molecules [chemo- (C)], or from oxidation of organic molecules [organo- (O)]. For example, photosynthetic plants are photoautotrophs (PA), and most animals are organoheterotrophs (OH). The designated activity may not be the only one possible for that organism.

as components of the mat community at temperatures representing their upper limit (Fig. 1; table). These distinctive thermal mats begin at 73°C (163°F) with *S. lividus*. Below 50°C (122°F) taxonomic and trophic diversity increases rapidly with decreasing temperatures. Protozoa, ostracods, and other invertebrates become components of the systems as herbivores at these lower temperatures. By the time the water has cooled to around 40°C (104°F), multispecies food chains and webs exist. *See* FOOD WEB.

The upstream and downstream distribution of a species may also depend on requirements for and tolerance of other abiotic factors (such as hydrogen sulfide concentration) and biotic factors (such as competition, grazing, and commensalism).

Mats below 40–50°C (104–122°F) often are thin (1–4 mm or 0.04–0.16 in.) and cutrose and show little or no subsurface stratification. Higher-temperature mats are thicker (3–100 mm or 0.12–3.9 in.) and gelatinous, and usually have distinct layers (**Fig. 2**). Species composition along the vertical aspect of gelatinous mats is determined by light and oxygen.

Light-dependent microorganisms are restricted to the upper regions of the gelatinous mats. Cyanobacterial photosynthesis oxygenates the upper layer diurnally and secretes photosynthate used for growth of aerobes and facultative anaerobes in this region. During nocturnal periods, low levels of atmospheric oxygen diffuse to the mat surface; the oxygen is consumed by microbial respiration, however, rendering the upper layers anoxic at night.

The undermat layer (Fig. 2), principally composed of *C. aurantiacus*, assimilates cyanobacterial photosynthate for photoheterotrophic growth during the day and for aerobic respiration and growth at night when it migrates toward the surface to utilize atmospheric oxygen. Alternation between surface growth of cyanobacteria and this nocturnal migration yields a lamellar mat structure similar in appearance to algal stromatolites. Cyanobacterial cells covered by migrations and continued surface growth become moribund and die, thereby contributing organic material for growth of anaerobes in the region below light penetration where anoxic conditions persist. Decomposition primarily occurs in the bottom layers.

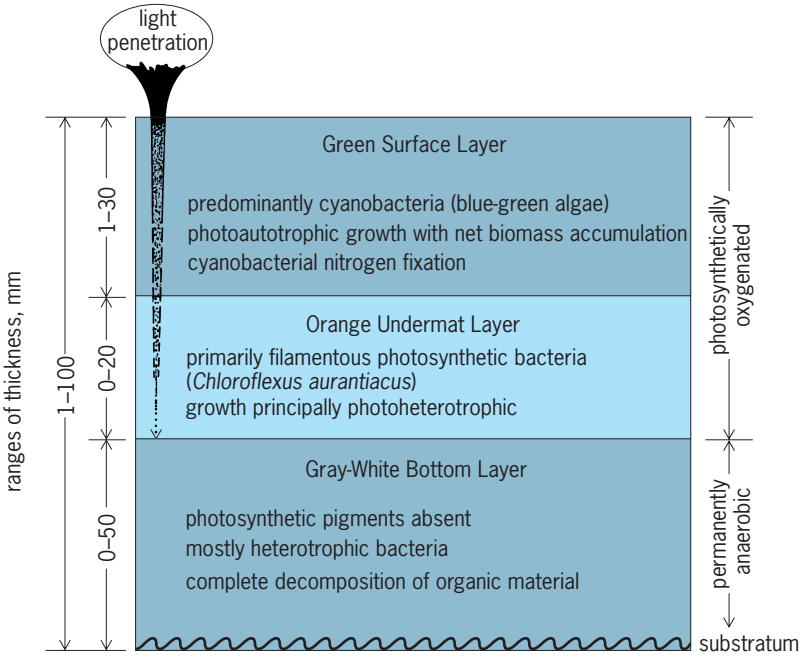


Fig. 2. Properties and distribution of activities within mature gelatinous mats along the vertical axis. The mats vary in thickness and composition with stream temperature. Above 70°C (158°F) only the surface layer is present; below about 45°C (113°F) the orange undermat area may be absent. Light reduction (due to self-shading) and extent of diffusion of oxygen from surface cyanobacterial photosynthesis depend on mat thickness. 1 mm = 0.04 in.

Cyanobacteria colonize barren thermal habitats; heterotrophic bacteria are limited by availability of dissolved organic compounds. Maximum cyanobacterial diversity is achieved within 3–4 weeks in thermal areas. Since the dominants of young mats are similar to those of mature mats, few successional patterns for microbial species are recognized. In some streams, *Chloroflexus* precedes *Synechococcus* and provides an attachment surface for the cyanobacterium; in low-nitrogen streams, nitrogen-fixing species (such as *Mastigocladus laminosus* and *Calothrix* spp.) are most active during early successional stages.

Cyanobacterial biomass accumulates rapidly, and within 2 weeks stable subsurface habitats are available to anaerobes. Some springs are as productive as eutrophic lakes. More than 97% of the community's biomass is prokaryotic. The standing crop biomass of mature mats ranges from 200 to 500 g/m^2 with peak values above 50°C (122°F). The undisturbed mats accumulate biomass and thicken until decomposition equals production. Grazers, such as the ostracod *Potamocypris*, may significantly decrease the standing crop, while others, such as ephydrid fly larvae, may stimulate productivity through nutrient turnover.

Adaptations of thermal biota. Many hot-spring organisms remain within their temperature range along the thermal gradient by movement. When suddenly exposed to high temperature shifts, insect larvae and the ostracod *Potamocypris* sp. drift downstream in the current to cooler temperatures. The ostracod returns upstream by positive thermotaxis until encountering its maximum (50°C or 122°F), when it reverses direction (thermophobic response). *Thermacarus nevadensis* moves upstream by positive rheotaxis. *Oscillatoria terebriformis* maintains its position along thermal and light-intensity gradients by thermo- and phototactic movements, respectively.

The temperature optimum for metabolic activity of nonmotile microorganisms usually is at or near the habitat temperature. This physiological adaptation ensures maximal activity where the organism is growing, and is known for cyanobacterial photosynthesis and nitrogenase activity and for incorporation of organic molecules by bacteria. This is even true for carbon dioxide fixation by the chemoautotrophs at boiling temperatures. Cyanobacterial cells at the mat surface are adapted to high light intensities, while those within the mat are adapted to low light.

Physiological adaptations have led to the selection of genetic strains in some species. Four stenothermal strains (ecotypes) of *S. lividus* with distinct cardinal (maximum, minimum, optimum) temperatures for growth comprise sequentially the upper *Synechococcus* layer from 50 to 73°C (122 to 163°F). Light-intensity ecotypes of the cyanobacterium *Plectonema notatum* also exist. In summer, the high-intensity strains dominate the mat, while low-intensity strains persist in subsurface layers; the latter strains dominate the mat during winter. The life cycles of some invertebrates are genetically adapted to their thermal niche. *Paracoenia bisetosa* adults oviposit on mat surfaces surrounded by lethal-temperature water. The larvae decimate mat structure through feeding activities, but prior to the reentry of hot water many of the adults emerge from pupae, and the remaining larvae and pupae are killed.

Distribution and dispersal of species. The microflora of thermal habitats is widespread within geographic regions, while the fauna is more restricted. Thermal-spring species seem as sensitive to unfavorable conditions encountered during dispersal as their mesophilic counterparts; additionally, thermophilic species must survive extended exposure to suboptimal temperatures.

Acid hot springs. Acid hot springs are less common and less well studied than the alkaline variety. Their main characteristics are described below.

Physicochemical properties. Most of these springs are acidic due to abiotic subterranean or surface biotic oxidation of hydrogen sulfide to sulfuric acid. Consequently, sulfate levels are high (often around 1 g/liter). Concentrations of metallic ions of aluminum, copper, iron, manganese, magnesium, and zinc are higher than in alkaline springs, as is ammonium (NH_4^+).

Biota. Compared to the alkaline streams, acid hot springs have less species diversity. No microorganisms live in boiling acid source waters, and cyanobacteria do not grow at pH < 4. *Cyanidium caldarium*, a eukaryote with red algal characteristics, forms emerald green mats (1–20 mm thick) from 35 to 55°C (95 to 131°F). These mats are not stratified, but filaments of the fungus *Dactylaria gallopava* and rod-shaped bacteria (such as *Bacillus* spp. and *Thiobacillus* spp.) are interspersed within them. These heterotrophs rely on *Cyanidium* photosynthate for growth.

Few microorganisms inhabit acid thermal streams above 55°C (131°F). *Bacillus acidocaldarius* occurs up to 65°C (149°F), and *Sulfolobus acidocaldarius*, a chemoautotroph, is found at temperatures up to 89°C (192°F). At less than 40°C (104°F) other photosynthetic organisms such as *Chlorella* sp. and *Euglena mutabilis* occur. A purple mat of *Zygogonium* sp. is present around 30°C (86°F). This mat comprises a community including the above algae, herbivorous brine flies, *Ephydra thermophila*, and carnivorous midges, *Bezzia setulosa*. *See* ECOLOGY.

Conrad Wickstrom

Bibliography. T. D. Brock, *Thermophilic Microorganisms and Life at High Temperatures*, 1978; C. Edwards (ed.), *Microbiology of Extreme Environments*, 1990; R. A. Herbert and G. A. Codd (eds.), *Microbes in Extreme Environments*, 1986; J. K. Kristjansson, *Thermophilic Bacteria*, 1991.

Thermal expansion

Solids, liquids, and gases all exhibit dimensional changes for changes in temperature while pressure is held constant. The molecular mechanisms at work

and the methods of data presentation are quite different for the three cases and are therefore discussed separately in this article.

Expansion of solids. The temperature coefficient of linear expansion α_l is defined by Eq. (1), where l is

$$\alpha_l = \frac{1}{l}\left(\frac{\partial l}{\partial t}\right)_{p=\text{const}} \tag{1}$$

the length of the specimen, t is the temperature, and p is the pressure. For each solid there is a Debye characteristic temperature Θ, below which α_l is strongly dependent upon temperature and above which α_l is practically constant. Many common substances are near or above Θ at room temperature and follow approximate equation (2), where l_0 is the length at 32°F

$$l = l_0(1 + \alpha_l t) \tag{2}$$

(0°C) and t is the difference between the temperature in °F and 32°F (the temperature in °C). The total change in length from absolute zero to the melting point has a range of approximately 2% for most substances. Typical room temperature values of a_l are given in **Table 1**.

Linear, harmonic vibration of the atoms in a solid cannot account for changes in volume, hence this must result from nonlinearity of the thermally excited vibration. The theory of E. Grüneisen takes this into account and shows the coefficient of expansion to be proportional to the constant-volume specific heat of the solid. At low temperatures (small amplitude vibration), the coefficient of expansion approaches zero.

Pure crystals may have different values of α_l along different axes, but substances such as structural steel have many crystals randomly oriented and are almost free from this effect. At certain temperatures, crystalline substances may change in lattice arrangement, and a sudden change of volume occurs at constant temperature, making α_l momentarily infinity. *See* LATTICE VIBRATIONS; SPECIFIC HEAT OF SOLIDS; THERMOCOUPLE.

TABLE 1. Temperature coefficients of linear expansion for typical substances at room temperature

Substance	Coefficient of linear expansion per °F (°C) × 10^6
Aluminum, commercial	13 (24)
Copper	9 (17)
Diamond	0.6 (1)
Glass, commercial	6 (11)
Glass, Pyrex	2 (3)
Granite	4.6 (8.3)
Ice	28 (50)
Iron	7 (12)
Invar alloy	0.5 (0.9)
Quartz, crystalline	3 (5)
Quartz, fused	0.3 (0.5)
Oak, along fiber	3 (5)
Oak, across fiber	30 (54)
Rubber, hard	44 (80)

Expansion of gases. So-called perfect gases follow the relation in Eq. (3), where p is absolute pressure, v

$$\frac{pv}{T} = \frac{R}{\text{molecular weight}} \tag{3}$$

is specific volume, T is absolute temperature, and R is a constant. The magnitude of R, the so-called gas constant, is 1544 ft-lb/(°R)(lb-mole) in the English system, or 8.3144 joules/(K)(g-mole) in the metric system. Real gases often follow this equation closely; for example, **Table 2** shows values of R at atmospheric pressure and 0°C. *See* GAS CONSTANT.

TABLE 2. Values for gas constant R at atmospheric pressure* and 32°F (0°C)

Gas	R
Air	1545
Hydrogen	1546
Nitrogen	1543
Oxygen	1544
Methane	1539

*1 atm = 10^2 kPa.

The coefficient of cubic expansion α_v is defined by Eq. (4), and for a perfect gas this is found to be $1/T$.

$$\alpha v = \frac{1}{v}\left(\frac{\partial v}{\partial t}\right)_{p=\text{const}} \tag{4}$$

The behavior of real gases is largely accounted for by van der Waals' equation (5), where $a$and b are

$$p = \frac{RT}{v - b} - \frac{a}{v^2} \tag{5}$$

constant for a given gas. When the specific volume is large, the effects of these constants are unimportant, and the real gas behaves as a perfect gas. In the regions where a and b have a dominant effect it is usually found desirable to use experimentally determined graphs or charts of properties. *See* GAS; KINETIC THEORY OF MATTER.

Expansion of liquids. For liquids, α_v is somewhat a function of pressure but is largely determined by temperature. Though α_v may often be taken as constant over a sizable range of temperature (as in the liquid expansion thermometer), generally some variation must be accounted for. For example, water contracts with temperature rise from 32 to 39°F (0 to 4°C), above which it expands at an increasing rate, as shown by the data in **Table 3**, which were taken at atmospheric pressure. One approach to this variation is to evaluate the constants α, β, and γ in Eq. (6),

$$v = v_0(1 + \alpha t + \beta t^2 + \gamma t^3) \tag{6}$$

where v_0 is the volume at 0°C (32°F), and v is the volume at temperature t in °C. Typical values of the coefficients appear in **Table 4**.

TABLE 3. Behavior of water at different temperatures

t, °F (°C)	Volume expansion, in.^{3}/oz (ml/g)
14 (−10)	1.73321 (1.00186)
32 (0)	1.73022 (1.00013)
39 (4)	1.72999 (1.00000)
50 (10)	1.73046 (1.00027)
212 (100)	1.742 (1.007)

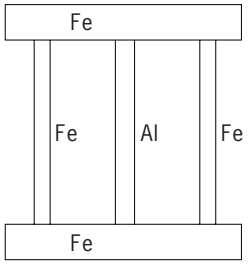
TABLE 4. Coefficients of volume expansion of gases

Liquid	$\alpha \times 10^3$	$\beta \times 10^6$	$\gamma \times 10^8$
Ethyl alcohol (99.3% by volume)	1.012	2.20	
500 atm*	0.866		
3000 atm*	0.524		
Carbon tetrachloride	1.184	0.899	1.351
Mercury	1.182	0.0078	
Petroleum	0.8994	1.396	
Water	−0.06427	8.5053	−6.7900

*1 atm = 10^2 kPa.

Thermal stresses. When a homogeneous body is subject to constant boundary loads and is raised uniformly in temperature, the stress pattern in it will not change unless its elastic properties change. In general, stresses arise if (1) the body is made up of substances having different coefficients of expansion, (2) changes of boundary dimensions are restrained, or (3) temperature distribution is not uniform. A simple example of the first case is shown in the **illustration**, where the aluminum bar, if heated, would tend to expand faster than the iron bars, thereby putting the iron in tension and the aluminum in compression. Considering the aluminum alone, its change of length would be restrained, and therefore stresses would arise in it. If the aluminum bar were replaced by an iron one and if this one alone were heated, again the other bars would be in tension and the center one in compression. More complex stress patterns may arise in continuous bodies; for example, if the bars were joined along the sides rather than at the ends, shear stresses would arise in the seams. In iron, 360 lb/in.^{2} (2.48 megapascals) tensile stress would produce the same elongation as would a temperature rise of 1.8°F (1°C).

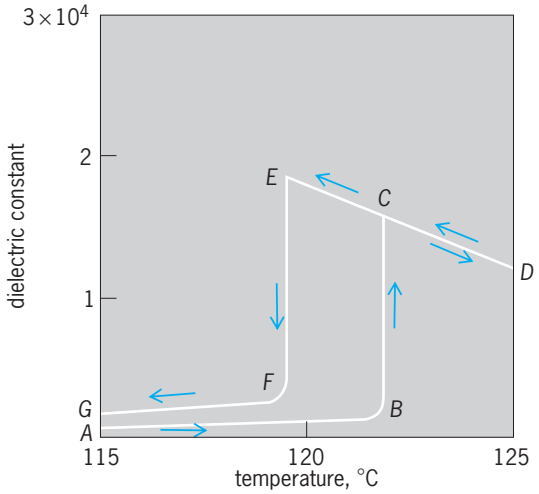
Since one source of temperature variation is the gradient necessary for heat transfer, thermal conductivity and heat capacity may both play a role in determining the stress pattern. *See* CONDUCTION (HEAT); HEAT CAPACITY; THERMAL CONDUCTION IN SOLIDS; THERMAL STRESS; THERMOMETER. Ralph A. Burton



Body composed of substances having different coefficients of expansion. Thermal stresses would arise if it were subjected to heat.

Thermal hysteresis

A phenomenon in which a physical quantity depends not only on the temperature but also on the preceding thermal history. It is usual to compare the behavior of the physical quantity while heating and the behavior while cooling through the same temperature range. The **illustration** shows the thermal hysteresis which has been observed in the behavior of the dielectric constant of single crystals of barium titanate. On heating, the dielectric constant was observed to follow the path *ABCD* and so on, cooling the path *DCEFG*. *See* FERROELECTRICS; PERMITTIVITY.



Plot of dielectric constant versus temperature for a single crystal of barium titanate. °F = (°C × 1.8) + 32. (*After M. E. Drougard and D. R. Young, Phys. Rev., 95:1152–1153, 1954*)

Perhaps the most common example of thermal hysteresis involves a phase change such as solidification from the liquid phase. In many cases these liquids can be dramatically supercooled. Elaborate precautions to eliminate impurities and outside disturbances can be instrumental in supercooling 110 to 150°F (60 to 80°C). On raising the temperature after freezing, however, the system follows a completely different path, with melting coming at the prescribed temperature for the phase change.

Solidification (or appropriate phase change) occurs by a nucleation mechanism, while melting (the reversal of phase) does not. The viable nucleus only forms below the melting temperature, being held up by the competing demands of surface and bulk free energy. Once formed, the nucleus grows rapidly to bring on the sudden precipitation of the solid phase. This rapid release of latent heat and the catastrophic nature of the solidification process illustrate the basic irreversibility

and hysteresis of thermally induced phase transformations. *See* CRYSTAL GROWTH; NUCLEATION; PHASE TRANSITIONS. H. B. Huntington; R. K. MacCrone

Thermal neutrons

Neutrons whose energy distribution is governed primarily by the kinetic energy distribution of molecules of the material in which the neutrons are found.

The molecules of the material usually have a kinetic energy distribution very close to a Maxwell-Boltzmann distribution. This distribution shows a peak at an energy equal to half the product of the temperature and the Boltzmann constant. At high energies it decreases exponentially, and at low energies it is proportional to the square root of the energy. When the material is large and very weakly absorbing, the neutron energy distribution closely approaches this maxwellian. *See* BOLTZMANN STATISTICS; KINETIC THEORY OF MATTER.

Deviations occur for the following reasons:

1. In nuclear reactors, neutrons are born at high energies and moderated by collisions with the material. These neutrons add a high-energy tail to the neutron energy distribution, which is inversely proportional to the neutron energy. Particularly when absorption is strong, the tail reaches down in energy, resulting in an increase in the peak energy of the neutron distribution relative to that in the moderator. It is common to describe this change in the peak by attributing a higher temperature to the neutrons. *See* REACTOR PHYSICS.

2. If an absorbing atom in the system has strong resonance absorption in the thermal energy range, it can create a depression in the neutron distribution around the peak resonant energy. *See* NEUTRON SPECTROMETRY.

3. If the loss of thermal neutrons is dominated by leakage rather than absorption, the neutron spectrum at low energies peaks at a lower temperature than otherwise. This phenomenon, called diffusion cooling, actually lowers the neutron temperature relative to that of the material.

The most common way of generating thermal neutrons is to allow neutrons from a source—reactor, accelerator, or spontaneous fission neutron emitter—to diffuse outward through a large block or tank of very weakly absorbing moderator. Neutrons in such a thermal column have very small tails to their spectral distribution and a temperature very close to that of the thermal-column medium.

Neutrons are sometimes scattered from cryogenic moderating materials in order to maximize the number of very low energy neutrons available for experiment. Here, although the spectrum is not quite maxwellian, it is still characterized by a temperature corresponding to the spectral peak. *See* NEUTRON. Bernard I. Spinrad

Bibliography. G. I. Bell and S. Glasstone, *Nuclear Reactor Theory*, 1970, reprint 1979; J. R. Lamarsh, *Introduction to Nuclear Reactor Theory*, 1966; M. M. R. Williams, *The Slowing Down and Thermalization of Neutrons*, 1966.

Thermal stress

Mechanical stress induced in a body when some or all of its parts are not free to expand or contract in response to changes in temperature. In most continuous bodies, thermal expansion or contraction cannot occur freely in all directions because of geometry, external constraints, or the existence of temperature gradients, and so stresses are produced. Such stresses caused by a temperature change are known as thermal stresses.

Structures subject to stress. Problems of thermal stress arise in many practical design problems, such as those encountered in the design of steam and gas turbines, diesel engines, jet engines, rocket motors, and nuclear reactors. The high aerodynamic heating rates associated with high-speed flight present even more severe thermal-stress problems for the design of spacecraft and missiles. These may be further complicated by transient heating phenomena, a complex geometry, and changes in the material properties at high temperature, as well as by the different materials employed as structure.

The study of the thermal-stress problem is not restricted to the classical case of finding the elastic thermal stresses produced by a given temperature distribution in a structure, but touches on all phases of structural design. One must also include in the analysis temperature distribution, elastic and inelastic deformation, and the behavior of materials at elevated temperature. In addition, thermal buckling, fatigue, shock and aeroelastic effects at high temperatures must be considered.

The method of solving these problems is to formulate the complete problem and then to simplify it by making assumptions based on the physical situation or on experimental data. From this simplified formulation, one obtains an analytic solution that identifies the basic parameters of the problem and then constructs charts showing how the stresses vary with these parameters. The solution is then refined by examining the simplifying assumptions and obtaining correction factors. In most cases, this procedure yields results sufficiently accurate for use in the design of structures for high-temperature applications.

Procedure at elevated temperatures. To design for thermal stress in aircraft, missiles, and spacecraft, it is necessary to know either the expected mission profiles or the flight histories of speed, altitude, angle of attack, and bank angle. From these predicted performance profiles for various vehicles, the most severe combinations of applied, thermal, and allowable stresses are selected to be used in the design.

At elevated temperatures, the designer is faced with two factors which tend to increase the weight of the structure. First, the strength and moduli of most materials decrease as temperature rises. Second, the thermal stresses may act in concert with the

applied stresses to induce a larger deflection than is estimated for the applied stresses alone. To obtain the minimum possible structural weight, both of these problems must be considered. For flights of short duration and with low heating rates, insulation and radiative cooling may help to keep the temperature of structural material low, and may reduce the effects of thermal stresses. For longer flights with low heating rates, insulation and cooling techniques can also minimize temperature in structure. For short-duration flights with high heating rates, such as those of ballistic missiles and low L/D (lift-to-drag ratio) spacecraft reentering the Earth's atmosphere, ablation cooling is the most effective means of reducing interior structural temperatures. *See* ATMOSPHERIC ENTRY.

A mechanical design that allows sufficient deflection for thermal expansion, such as a corrugated skin structure in spacecraft, can relieve some thermal stress, but maintaining structural stiffness is not always a simple problem. Corrugated webs, ribs, and clip attachments may relieve some thermal stresses and still maintain stiffness; however, expansion joints for the corrugated structure may be difficult to design. In addition, heavy spar caps or thick skin with integral stiffeners can produce high thermal stresses. One possible way to relieve thermal stress is to design structural members in the inelastic portion of the stress-strain relation to provide the deflection and to absorb the thermal expansion. This procedure involves the concept of strain design rather than stress design, but it is possible in many cases to design for the applied loads without regard to thermal stresses, and then to add the thermal strains without obtaining appreciable permanent set. In other cases, a design which allows for some permanent set may be feasible. However, if the structure is exposed to numerous temperature cycles, thermal fatigue must be considered in such designs. *See* STRESS AND STRAIN.

Allowable stresses. Combined thermal and applied stresses at elevated temperatures may occur in either the elastic or inelastic range of the stress-strain curve for structural materials. If they are in the inelastic range, the mechanical properties of the structural material must be determined for various temperatures. To design the structure to support the applied and thermal stresses, it is also necessary to know the allowable stresses of the material under various loading and temperature conditions. A honeycomb-sandwich structure can increase the strength of a material at elevated temperature for a given weight.

Creep. The creep rate of most materials increases substantially with temperature and stress. Materials used in aircraft and missile structures creep at elevated temperatures, so that the structure may undergo large deformations if the load is applied for long periods. To keep these deformations within permissible limits, it may be necessary to reduce the applied stresses on the structure. *See* CREEP (MATERIALS).

Thermal fatigue. At room temperature, fatigue limits the number of stress cycles a material can withstand before it ruptures. At high temperatures, thermal fatigue is caused by both the stress and the temperature cycles, and can result in rupture and deformation arising from creep. When a structure is restrained, it is possible at sufficiently high temperatures to produce thermal stresses in the inelastic region of the stress-strain curve that exceed yield stress and result in plastic flow or rupture.

Thermal shock. If a body is subjected to sudden heating or cooling, so that local large thermal stresses are induced, thermal shock may result. Such sudden heating may occur when a body enters the atmosphere from space. For a thin structural member, such as a plate or a shell, thermal stress can be either normal or bending or a combination of both. In general, if the temperature distributions are symmetrical with respect to axes of symmetry in the symmetrical cross section, the thermal stress induced will not be of the bending type. However, when the temperature of the structure is nonuniform, bending stresses may be induced by an unsymmetrical temperature distribution, by an unsymmetrical cross section, by the use of different materials in the structure, or by the variation of physical properties with respect to temperature. Brittle and ductile materials react differently to such thermal stresses. Because the thermal stress arises from the strain due to thermal expansion, brittle materials, which can endure little strain before rupture, may fail under the thermal shock. *See* AEROELASTICITY; SPACECRAFT STRUCTURE.

Shih-Yuan Chen

Bibliography. B. A. Boley and J. H. Weiner, *Theory of Thermal Stresses,* 1960, reprint 1997; D. Burgreen, *Elements of Thermal Stress Analysis*, 1971, reprint 1987; R. B. Hetnarski (ed.), *Thermal Stresses,* 3 vols., 1986, 1987, 1989.

Thermal wind

The difference in the geostrophic wind between two heights in the atmosphere over a given position on Earth. It approximates the variation of the actual winds with height for large-scale and slowly changing motions of the atmosphere. Such structure in the wind field is of fundamental importance to the description of the atmosphere and to processes causing its day-to-day changes. The thermal wind embodies a basic relationship between vertical fluctuations of the horizontal wind and horizontal temperature gradients in the atmosphere. This relationship arises from the combination of the geostrophic wind law, the hydrostatic equation, and the gas law.

The geostrophic wind law applies directly to steady, straight, and unaccelerated horizontal motion and is a good approximation for large-scale and slowly changing motions in the atmosphere. The law specifies that the wind direction is perpendicular to the horizontal pressure-gradient force with a direction such that lower pressure is to the left and higher pressure to the right in the Northern Hemisphere (the opposite sense applies in the Southern Hemisphere). Wind speed is proportional to the horizontal pressure-gradient force per unit mass (the speed

also depends on the Coriolis parameter, which varies only with latitude). This law, thus, relates horizontal motions to horizontal variations of pressure in the atmosphere.

The hydrostatic equation combined with the gas law relates the atmospheric pressure and temperature fields. The relationship is accurate for most atmospheric situations but not for small-scale and rapidly changing conditions such as in turbulence and thunderstorms. The equation gives the change of pressure in the vertical direction as a function of pressure and temperature. The key conclusion is that at a given level in the atmosphere the pressure change (decrease) with height is more rapid in cold air than in warm air.

The thermal wind relationship is obtained by combining the geostrophic wind law and hydrostatic equation. The geostrophic wind law provides that the vertical change in horizontal wind depends upon the vertical change in horizontal pressure-gradient force per unit mass. In turn, the hydrostatic equation shows that the vertical change in this pressure-gradient force depends primarily on the horizontal temperature gradient at the level in question. Thus, the vector of wind change with height is in a direction essentially perpendicular to the temperature gradient and parallel to the isotherms of temperature on a horizontal surface with the cold air to the left and warm air to the right in the Northern Hemisphere (the opposite sense applies in the Southern Hemisphere). The magnitude of the change is proportional to the temperature gradient. The thermal wind relationship explains the tendency for westerly winds to increase with height in the middle latitude troposphere, reflecting the general fact that tropospheric temperatures are cooler in the polar regions than in the tropics. *See* ATMOSPHERIC GENERAL CIRCULATION; CORIOLIS ACCELERATION; GEOSTROPHIC WIND; HYDROSTATICS; TROPOSPHERE.

David D. Houghton

Bibliography. D. G. Andrews, J. R. Holton, and C. B. Conway (eds.), *Middle Atmosphere Dynamics*, 1987; D. L. Hartmann, *Global Physical Climatology*, 1994; J. R. Holton, *An Introduction to Dynamic Meteorology*, 3d ed., 1992; D. D. Houghton (ed.), *Handbook of Applied Meteorology*, 1985; C. Riegel, *Fundamentals of Atmospheric Dynamics and Thermodynamics*, 1992; G. T. Trewartha and L. H. Horn, *Introduction to Climate*, 5th ed., 1980.

Thermionic emission

The emission of electrons into vacuum by a heated electronic conductor. In its broadest meaning, thermionic emission includes the emission of ions, but since this process is quite different from that normally understood by the term, it will not be discussed here. Thermionic emitters are used as cathodes in electron tubes and hence are of great technical and scientific importance. Although in principle all conductors are thermionic emitters, only a few materials satisfy the requirements set by practical applications. Of the metals, tungsten is an important practical thermionic emitter; in most electron tubes, however, the oxide-coated cathode is used to great advantage.

Richardson-Dushman equation. The thermionic emission of a material may be measured by using the material as the cathode in a vacuum tube and collecting the emitted electrons on a positive anode. If the anode is sufficiently positive relative to the cathode, space charge (a concentration of electrons near the cathode) can be avoided and all electrons emitted can be collected; the saturation thermionic current is then measured. Actually, the emission current increases slightly with increasing field strength at the cathode, and in order to obtain the true saturation current one should extrapolate to zero applied field. *See* SCHOTTKY EFFECT.

The emission current density J increases rapidly with increasing temperature; this is illustrated by the following approximate values for tungsten:

T (K)	1000	2000	2500	3000
J (A/cm^2)	10^{-15}	10^{-3}	0.3	15

The temperature dependence of J is given by Eq. (1), the Richardson-Dushman (or Richardson)

$$J = AT^2 e^{-(\phi/kT)} \tag{1}$$

equation. Here A is a constant, k is Boltzmann's constant (=1.38 × 10^{-23} joule per degree), and ϕ is the work function of the emitter. The work function has the dimensions of energy and is a few electronvolts for thermionic emitters.

The temperature dependence of J is essentially determined by the exponential factor, since its temperature dependence predominates strongly over that of the factor T^2. Both A and ϕ may be obtained

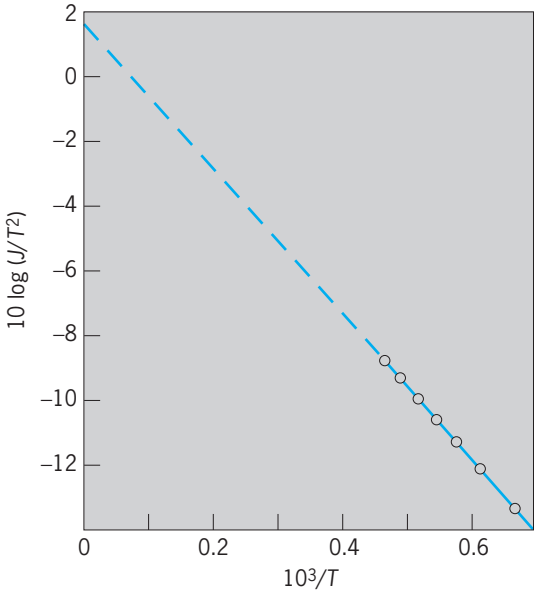


Fig. 1. Richardson-Dushman plot for tungsten, an important thermionic emitter. (*After G. Herrmann and S. Wagener, The Oxide-Coated Cathode, vol. 2, Chapman and Hall, 1951*)

experimentally by plotting the logarithm of J/T^2 versus $1/T$, as for tungsten in **Fig. 1**. The Richardson-Dushman formula can be derived for metals and semiconductors on the basis of relatively simple physical models.

Metals. According to quantum theory the electrons in a free atom occupy a set of discrete energy levels. When atoms are brought together to form a solid, these energy levels broaden into energy bands; the broadening is a result of the perturbing fields produced by neighboring atoms on the electrons and is most pronounced for the outer or valence electrons. In a metal, the perturbing influence on the valence electrons is so strong that they can no longer be associated with particular atoms but must be considered as moving freely throughout the crystal. These so-called free, or conduction, electrons are responsible for the high electrical and thermal conductivity of metals and also for the thermionic emission. *See* BAND THEORY OF SOLIDS; FREE-ELECTRON THEORY OF METALS.

The free electrons may be assumed to move in an approximately constant potential as indicated in **Fig. 2**. The bottom of the box corresponds to the energy of a conduction electron at rest in the metal; the "vacuum" level represents the energy of an electron at rest in free space. According to quantum mechanics, the electrons in this model can assume only particular states of motion which correspond to a set of very closely spaced energy levels. The probability for a given state to be occupied depends on the energy E of the state and on the absolute temperature T in accordance with Eq. (2), the so-called

$$F(E) = \frac{1}{1 + \exp\left[(E - E_F)/kT\right]} \qquad (2)$$

Fermi-Dirac distribution function. The quantity E_F is called the Fermi energy; it is determined by the number of electrons per unit volume in the metal and is of the order of a few electronvolts. Since kT at room temperature ($T = 300$ K) is only about 0. 025 eV, $E_F > kT$ for all temperatures below the melting point of metals. Note that for $T = 0$, $F(E) = 1$ for $E < E_F$, and $F(E) = 0$ for $E > E_F$. Hence, at absolute zero all energy levels up to E_F are occupied by electrons, whereas those above E_F are empty. For temperatures different from zero, some electrons have energies larger than E_F and the thermionic emission is due to those electrons in the "tail" of the Fermi distribution for which the energy lies above the vacuum level in Fig. 1. Note that when $E = E_F$, $F(E) = 0.5$; that is, the Fermi energy corresponds to those states for which the probability of being occupied is equal to 0.5.

When these ideas are put in a quantitative form, one arrives at the Richardson-Dushman equation with the specific value $A = 120$ A/cm^2 (if one takes into account reflection of electrons against the surface potential barrier, the theoretical value of A is <120 A/cm^2).

Experiments by M. N. Nichols in 1940 and by G. F. Smith in 1954 on single crystals of tungsten have shown that experimental values for A and ϕ depend on the crystallographic plane from which the emission is measured; values for A (in A/cm^2) and ϕ (in electronvolts) for two crystallographic directions are given in the **table**. For polycrystalline metals, the experimental values for A and ϕ are thus average values for the particular specimen.

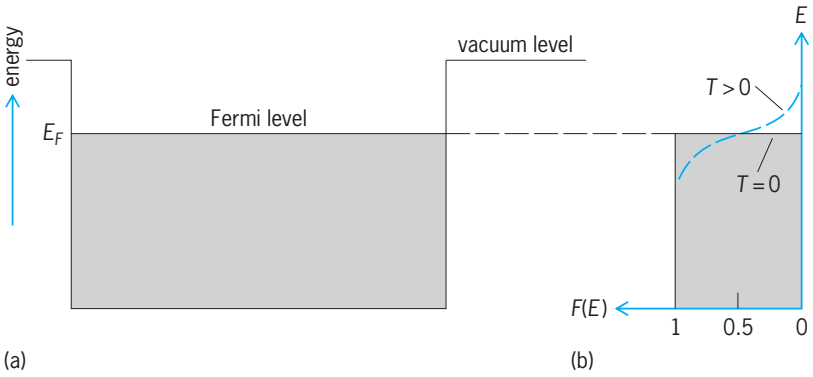


Fig. 2. Free electrons are assumed to move in approximately constant potential. (*a*) Occupation of electron states between the bottom of the conduction band and the Fermi level of a metal is indicated for $T = 0$ by the shaded area. (*b*) Fermi distribution function is represented schematically for $T = 0$ and for $T > 0$.

Experimental values for single crystals of tungsten

	Nichols		Smith	
Direction	A	ϕ	A	ϕ
(111)	35	4.39	52	4.38
(100)	117	4.56	105	4.52

Semiconductors. For semiconductors, the thermionic emission is also due to the escape of electrons which have energies above the vacuum level. The theory leads to the Richardson-Dushman formula, as it does for metals. The work function measures again the difference between the Fermi level of the semiconductor and the vacuum level. *See* SEMICONDUCTOR.

A. J. Dekker

Thermionic power generator

A device for converting heat into electricity through the use of thermionic emission and no working fluid other than electric charges. An elementary thermionic generator, or thermionic converter, consists of a hot metal surface (emitter) separated from a cooler electrode (collector) by an insulator seal (**Fig. 1**). The interelectrode gap is usually a fraction of a millimeter in width. The hermetic enclosure contains a small amount of an easily ionizable gas, such as cesium vapor maintained by a liquid-cesium reservoir. In some experimental devices, the enclosure may be evacuated.

Electrons evaporated from the emitter cross the interelectrode gap, condense on the collector, and are returned to the emitter via the external electrical load circuit. The thermionic generator is essentially a heat engine utilizing an electron gas as the

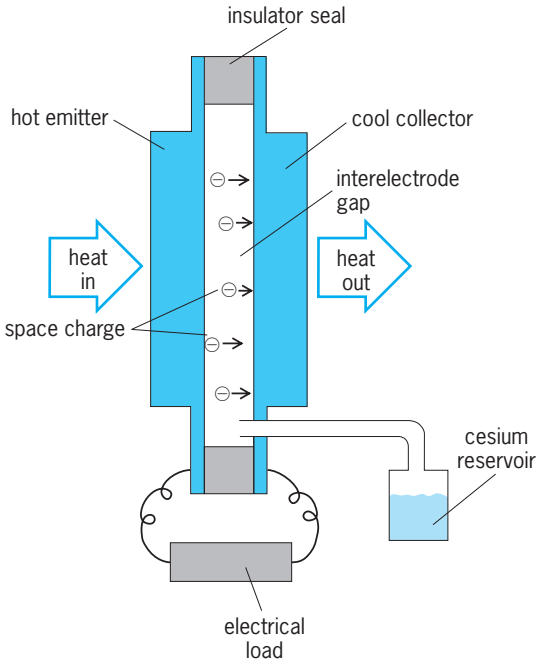


Fig. 1. Diagram of thermionic converter.

working fluid. The temperature difference between the emitter and the collector drives the electron current.

Thermionic generators are characterized by high operating temperatures, typically emitter temperatures between 1600 and 2500 K (2420 and 4040°F) and collector temperatures ranging from 800 to 1100 K (980 to 1520°F); low output voltage (approximately 0.5 V per converter); high current density (around 5–10 A/cm^2); and high conversion efficiency (about 10–15%). These characteristics, especially the relatively high heat-rejection temperature, make the thermionic generator attractive for producing electric power in space applications with nuclear-reactor or radioisotope energy sources. The high electrode temperatures make thermionic generators also attractive as topping units for steam power plants, and for the cogeneration of electricity in combination with heat for intermediate-temperature industrial processes. Topping units increase the overall system efficiency. *See* COGENERATION; NUCLEAR BATTERY.

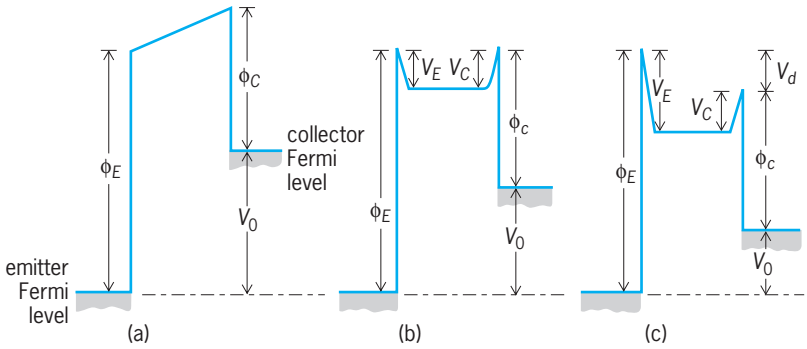


Fig. 2. Motive diagrams of a thermionic converter operating in (*a*) ideal, (*b*) unignited, and (*c*) ignited modes.They give the spatial variation of the energy of an electron from emitter to collector. Here, ϕ_E and ϕ_C are emitter and collector work functions, V_0 is the output voltage, V_E and V_C are voltages across sheath regions at emitter and collector, and V_d is the arc drop.

Vacuum devices. Because the energy of the electrons in the collector is greater than that in the emitter (**Fig. 2***a*), the collected electrons perform work as they flow back to the emitter through the electrical load. The output voltage V_0 is given by the difference in the Fermi levels between the emitter and the collector. The Fermi level is a characteristic energy of electrons in a material. It corresponds to the highest energy of the electrons in the material at zero temperature. The difference between the motive just outside a surface and the Fermi level is a property of the surface and is called the work function. The emitter work function ϕ_E is given by the difference of the motive of a point just outside the emitter and the emitter Fermi level. Likewise, the collector work function ϕ_C is given by the difference between the motive of a point just outside the collector and the collector Fermi level. *See* FREE-ELECTRON THEORY OF METALS; WORK FUNCTION (ELECTRONICS).

The electron emission from the emitter or collector electrodes is given by the Richardson equation (1), where J is the current density in am-

$$J = 120T^2 e^{-(11604\phi/T)} \tag{1}$$

peres per square centimeter, ϕ is the work function of the electrode in electronvolts, and T is the electrode temperature in kelvins. *See* THERMIONIC EMISSION.

The output power density of the converter is the product of V_0 times the load current density J that flows through the converter. Usually J is substantially less than the emitter current density J_E. In the ideal mode (Fig. 2*a*), J is given by Eq. (2) for $V_0 < \phi_E - \phi_C$;

$$J = J_E e^{-[11604(V_0+\phi_C-\phi_E)/T_E]} \tag{2}$$

for $V_0 > \phi_E - \phi_C$, $J = J_E$. In Eq. (2), T_E is the emitter temperature in kelvins, and it is assumed that the collector temperature T_C is low enough that back emission from the collector is negligible. The output voltage V_0 can be varied between zero and its open circuit value by changing the load resistance, giving rise to the ideal-mode current density-voltage characteristic (**Fig. 3**).

The ideal mode (Fig. 2*a*) corresponds to an evacuated interelectrode spacing of 0.01 mm or less. Small thermionic diodes with such spacings have been built. However, the difficulty of maintaining the necessary close spacing over large areas at high temperatures makes the vacuum diode mostly of academic interest.

Cesium-vapor devices. The problems of extremely close spacing can be circumvented by introducing low-pressure cesium vapor between the emitter and the collector. The cesium serves two functions. First, it adsorbs on the electrode surfaces and reduces their work functions. Each work function is reduced to a value that depends on the ratio of the electrode temperature and the cesium reservoir temperature. Second, the cesium supplies positive ions to neutralize

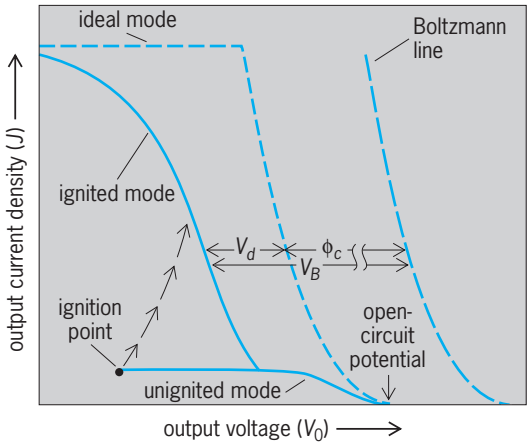


Fig. 3. Current density-voltage characteristics of a thermionic converter. V_B is the barrier index; other symbols are as in Fig. 2.

the electron space charge. The ions are produced by surface ionization and electron-impact ionization in the interelectrode gap. Surface ionization occurs when the cesium vapor contacts the hot emitter. The ionization efficiency is high when the emitter work function is comparable to the 3.89-eV ionization potential of cesium. The addition of barium vapor to the cesium vapor may lead to better converter performance.

Unignited mode. A thermionic converter in which the predominant source of cesium ions is surface ionization operates at low cesium pressure, less than about 10 pascals or 0.1 torr, and is said to operate in the unignited, lower, extinguished, or Knudsen mode (Figs. 2*b* and 3). Sheath regions at the emitter and collector couple the neutral plasma (which has essentially equal ion and electron concentrations) to the adjacent electrodes. Unfortunately, in the unignited mode, a converter provides useful power densities at emitter temperatures above 2500 K (4040°F), a region for which there are few practical energy sources.

Ignited mode. In order to obtain practical power densities at more moderate emitter temperatures, say 1600 to 2000 K (2420 to 3140°F), the cesium pressure must be increased to the order of 10^2 Pa (1 torr). Then cesium adsorption on the emitter reduces its work function so that the current density is greater than 5 A/cm^2, and the converter is said to operate in the ignited mode (Figs. 2*c* and 3). To obtain this mode, not only is the cesium pressure increased but the output voltage is reduced along the unignited-mode characteristic curve until the ignition point is reached and electron impact ionization in the interelectrode space results in a discharge which triggers the thermionic converter into the ignited mode. It is possible to operate back and forth along the ignited-mode curve. However, if the output potential is increased to the intersection with the unignited mode, the discharge is extinguished, and the ignited mode can only be reestablished by reducing the output voltage along the unignited-mode curve to the ignition point.

The electrons are scattered many times by the cesium as they cross to the collector, resulting in resistive and current losses in the plasma. There is an additional loss due to the energy required to ionize the cesium atoms. The sum of these losses is usually called the arc drop and is denoted V_d. It represents the loss in output voltage that must be supplied to the plasma in order to operate the converter in the ignited mode. Thus far, this mode has been the only practical means of operating a thermionic generator. For good performance, it is clear that V_d and ϕ_C must be small. *See* ELECTRICAL CONDUCTION IN GASES.

Barrier index. The barrier index V_B is defined by Eq. (3). This index serves as an inverse figure of merit of the thermionic converter performance; the lower the value of V_B, the higher the converter performance. Reductions in V_B can be translated into higher efficiency at a given temperature or a lower emitter temperature at a given efficiency. Present converters have V_B values between 1.9 and 2.2 eV. Operationally, V_B is defined as the minimum potential difference between the ignited-mode characteristic and the Boltzmann line. The Boltzmann line represents the ideal current density-voltage characteristic assuming zero arc drop, zero collector work function, and zero collector temperature.

$$V_B = V_d + \phi_C \tag{3}$$

Elias P. Gyftopoulos; George N. Hatsopoulos

Bibliography. S. W. Angrist, *Direct Energy Conversion,* 4th ed., 1982; F. G. Baksht et al., *Thermionic Converters and Low-Temperature Plasma*, 1978; R. Decher, *Direct Energy Conversion: Fundamentals of Electric Power Production*, 1997; G. N. Hatsopoulos and E. P. Gyftopoulos, *Thermionic Energy Conversion*, vol. 1, 1974, vol. 2, 1979; *Proceedings of the 11th Symposium on Space Power and Propulsion*, Albuquerque, 1994; N. S. Rasor, *Applied Atomic Collision Physics*, vol. 5, 1982.

Thermionic tube

An electron tube that relies upon thermally emitted electrons from a heated cathode for tube current. Thermionic emission of electrons means emission by heat. In practical form an electrode, called the cathode because it forms the negative electrode of the tube, is heated until it emits electrons. The cathode may be either a directly heated filament or an indirectly heated surface. With a filamentary cathode, heating current is passed through the wire, which either emits electrons directly or is covered with a material that readily emits electrons. Some typical filament structures are shown in **Fig. 1**. Filaments of tungsten or thoriated tungsten are commonly used in high-power transmitting tubes where their ruggedness and ability to withstand high voltages are essential. Oxide-coated filaments are used in a few small high-voltage rectifier tubes.

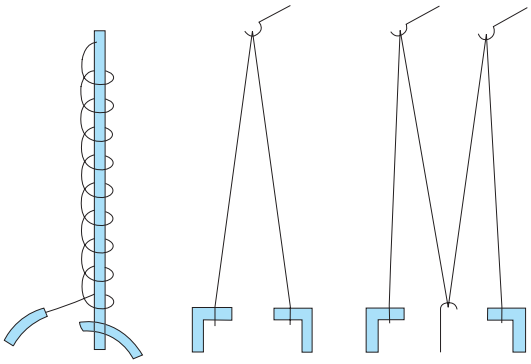


Fig. 1. Typical filamentary cathode structures for thermionic tubes.

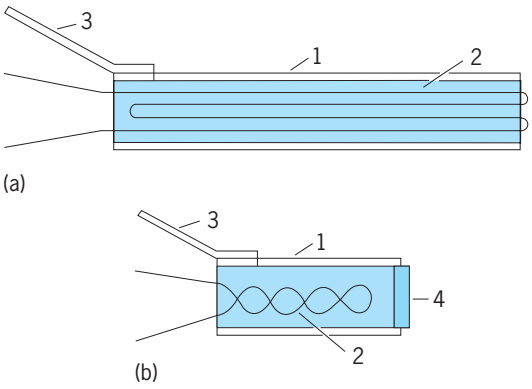


Fig. 2. Cathodes. (*a*) Receiving-tube cathode; 1, cathode sleeve, oxide-coated on exterior; 2, folded heater, insulated with refractory oxide; 3, cathode tab, for electrical connection. (*b*) Kinescope cathode: 1, cathode sleeve; 2, heater, insulated with refractory oxide; 3, cathode tab, for electrical connection; 4, emitting "button," oxide-coated on right surface.

Indirectly heated cathodes have a filament, commonly called the heater, located within the cathode electrode to bring the surface of the cathode to emitting temperature. Some common forms are shown in **Fig. 2**. They are usually coated with barium-strontium oxide, on the periphery in receiving tubes and on the end in kinescopes. Because the emitting surface carries no heating current, there is no voltage drop along the surface. Hence such cathodes are usually known as equipotential cathodes. The high emission capability, the equipotential surface, and the favorable geometry of these cathodes make possible the close-spaced tube structures that lead to the high transconductances required in modern applications. Hence, oxide-coated equipotential cathodes are used in almost all receiving and medium-power transmitting tubes. They are also used in some high-power pulsed transmitting tubes, where the remarkable ability of the oxide cathode to emit very high current densities (tens of amperes per square centimeter, for microsecond periods at low repetition rates) is exploited. The majority of all vacuum tubes are thermionic tubes. It is possible to make so-called cold-cathode tubes, but they tend to be unstable in vacuum and find their main application in gas tubes, not vacuum tubes. *See* ELECTRON TUBE; GAS TUBE; THERMIONIC EMISSION; VACUUM TUBE.

Leon S. Nergaard

Thermistor

An electrical resistor with a relatively large temperature coefficient of resistance. Thermistors are useful for measuring temperature and gas flow or wind velocity. Often they are employed as bolometer elements to measure radio-frequency, microwave, and optical power. They also are used as electrical circuit components for temperature compensation, voltage regulation, circuit protection, time delay, and volume control. A common type of thermistor is a semiconducting ceramic composed of a mixture of several metal oxides. Metal electrodes or wires are attached to the ceramic material so that the thermistor resistance can be measured conveniently. The temperature coefficient of resistance is negative for these thermistors. Other types can have either negative or positive temperature coefficients. *See* ELECTRICAL RESISTIVITY; VOLTAGE REGULATOR; VOLUME CONTROL SYSTEMS.

At room temperature the resistance of a thermistor may typically change by several percent for a variation of 1°C of temperature, but the resistance does not change linearly with temperature. The temperature coefficient of resistance of a thermistor is approximately equal to a constant divided by the square of the temperature in kelvins. The constant is equal to several thousand kelvins and is specified for a given thermistor and the temperature range of intended use.

The electrical and thermal properties of a thermistor depend upon the material composition, the physical dimensions, and the environment provided by the thermistor enclosure. Thermistors range in form from small beads and flakes less than 25 micrometers (10^{-3} in.) thick to disks, rods, and washers with centimeter dimensions. The small beads are often coated with glass to prevent changes in composition or encased in glass probes or cartridges to prevent damage. Beads are available with room-temperature resistances ranging from less than 100 Ω to tens of megohms, and with time constants that can be less than a second. Large disks and washers have a similar resistance range and can have time constants of minutes. *See* TIME CONSTANT.

Temperature measurement. An instrument can be made to measure temperature by connecting a battery, a thermistor, and a current meter in series. The current indicated by the meter is determined by the thermistor resistance and therefore is an indication of the temperature. A thermistor across the terminals of an ohmmeter or a Wheatstone bridge will also provide an indication of the temperature. By careful design, a thermistor can be used to measure temperature changes of less than 10^{-3} °C. *See* TEMPERATURE MEASUREMENT; THERMOMETER.

Velocity or flow measurement. Wind velocity or gas flow can be measured by increasing the current in

the thermistor of a temperature-measuring instrument so that the thermistor temperature is considerably higher than the ambient temperature. The thermistor temperature then will depend on how fast the thermistor can dissipate the heat and therefore how fast the air is flowing over the thermistor surface. If the ambient temperature is varying, a second flow meter with the thermistor shielded from air currents can be used to determine the effect of temperature variations alone. *See* FLOW MEASUREMENT; VELOCIMETER.

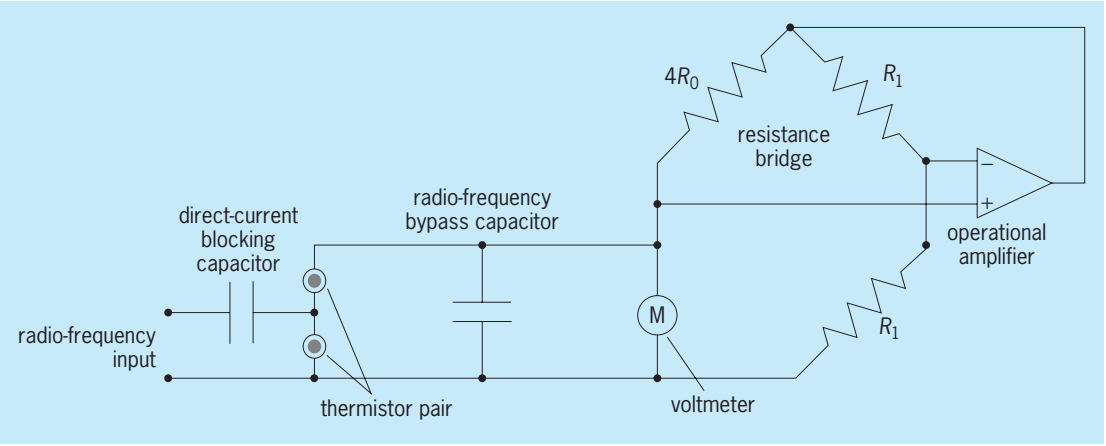
Level measurements. The same type of instrument used for measuring gas flow can be used to sense liquid level. The thermistor is placed above the surface of the liquid. If the liquid level rises to contact the thermistor, heat will be dissipated rapidly from the thermistor and the thermistor temperature will decrease, giving a decrease in the electric current. With great care, thermistors can be made sufficiently sensitive to detect objects very close to, but not touching, the thermistor. *See* LEVEL MEASUREMENT.

Radio-frequency power measurement. One of the most accurate techniques for measuring radio-frequency and microwave power is to use thermistors to measure a substituted change in direct-current power. Initially, direct-current power is dissipated in the thermistor, and the amount of power needed to achieve a specified resistance, and therefore temperature, is recorded. The thermistor is then used to dissipate both radio-frequency and direct-current power. In order to achieve the same resistance, less direct-current power is needed, and this decrease in direct-current power is approximately equal to the radio-frequency power dissipated by the thermistor. The technique is very effective because direct-current power can be measured very accurately.

The direct-current portion of the circuit is a bridge (see **illustration**). The radio-frequency signal is incident from the left side of the illustration through a coaxial transmission line with a characteristic impedance of R_0. Capacitors are used to isolate the radio-frequency and direct-current portions of the circuit except for two identical thermistors, which are part of both portions. The two thermistors in series constitute one arm of the bridge circuit. The operational amplifier provides feedback that keeps the voltage difference between its input terminals (+ and − in the illustration) very small. This is accomplished when the resistor ratios on the left and right are identical. Since the two resistors on the right have the same value R_1, the combined thermistor resistance will be made equal to $4R_0$. Thus, the individual thermistors have a resistance of $2R_0$. The thermistors appear in parallel with the radio-frequency circuit, and their combined resistance matches the characteristic impedance of the transmission line, which maximizes the absorbed radio-frequency power. The voltage V across the thermistor pair is recorded by the voltmeter without any radio-frequency input. The radio-frequency signal is then turned on and the voltage recorded again. The power dissipated in the thermistor pair is $V^2/4R_0$. The difference in power dissipated by the thermistor pair in the two cases is calculated, and this is approximately equal to the radio-frequency power. *See* BRIDGE CIRCUIT; TRANSMISSION LINES.

There are small differences between the radio-frequency power and the change in direct-current power due to radio-frequency losses in the transmission line, differences in the current distributions, and reactances in the circuit. This difference is characterized by an effective efficiency that is measured with other techniques. Coaxial thermistor sensors, also known as bolometer mounts, typically have effective efficiencies of greater than 95%. Similar devices using single thermistors are also available with rectangular waveguides. *See* BOLOMETER; WAVEGUIDE.

For radio-frequency measurements, because the thermistor resistance is held at a fixed value, its nonlinear characteristics do not affect the measurement, and the radio-frequency power measurement is very linear over most of its operating range, which is generally up to about 10 milliwatts. A slight nonlinearity does arise at higher power levels because the two thermistors are not exactly the same. The bridge will keep their series resistance equal to $4R_0$, but their parallel resistance will no longer



Bridge circuit for measuring radio-frequency power by use of a pair of thermistors.

match the transmission line impedance of R_0. *See* ELECTRIC POWER MEASUREMENT; MICROWAVE POWER MEASUREMENT.

Thomas P. Crowley

Bibliography. A. Fantom, *Radio Frequency and Microwave Power Measurement*, Peter Peregrinus, London, 1990; C. J. Kaiser, *The Resistor Handbook*, 2d ed., Saddleman Press, 1994; E. D. Macklen, *Thermistors*, Electrochemical Publications, 1979; F. Zandman, P.-R. Simon, and J. Szwarc, *Resistor Theory and Technology*, Vishay Intertechnology, Malvern, PA, 2001.

Thermoacoustics

The study of phenomena that involve both thermodynamics and acoustics. A sound wave in a gas is usually regarded as consisting of coupled pressure and displacement oscillations, but temperature oscillations accompany the pressure oscillations. When there are spatial gradients in the oscillating temperature, oscillating heat flow also occurs. The combination of these four oscillations produces a rich variety of thermoacoustic effects. *See* ACOUSTICS; OSCILLATION; SOUND; THERMODYNAMIC PRINCIPLES.

Although the oscillating heat transfer at solid boundaries does contribute significantly to the dissipation of sound in enclosures such as buildings, thermoacoustic effects are usually too small to be obviously noticeable in everyday life. For example, the amplitude of the temperature oscillation in conversational levels of sound is only about 0.0001°C (0.0002°F). However, thermoacoustic effects in intense sound waves inside suitable cavities can be harnessed to produce extremely powerful pulsating combustion, thermoacoustic refrigerators, and thermoacoustic engines. Gas compressibility and inertia, heat transport, and thermal expansion and contraction are important in all of these systems.

Pulsating combustion. Oscillations can occur whenever combustion takes place in a cavity. In industrial equipment and residential appliances, these oscillations are sometimes encouraged in order to stir or pump the combustion ingredients, while in rocket engines such oscillations must usually be suppressed because they can damage the rocket structure. The oscillations occur spontaneously if the combustion progresses more rapidly or efficiently during the compression phase of the pressure oscillation than during the rarefaction (expansion) phase—the Rayleigh criterion. *See* COMBUSTION; GAS DYNAMICS.

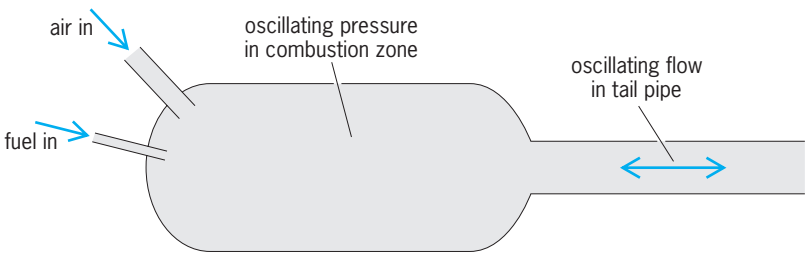


Fig. 1. Acoustic resonance in a pulse combustor. The mass of gas in the tail pipe bounces against the "spring" of compressibility in the combustion zone, forming a resonant oscillator. The oscillations are driven by periodic, pressure-enhanced combustion.

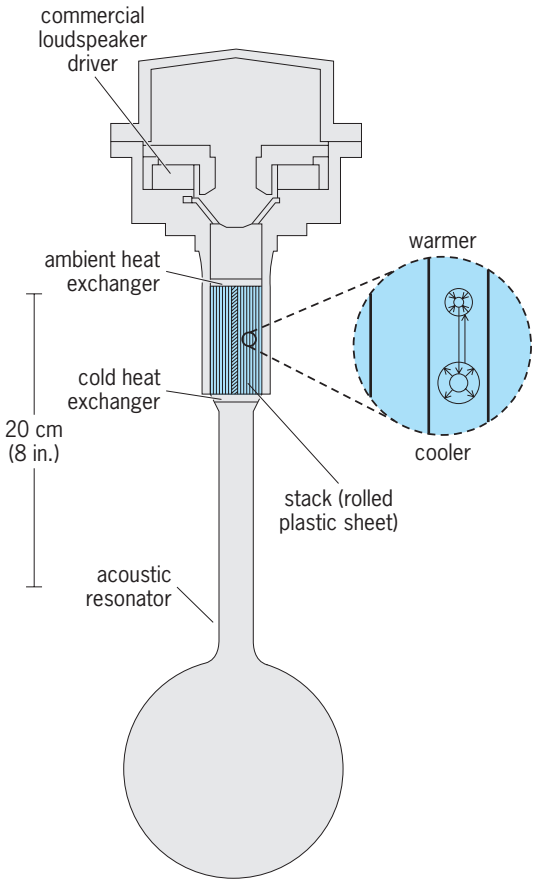


Fig. 2. An early standing-wave thermoacoustic refrigerator that cooled to −60°C (−76°F). Heat is carried up the temperature gradient in the stack. At the right is a magnified view of the oscillating motion of a typical parcel of gas. The volume of the parcel depends on its pressure and temperature. (*After T. J. Hofler, Thermoacoustic Refrigerator Design and Performance, Ph.D. thesis, University of California at San Diego, 1996*)

The geometry of the cavity determines the oscillation frequency, just as the length of an organ pipe determines its pitch (**Fig. 1**). Stored resonance energy shifts back and forth between kinetic energy of moving gas in the tail pipe and compressive energy of pressurized gas in the combustion zone, much as the energy in a pendulum shifts back and forth between kinetic energy and gravitational potential energy. At one time of the oscillation, the air is motionless and the combustion zone is pressurized above atmospheric pressure. Combustion is enhanced at this time, and all the resonance energy is stored in the compressibility of the pressurized air. A quarter cycle later, after the pressure has accelerated the air in the tail pipe rightward, the air velocity in the tail pipe reaches its largest (rightward) value when the pressure in the combustion zone reaches its ambient value, so all the resonance energy is stored kinetically in the moving air. Next, inertia of the moving air keeps it moving rightward, pulling air out of the combustion zone and lowering the pressure there, so that the air soon comes to rest with the combustion

zone below ambient pressure, shifting the resonance energy back to compressibility and suppressing the combustion. This low pressure then pulls air back into the tail pipe, so that after another quarter cycle the air is moving leftward most rapidly, storing kinetic energy, and the combustion zone has been refilled to ambient pressure. Finally, inertia keeps the air moving leftward, pressurizing the combustion zone, so that the system comes to rest again at the starting condition. *See* INERTIA.

Thermoacoustic refrigerators. Thermoacoustic refrigerators use acoustic power to pump heat from a low temperature to ambient temperature (**Fig. 2**). The heat-pumping mechanism takes place in the pores of a structure called a stack. As a typical parcel of the gas oscillates along a pore, it experiences changes in temperature. Most of the temperature change comes from adiabatic compression and expansion of the gas by the sound pressure, and the rest is a consequence of the local temperature of the solid wall of the pore. A thermodynamic cycle results from the coupled pressure, temperature, position, and heat oscillations. At the topmost position of the parcel, it dumps heat to the pore wall, since the parcel temperature was raised above the local pore temperature by adiabatic compression by the acoustic wave. Similarly, at its lowermost position, the parcel absorbs heat from the pore, since the parcel temperature is below the local stack temperature. Thus, each parcel moves a little heat a small distance from the bottom toward the top along the stack, a small distance up the temperature gradient, during each cycle of the acoustic wave. The overall effect, much as in a bucket brigade, is the net transport of heat from the cold heat exchanger to room temperature. *See* ADIABATIC PROCESS; SOUND PRESSURE; THERMODYNAMIC CYCLE; THERMODYNAMIC PROCESSES.

The extrema in pressure and gas motion are in phase in this thermoacoustic refrigerator and in the pulsating combustion discussed above. A wave with such time phasing is called a standing wave.

Thermoacoustic engines. While standing-wave thermoacoustic systems have matured only recently, Stirling engines and refrigerators have a long, rich history. Recently, new insights have resulted from applying thermoacoustics to Stirling systems, treating them as traveling-wave thermoacoustic systems in which the extrema in pressure and gas motion are approximately 90° out of phase in time. In the thermoacoustic-Stirling engine, the thermodynamic cycle is accomplished in a traveling-wave acoustic network, and acoustic power is produced from heat with an efficiency of 30% (**Fig. 3**). The wave circulating around the torus causes the gas in the regenerator to experience oscillating pressure and oscillating motion, with traveling-wave time phasing. Excellent thermal contact between the typical parcel of gas and the wall of its pore ensures that the gas temperature always matches the local solid temperature. Hence, downward motion of the parcel causes thermal expansion, and upward motion causes thermal contraction. The time phasing between this motion and the oscillating pressure ensures that the thermal expansion occurs at high pressure and the thermal contraction occurs at low pressure, so that the parcel does net work during each cycle. Since this

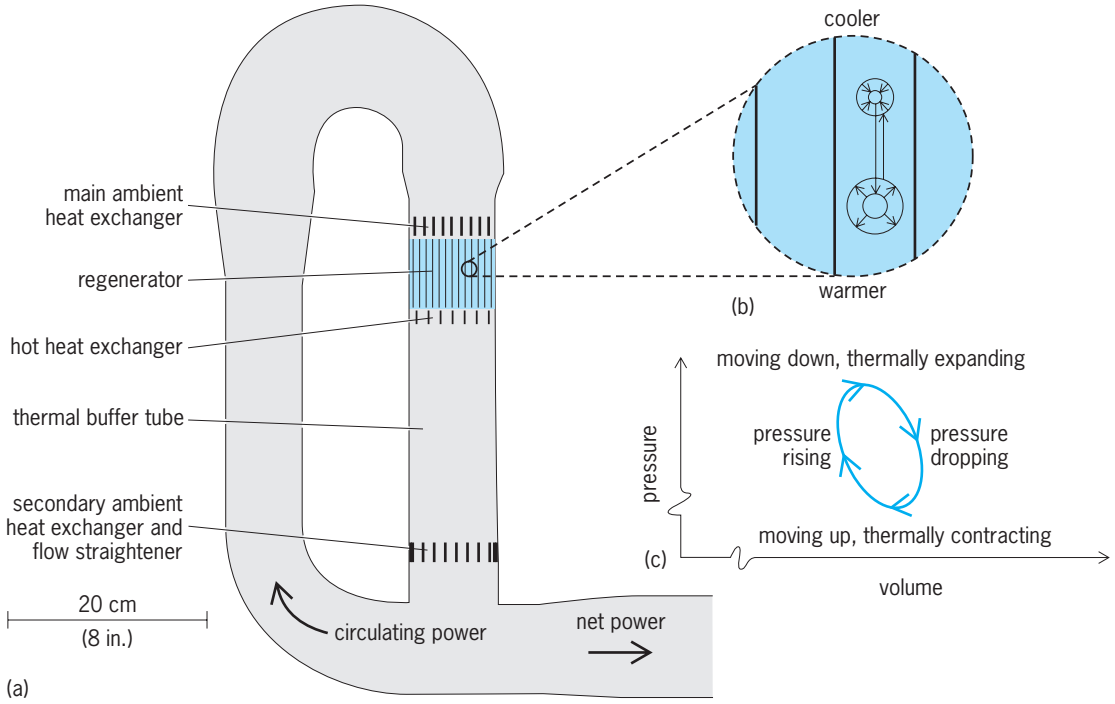


Fig. 3. Thermoacoustic-Stirling hybrid engine. (*a*) A traveling wave runs clockwise around the toroidal loop. Its acoustic power is amplified in the regenerator, so that net power up to 1 kW (1.3 horsepower) can be extracted at the tee and delivered to a load. (*b*) Magnified view of the oscillating motion of a typical parcel of gas. (*c*) Thermodynamic cycle of the typical parcel of gas. (*After S. Backhaus and G. Swift, Thermoacoustic-Stirling heat engine: Detailed study, J. Acous. Soc. Amer., 107:3148–3166, 2000*)

net work is produced at the acoustic frequency, it increases the acoustic power of the wave passing through the regenerator, so net power is available for extraction elsewhere. *See* STIRLING ENGINE; TORUS; WAVE (PHYSICS).

Current developments. Typical engines and refrigerators have crankshaft-coupled pistons or rotating turbines, whereas thermoacoustic systems often have no moving parts or flexing moving parts (as in a loudspeaker), with no sliding seals. Such systems appear attractive because of their elegance, reliability, and low cost. Some product-oriented research and development is under way in pulsating combustion and thermoacoustics, and many academic groups are studying nonlinear behavior at a fundamental level. *See* NONLINEAR ACOUSTICS.

Greg Swift

Bibliography. S. Backhaus and G. Swift, Thermoacoustic-Stirling heat engine: Detailed study, *J. Acous. Soc. Amer.*, 107:3148–3166, 2000; T. J. Hofler, *Thermoacoustic Refrigerator Design and Performance*, Ph.D. thesis, University of California at San Diego, 1986; L. E. Kinsler et al., *Fundamentals of Acoustics*, 4th ed., 1999; A. J. Organ, *Thermodynamics and Gas Dynamics of the Stirling Cycle Machine*, 1992; J. W. S. Rayleigh, *The Theory of Sound*, vol.2, 1976; G. Walker, *Cryocoolers*, 1983.

Thermochemistry

A branch of physical chemistry concerned with the absorption or evolution of heat that accompanies chemical reactions. Closely related topics are the latent heat associated with a change in phase (crystal, liquid, gas), the chemical composition of reacting systems at equilibrium, and the electrical potentials of galvanic cells. Thermodynamics provides the link among these phenomena.

A knowledge of such heat effects is important to the chemical engineer for the design and operation of chemical reactors, the determination of the heating values of fuels, the design and operation of refrigerators, the selection of heat storage systems, and the assessment of chemical hazards. Thermochemical information is used by the physiologist and biochemist to study the energetics of living organisms and to determine the calorific values of foods. Thermochemical data give the chemist an insight to the energies of, and interactions among, molecules. *See* CHEMICAL EQUILIBRIUM; CHEMICAL REACTOR; PHASE EQUILIBRIUM; POTENTIALS; REFRIGERATOR.

Thermodynamic principles. The first law of thermodynamics expresses the principle of conservation of energy. When a closed system changes from an initial state to a final state, its internal energy, U, changes by the amount shown in Eq. (1), where q is the heat

$$U(\text{final}) - U(\text{initial}) = \Delta U = q + w \qquad (1)$$

energy transferred to the system from the outside and w is the work done on the system by external forces. A positive sign of q, w, or ΔU means energy is transferred to the system, while a negative sign means energy is removed. Some authors use the opposite sign convention for w. *See* CONSERVATION OF ENERGY; THERMODYNAMIC PRINCIPLES.

Internal energy. Internal energy (U) is a variable of state. This means that its value depends only on the state of the system and not on its previous history. The value of ΔU, the change in internal energy, depends only on the initial and final states. Such states are identified by chemical composition, physical phase, temperature, pressure, and sometimes other relevant variables. The values of q and w, however, depend both on the states and on the way the transformation is brought about. Their sum must always satisfy Eq. (1). The first law governs any system and any combination of states. If a chemical reaction occurs during the transformation, the initial and final states will have different compositions. Then q is called the heat of reaction.

If the change in states takes place with no work done, then $w = 0$ and $\Delta U = q$. The most common example is a process which takes place irreversibly at constant volume. However, in the laboratory, chemical reactions are usually conducted at constant pressure. In this case, $w = -P\Delta V$, where P is the pressure and ΔV is the change in volume for the process. Under such conditions, the value for ΔU is given by Eq. (2).

$$\Delta U = q - P\Delta V \qquad (2)$$

See INTERNAL ENERGY.

Enthalpy. The property H is called enthalpy. It is defined in general as $H = U + PV$. When Eq. (2) is solved for the term q, the result gives the value for ΔH [Eq. (3)], the change in enthalpy.

$$q = \Delta U + P\Delta V = \Delta H \qquad (3)$$

See ENTHALPY.

If the change in states is brought about reversibly, then w is algebraically a minimum and q is a maximum. In this case, $q = T\Delta S$, where ΔS is the change of entropy of the system and T is the temperature. Heats of reaction are seldom measured directly under such conditions, however.

If q is positive for the irreversible process (energy transferred to the system) the reaction is called endothermic, and if q is negative (heat given off) the reaction is exothermic.

Change in enthalpy (ΔH) and change in internal energy (ΔU) for a chemical reaction are reported as a certain quantity of energy for the number of moles indicated in the balanced chemical equation. Equation (3) provides the relationship between the two quantities. If all reactants and products are liquids or solids, the difference is negligible (except at very high pressures). If gases are involved, the difference is significant only when the number of moles of product gases differs from the number of moles of reactant gases.

A chemical transformation may take place in a series of steps, each corresponding to a certain reaction. Thus, for example, at 1 bar (10^5 pascals)

and 25°C (77°F), the reaction steps shown as (*4a–d*) took place. (The subscript *r* is the symbol for chemical reaction.) The net result of all of these steps is shown in reaction (*4e*).

$$CH_4(gas) \rightarrow C(graphite) + 2H_2(gas) \qquad (4a)$$
$$\Delta_r H = 74.52 \quad \Delta_r U = 72.04$$

$$C(graphite) + {}^1/_2O_2(gas) \rightarrow CO(gas) \qquad (4b)$$
$$\Delta_r H = -110.525 \quad \Delta_r U = -111.764$$

$$CO(gas) + {}^1/_2O_2(gas) \rightarrow CO_2(gas) \qquad (4c)$$
$$\Delta_r H = -282.984 \quad \Delta_r U = -281.744$$

$$2H_2(gas) + O_2(gas) \rightarrow 2H_2O(liquid) \qquad (4d)$$
$$\Delta_r H = -571.66 \quad \Delta_r U = -564.22$$

$$CH_4(gas) + 2O_2(gas) \rightarrow CO_2(gas) + 2H_2O(liquid) \qquad (4e)$$
$$\Delta_r H = -890.64 \quad \Delta_r U = -885.69$$

The values of $\Delta_r H$ and $\Delta_r U$ are in kilojoules for the reaction specified. Since both U and H are variables of state, values of ΔU and ΔH are sums of the corresponding values for the intermediate steps for any path which leads from the initial to the final state. This result is called Hess's law, but it is simply a consequence of the first law of thermodynamics.

The change of any property, symbolized by X, associated with a chemical reaction may be equated to a sum of terms, one for each reactant and each product, by Eq. (5), where the $X(i)$ are the corresponding properties of reactants and products, i; and ν_i are the coefficients in the balanced chemical equation (positive for products, negative for reactants). The enthalpy of formation, sometimes called heat of formation, of a compound is the change in enthalpy for a reaction in which the compound is synthesized from its component elements. These quantities are seldom measured directly but are calculated from enthalpies of other reactions through the application of Hess's law. The enthalpy change for any reaction can be calculated from the enthalpies of formation of the reactants and products by the substitution of $\Delta_f H$ for X in the right side of Eq. (5) [the subscript f is the symbol for formation from elements]. *See* ENTHALPY.

$$\Delta_r X = \Sigma \nu_i X(i) \qquad (5)$$

Heats of reaction. Enthalpies of reactions may be obtained from several types of measurement. Three classes are recognized. The first-law heat of reaction is measured directly in a calorimeter. The second-law heat of reaction is calculated from the effect of temperature change on the equilibrium constant by the use of the Van't Hoff equation. The third-law heat of reaction is calculated by $\Delta_r H = \Delta_r G + T\Delta_r S$. $\Delta_r G$ is the change in Gibbs energy for the reaction calculated from an equilibrium constant or electrical cell potential, and $\Delta_r S$ is the corresponding change in entropy calculated from measured heat capacities and the third law of thermodynamics. In principle, all three methods give the same result for any particular reaction. The ease of measurement and the attainable accuracy may vary for different situations.

Enthalpies of reactions are slowly changing functions of temperature. The enthalpies of a reaction at the two temperatures T_1 and T_2 may be related to the difference in enthalpies for each component of the reaction between the two temperatures by Eq. (6).

$$\Delta_r H(T_2) = \Delta_r H(T_1) + \Delta_r[H(i, T_2) - H(i,T_1)] \qquad (6)$$

The quantities in the last term of Eq. (6) may be calculated from the corresponding heat capacities by Eq. (7).

$$H(i,T_2) - H(i,T_1) = \int_{T_1}^{T_2} C_p(i)\, dT \qquad (7)$$

See CHEMICAL THERMODYNAMICS; ENTROPY; FREE ENERGY.

Calorimetric measurements. A calorimeter is an instrument for measuring the heat added to or removed from a process. There are many designs, but the following parts can generally be identified: the vessel in which the process is confined, the thermometer which measures its temperature, and the surrounding environment called the jacket. The heat associated with the process is calculated by Eq. (8), where

$$q = C[(T(\text{final}) - T(\text{initial})] - q_{ex} - w \qquad (8)$$

T is the temperature. The quantity C, the energy equivalent of the calorimeter, is obtained from a separate calibration experiment. The work transferred to the process, w, is generally in the form of an electric current (as supplied to a heater, for example) or as mechanical work (as supplied to a stirrer, for example) and can be calculated from appropriate auxiliary measurements. The quantity q_{ex} is the heat exchanged between the container and its jacket during the experiment. It is calculated from the temperature gradients in the system and the measured thermal conductivities of its parts.

Two principal types of calorimeters are used to measure heats of chemical reactions. In a batch calorimeter, known quantities of reactants are placed in the vessel and the initial temperature is measured. The reaction is allowed to occur and then the final equilibrium temperature is measured. If necessary, the final contents are analyzed to determine the amount of reaction which occurred.

In a flow calorimeter, the reactants are directed to the reaction vessel in two or more steady streams. The reaction takes place quickly and the products emerge in a steady stream. The rate of heat production is calculated from the temperatures, flow velocities, and heat capacities of the incoming and outgoing streams, and the rates of work production and heat transfer to the jacket. Dividing this result by the rate of reaction gives the heat of reaction.

The combustion of a substance in oxygen is often studied in a specially designed reaction calorimeter. The heats of combustion of liquid or solid samples are usually measured in a batch-type calorimeter. The

vessel is a strong steel alloy bomb which is placed in a container of water fitted with appropriate thermometers and stirrers. The sample is placed in the bomb in an atmosphere of oxygen at high pressure (around 30 atm or 3 megapascals). It is then ignited by an electrical fuse. Heats of combustion of gases are usually measured in a type of flow calorimeter called a flame calorimeter. Heats of combustion can be measured to accuracies of 1 part in 10,000. The primary limiting factor in such measurements is the purity of the samples. Heats of combustion in fluorine have also been measured. *See* CALORIMETRY.

Units and symbols. The International Union of Pure and Applied Chemistry (IUPAC) Commission on Thermodynamics has recommended the general symbol $\Delta_\alpha^\beta X$ to represent the change of any property of a system when it changes from an initial state (denoted by α) to a final state (denoted by β). In addition, certain commonly observed processes are given special symbols of the type $\Delta_b X$, where the subscript b represents the process. Other symbols are vap, vaporization of a liquid; sub, sublimation of a solid; fus, fusion of a solid; mix, mixing without reaction; r, chemical reaction; f, formation from elements; and c, combustion. The subscript m placed after the property symbol indicates 1 mole of substance. Additional specifications can be placed within parentheses. For example, $\Delta_{vap}H_m(H_2O, 298.15\ K)$ symbolizes the heat of vaporization of 1 mole of water at 298.15 K, $\Delta_r G(1000\ K)$ the change in Gibbs energy for a chemical reaction at 1000 K, and $\Delta_c H_m(C_4H_{10}, g, 300\ K)$ the enthalpy of combustion of 1 mole of butane gas at 300 K.

Thermochemical quantities are usually reported and tabulated for substances in their standard states. The standard state of a solid is the thermodynamically stable crystal, of a liquid the liquid, and of a gas the hypothetical ideal gas, all at unit pressure. For the past century the pressure unit has been the atmosphere. It has been suggested that the bar is more suitable for this role as it is more compatible with the International System (SI) of units. Standard states can be defined for any temperature, but 25°C (298.15 K) has been traditional. In Customary units, this standard state is given at 77°F (536.67°R). The ideal gas is a hypothetical state, but its properties can readily be calculated from the equation of state of the real gas. The internal energy and enthalpy of an ideal gas are independent of its pressure.

The standard state for the solvent in a solution is the pure liquid. The standard states for the solutes are hypothetical ideal solutions at unit concentrations. Concentrations are usually expressed as molalities or mole fractions. Properties of real solutions can be related to those of the hypothetical ideal solutions by appropriate auxiliary data. Standard states for individual ions in solution are defined with the help of the additional conventions that the enthalpy and Gibbs energies of formation of the hydrogen ion are zero. A degree symbol (°) designates a property of a standard state. The standard-state concept promotes compactness and explicitness for the tabulation of data.

In the past, thermochemical quantities usually have been given in units of calories. A calorie is defined as the amount of heat needed to raise the temperature of 1 gram of water 1°C. However, since this depends on the initial temperature of the water, various calories have been defined, for example, the 15° calorie, the 20° calorie, and the mean calorie (average from 0 to 100°C). In addition, a number of dry calories have been defined. Those still used are the thermochemical calorie (exactly 4.184 joules) and the International Steam Table calorie (exactly 4.1868 J).

Thermochemical quantities have also been reported in terms of British thermal units (Btu). This unit is the amount of heat required to raise the temperature of 1 lb of water 1°F. A proliferation of Btu's similar to that for calories has occurred. The Btu in common use is the International Steam Table Btu (1055.056 J).

The SI rules do not recognize either the calorie or the Btu. The energy unit is the absolute joule (J). Most modern literature uses this unit. *See* PHYSICAL MEASUREMENT.

Sources of data. Original reports of measured values of thermochemical quantities are widely scattered among the world's scientific literature. A number of compilations of enthalpy of formation $\Delta_f H^\circ$, Gibbs energy of formation $\Delta_f G^\circ$, absolute entropy S°, and heat capacity at constant pressure C_p° of pure compounds at 298.15 K have been published during the past century. Some of them contain data at other temperatures and values for mixtures and ions as well.

A unique example is the International Critical Tables which appeared as a series of seven volumes between 1926 and 1930. The series was the result of an international cooperation among scientists to collect all reliable physical and chemical properties of materials available at that time. Volume V contains thermochemical data. More recent compilations have been made for inorganic compounds. The Landolt-Börnstein Tables, which have undergone a series of revisions since the 1890s, contain extensive thermodynamic data. Other compilations which are regularly updated by supplements and revisions are the JANAF Tables for low-molecular-weight inorganic compounds and the Thermodynamics Research Center publications for organic and some nonmetallic inorganic compounds.

A consequence of Hess's law is that thermochemical values such as heats of formation, combustion, reaction, and phase transition at a fixed temperature are interrelated through a system of linear algebraic equations. Hundreds, or even thousands, of such equations are available for even limited sets of compounds. They usually form an overdetermined set. The compiler has the job of selecting values of heats of formation which best fit the experimental data with consideration of the assigned uncertainties. If data at different temperatures and second- and third-law heats of reactions are included, the system of equations becomes nonlinear. Formerly the selection was made manually by a series of iterations; more recently computer programs have been written to help

in the task of data management and equation solving. The whole process must be repeated to incorporate new data.

To promote internal consistency among thermochemical compilations, a division of the Committee on Data for Science and Technology (CODATA) has recommended certain values of key properties. These represent a basic starting point for most other compilations.

Heating values. The heating value (also called calorific value) of a fuel is the heat of combustion (with a positive sign) of a certain quantity of fuel expressed in some units when burned under given conditions. The price for wholesale commodity transfer is based on the heating value, rather than mass or volume. Engineers use heating values to carry out heat balance calculations for furnaces, engines, and chemical processes.

Many specific definitions of heating value have been issued by trade and standards organizations around the world. An organization may give different definitions for solid, liquid, and gaseous fuels. Some of these have found their way into long-term sales contracts and legal systems of many countries. Therefore the term heating value does not have a universally recognized quantitative meaning.

In many English-speaking countries, heating values have been reported in Btu per pound at 15°F. For a gross heating value the water produced by the combustion is assumed to be liquid. For a net heating value it is assumed to be a gas. A precise definition requires additional specifications such as the nature of other products (for example, those formed from nitrogen or sulfur if present), the amount of water in the fuel (dry or wet basis), and whether the fuel is burned in air or oxygen. If the fuel is a gas, the state, real or ideal, must be indicated. If it is in a real state, the pressure must also be specified.

Commercial calorimeters are available for the measurement of heating values, but accurate measurements are difficult to make and require skilled technicians. Many fuels are complicated mixtures whose composition is not completely known. If the composition is known, the heating value can be calculated from the standard-state enthalpies of combustion of the pure components. This may require a large amount of auxiliary data such as heat capacities, heats of mixing, equation of state of pure and mixed systems, and values of unit conversions.

Organizations concerned with definitions of heating values are the U.S. National Bureau of Standards and its counterpart in other countries, American Society for Testing and Materials, Gas Processors Association, International Standards Organization, and Groupe International des Importateurs de Gaz Naturel Liquifie (GIIGNAL). *See* FOSSIL FUEL

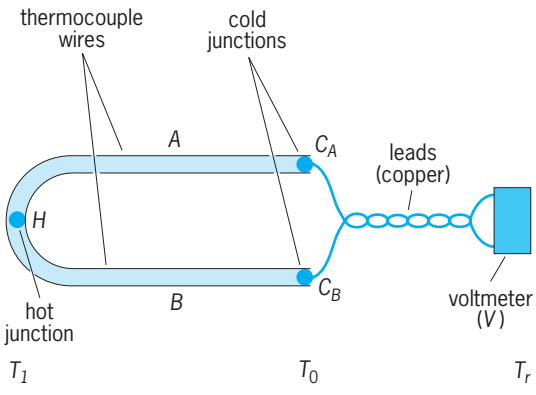
Randolph C. Wilhoit

Bibliography. P. W. Atkins, *Physical Chemistry*, 6th ed., 1998; H. Brodowsky and H. J. Schaller, *Thermochemistry of Alloys*, 1989; *Bulletin of Chemical Thermodynamics*, Thermochemistry Inc., Stillwater, Oklahoma; J. D. Cox, D. P. Wagman, and V. A. Mevdev (eds.), *CODATA Key Values for Thermodynamics*, 1989; M. Frenkel (ed.), *Thermochemistry and Equilibria of Organic Compounds*, 1993; *JANAF Thermochemical Tables*, Dow Chemical Co., accumulated supplements; I. Mills et al., *Quantities, Units and Symbols in Physical Chemistry*, 2d ed., 1993; *Physical and Thermodynamic Properties of Pure Compounds*, TRC Hydrocarbon Project, and *Selected Values of Properties of Chemical Compounds*, TRC Data Project, Thermodynamics Research Center, Texas A&M University, semiannual supplements; A report of IUPAC Commission 1.2 on Thermodynamics, *J. Chem. Thermodyn.*, 14:805–815, 1982.

Thermocouple

A device in which the temperature difference between the ends of a pair of dissimilar metal wires is deduced from a measurement of the difference in the thermoelectric potentials developed along the wires. The presence of a temperature gradient in a metal or alloy leads to an electric potential gradient being set up along the temperature gradient. This thermoelectric potential gradient is proportional to the temperature gradient and varies from metal to metal. It is the fact that the thermoelectric emf is different in different metals and alloys for the same temperature gradient that allows the effect to be used for the measurement of temperature.

Circuit. The basic circuit of a thermocouple is shown in the **illustration**. The thermocouple wires, made of different metals or alloys A and B, are joined together at one end H, called the hot (or measuring) junction, at a temperature T_1. The other ends, C_A and C_B (the cold or reference junctions), are maintained at a constant reference temperature T_0, usually but not necessarily 32°F (0°C). From the cold junctions, wires, usually of copper, lead to a voltmeter V at room temperature T_r. Due to the thermoelectric potential gradients being different along the wires A and B, there exists a potential difference between C_A and C_B. This can be measured by the voltmeter, provided that C_A and C_B are at the same temperature and that the lead wires between C_A and V and C_B and V are identical (or that V is at the temperature T_0, which is unusual).



Basic circuit of a thermocouple.

Such a thermocouple will produce a thermoelectric emf between C_A and C_B which depends only upon the temperature difference $T_1 - T_0$. If, however, the wires *A* or *B* are not physically or chemically homogeneous, this is no longer the case and the thermoelectric emf will also depend upon the shape of the temperature profile along the wires between *H* and *C*. Herein lies the principal limitation in the use of thermocouples for accurate temperature measurement. It is impossible, in practice, to obtain perfectly homogeneous wires and, moreover, as soon as one end is heated whatever homogeneity had previously existed is degraded to some extent. These problems limit the accuracy with which temperature measurements can be made by means of even the very best thermocouples to about 0.9°F (0.5°C) in the range from 32 to 1800°F (0 to 1000°C).

The thermoelectric emf in a thermocouple is developed in the temperature gradient and is in no way a junction phenomenon. Indeed, to avoid the effects of inhomogeneities near the junctions it is always advisable to try to arrange for both the hot and cold junctions to be in regions of uniform temperature. Under these conditions the way in which the junctions are made, whether it be by soldering, by welding, or simply by twisting the wires together, is of no consequence. *See* THERMOELECTRICITY.

Types. A large number of pure metal and alloy combinations have been studied as thermocouples, and the seven most widely used are listed in the **table**. The letter designations were originally introduced by the Instrument Society of America and have now gained worldwide acceptance.

The thermocouples in the table together cover the temperature range from about −420°F (−250°C or 20 K) to about 3300°F (1800°C). The most accurate and reproducible are the platinum/rhodium thermocouples, types R and S, while the most widely used industrial thermocouples are probably types K, T, and E. Each has its own special advantages and applications. For very low temperatures, below −420°F (−250°C or 20 K), thermocouples made from copper-nickel and gold-iron alloys have been developed. These make use of the Kondo effect and give useful thermoelectric emf's down to 1 K (−458°F or −272°C). For temperatures above the maximum range of type B platinum/rhodium thermocouples (about 3300°F or 1800°C), recourse has to be made to thermocouples using wires of alloys of tungsten and rhenium. These can be used up to 4900°F (2700°C) but must be very carefully protected from oxidation. For industrial applications these and many other thermocouples are manufactured in the so-called mineral-insulated (MI) version. The thermocouple wire, the refractory metal oxide insulator (magnesium oxide or beryllium oxide, for example), and the hermetically sealed metal sheath are assembled and sealed during manufacture. Such mineral-insulated thermocouples are thus to a very large extent protected from chemical contamination and mechanical damage. *See* KONDO EFFECT; LOW-TEMPERATURE THERMOMETRY.

Letter designations and compositions for standardized thermocouples*

Type designation	Materials
B	Platinum-30% rhodium/platinum-6% rhodium
E	Nickel-chromium alloy/a copper-nickel alloy
J	Iron/another slightly different copper-nickel alloy
K	Nickel-chromium alloy/nickel-aluminum alloy
R	Platinum-13% rhodium/platinum
S	Platinum-10% rhodium/platinum
T	Copper/a copper-nickel alloy

*After T. J. Quinn, *Temperature*, Academic Press, 1983.

Calibration. For all of the thermocouples listed in the table, international agreement has been obtained on standard reference tables of emf versus temperature. These are given in the International Electrotechnic Commission Standard, IEC 584, and are identical to those given in ASTM E-230-77 and many other national standards. Calibration of a thermocouple is best carried out by measuring the difference between the thermoelectric emf given by the thermocouple and that predicted by the standard reference table. Having measured this difference at a small number of fixed points, a complete calibration table is easily obtained by interpolating these differences from the standard reference table between the measured fixed points.

Pressure and magnetic field corrections. Due to the detailed mechanism of the thermoelectric effects, the magnitude of the thermoelectric emf given by a thermocouple is also affected by pressure and by the presence of magnetic fields. At very high pressures, for example, a type S thermocouple at 1800°F (1000°C) and at a pressure of 4 gigapascals (approximately 40 kilobars) is subject to a correction of about 36°F (20°C). For use in magnetic fields of 8 teslas and above, type E thermocouples are recommended as being one of the least affected combinations.

Extension and compensating wires. In many large industrial applications, the hot and cold junctions are widely separated, but the measured thermoelectric emf may be almost entirely developed in the first several feet (few meters) of wire. The remaining length of wire serves mainly to transmit the emf to the measuring system. The thermoelectric properties of this long length of wire near room temperature, or at least at temperatures below 212°F (100°C), are very much less critical than those of that part of the wire in the steep temperature gradient. Considerable economies can, therefore, be gained by using, in this less critical section, not the high-specification thermocouple wire, but another, cheaper wire whose thermoelectric properties are a reasonable match over the temperature range, say from 68 to 212°F (20 to 100°C). If these wires are made to the same nominal composition of the thermocouple wire but not to the same high specification, they are called extension wires. If, on the other hand, they are of a different

composition but one chosen to match the thermoelectric properties over this restricted temperature range, they are known as compensating wires. *See* TEMPERATURE MEASUREMENT. T. J. Quinn

Bibliography. T. W. Kerlin, *Practical Thermocouple Thermometry*, 1999; T. D. McGee, *Principles and Methods of Temperature Measurement*, 1988; R. M. Parks (ed.), *Manual on the Use of Thermocouples in Temperature Measurement*, American Society for Testing and Measurement, MNL 12, 1993; D. D. Pollock, *Thermocouples: Theory and Practice*, 1991; T. J. Quinn, *Temperature*, 2d ed., 1991.

Thermodynamic cycle

A procedure or arrangement in which one form of energy, such as heat at an elevated temperature from combustion of a fuel, is in part converted to another form, such as mechanical energy on a shaft, and the remainder is rejected to a lower-temperature sink as low-grade heat.

Common features of cycles. A thermodynamic cycle requires, in addition to the supply of incoming energy, (1) a working substance, usually a gas or vapor; (2) a mechanism in which the processes or phases can be carried through sequentially; and (3) a thermodynamic sink to which the residual heat can be rejected. The cycle itself is a repetitive series of operations.

There is a basic pattern of processes common to power-producing cycles. There is a compression process wherein the working substance undergoes an increase in pressure and therefore density. There is an addition of thermal energy from a source such as a fossil fuel, a fissile fuel, or solar radiation. There is an expansion process during which work is done by the system on the surroundings. There is a rejection process where thermal energy is transferred to the surroundings. The algebraic sum of the energy additions and abstractions is such that some of the thermal energy is converted into mechanical work. *See* HEAT.

A steam cycle that embraces a boiler, a prime mover, a condenser, and a feed pump is typical of the cyclic arrangement in which the thermodynamic fluid, steam, is used over and over again. An alternative procedure, after the net work flows from the system, is to employ a change of mass within the system boundaries, the spent working substance being replaced by a fresh charge that is ready to repeat the cyclic events. The automotive engine and the gas turbine illustrate this arrangement of the cyclic processes, called an open cycle because new mass enters the system boundaries and the spent exhaust leaves it.

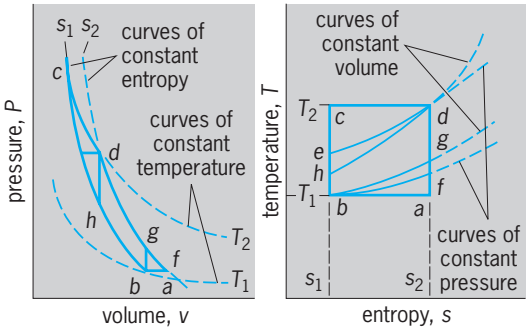
The basic processes of the cycle, either open or closed, are heat addition, heat rejection, expansion, and compression. These processes are always present in a cycle even though there may be differences in working substance, the individual processes, pressure ranges, temperature ranges, mechanisms, and heat transfer arrangements.

Air-standard cycle. It is convenient to study the various power cycles by using an ideal system such as the air-standard cycle. This is an ideal, frictionless mechanism enveloping the system, with a permanent unit charge of air behaving in accordance with the perfect gas relationships.

The unit air charge is assumed to have an initial state at the start of the cycle to be analyzed. Each process is assumed to be perfectly reversible, and all effects between the system and the surroundings are described as either a heat transfer or a mechanical work term. At the end of a series of processes, the state of the system is the same as it was initially. Because no chemical changes take place within the system, the same unit air charge is conceivably capable of going through the cyclic processes repeatedly.

Whereas this air-standard cycle is an idealization of an actual cycle, it provides an amenable method for the introductory evaluation of any power cycle. Its analysis defines the upper limits of performance toward which the actual cycle performance may approach. It defines trends, if not absolute values, for both ideal and actual cycles. The air-standard cycle can be used to examine such cycles as the Carnot and those applicable to the automobile engine, the diesel engine, the gas turbine, and the jet engine.

Cyclic standards. Many cyclic arrangements, using various combinations of phases but all seeking to convert heat into work, have been proposed by many investigators whose names are attached to their proposals, for example, the Diesel, Otto, Rankine, Brayton, Stirling, Ericsson, and Atkinson cycles (see **illus.**). All proposals are not equally efficient in the conversion of heat into work. However, they may offer other advantages which have led to their practical development for various applications. Nevertheless, there is one overriding limitation on efficiency. It is set by the dictates of the Carnot cycle, which states that no thermodynamic cycle can be projected whose thermal efficiency exceeds that of the Carnot cycle between specified temperature levels for the heat source and the heat sink. Many cycles may approach and even equal this limit, but none can exceed it. This is the uniqueness of the Carnot principle and is basic to the second law



Comparison of principal thermodynamic cycles. Cycles are, in the order of decreasing efficiency, Carnot cycle (*a-b-c-d-a*), Brayton cycle (*b-e-d-f-b*), Diesel cycle (*b-e-d-g-b*), Otto cycle (*b-h-d-g-b*).

of thermodynamics on the conversion of heat into work. *See* BRAYTON CYCLE; CARNOT CYCLE; DIESEL CYCLE; OTTO CYCLE; STIRLING ENGINE; THERMODYNAMIC PROCESSES.

Theodore Baumeister

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; R. T. Balmer, *Thermodynamics*, 1989; I. Granet, *Thermodynamics and Heat Power*, 5th ed., 1995; W. C. Reynolds and H. C. Perkins, *Engineering Thermodynamics*, 2d ed., 1977; M. W. Zemansky, *Heat and Thermodynamics*, 6th ed., 1981.

Thermodynamic principles

Laws governing the transformation of energy. Thermodynamics is the science of the transformation of energy. It differs from the dynamics of Newton by taking into account the concept of temperature, which is outside the scope of classical mechanics. In practice, thermodynamics is useful for assessing the efficiencies of heat engines (devices that transform heat into work) and refrigerators (devices that use external sources of work to transfer heat from a hot system to cooler sinks), and for discussing the spontaneity of chemical reactions (their tendency to occur naturally) and the work that they can be used to generate.

The subject of thermodynamics is founded on four generalizations of experience, which are called the laws of thermodynamics. Each law embodies a particular constraint on the properties of the world. The connection between phenomenological thermodynamics and the properties of the constituent particles of a system is established by statistical thermodynamics, also called statistical mechanics. Classical thermodynamics consists of a collection of mathematical relations between observables, and as such is independent of any underlying model of matter (in terms, for instance, of atoms). However, interpretations in terms of the statistical behavior of large assemblies of particles greatly enriches the understanding of the relations established by thermodynamics, and a full description of nature should use explanations that move effortlessly between the two modes of discourse. *See* STATISTICAL MECHANICS.

There is a handful of very primitive concepts in thermodynamics. The first is the distinction between the system, which is the assembly of interest, and the surroundings, which is everything else. The surroundings is where observations on the system are carried out and attempts are made to infer its properties from those measurements. The system and the surroundings jointly constitute the universe and are distinguished by the boundary that separates them. If the boundary is impervious to the penetration of matter and energy, the system is classified as isolated. If energy but not matter can pass through it, the system is said to be closed. If both energy and matter can penetrate the boundary, the system is said to be open. Another primitive concept of thermodynamics is work. By work is meant a process by which a weight may be raised in the surroundings. Work is the link between mechanics and thermodynamics. *See* WORK.

None of these primitive concepts introduces the properties that are traditionally regarded as central to thermodynamics, namely temperature, energy, heat, and entropy. These concepts are introduced by the laws and are based on the foundations that these primitive concepts provide.

Zeroth law of thermodynamics. The zeroth law of thermodynamics establishes the existence of a property called temperature. This law is based on the observation that if a system A is in thermal equilibrium with a system B (that is, no change in the properties of B take places when the two are in contact), and if system B is in thermal equilibrium with a system C, then it is invariably the case that A will be found to be in equilibrium with C if the two systems are placed in mutual contact. This law suggests that a numerical scale can be established for the common property, and if A, B, and C have the same numerical values of this property, then they will be in mutual thermal equilibrium if they were placed in contact. This property is now called the temperature.

In thermodynamics it is appropriate to report temperatures on a natural scale, where 0 is ascribed to the lowest attainable temperature. Temperatures on this thermodynamic temperature scale are denoted T and are commonly reported in kelvins. The relation between the Kelvin scale and the Celsius scale (θ) is $T(\mathrm{K}) = \theta(^\circ\mathrm{C}) + 273.15$. *See* TEMPERATURE.

First law of thermodynamics. The first law of thermodynamics establishes the existence of a property called the internal energy of a system. It also brings into the discussion the concept of heat.

The first law is based on the observation that a change in the state of a system can be brought about by a variety of techniques. Indeed, if attention is confined to an adiabatic system, one that is thermally insulated from its surroundings, then the work of J. P. Joule shows that same change of state is brought about by a given quantity of work regardless of the manner in which the work is done (**Fig. 1**). This observation suggests that, just as the height through which a mountaineer climbs can be calculated from the difference in altitudes regardless of the path the climber takes between two fixed points, so the work, w, can be calculated from the difference between the final and initial properties of a system. The relevant property is called the internal energy, U. However, if the transformation of the system is taken along a path that is not adiabatic, a different quantity of work may be required. The difference between the work of adiabatic change and the work of nonadiabatic change is called heat, q. In general, Eq. (1) is

$$\Delta U = w + q \tag{1}$$

satisfied, where ΔU is the change in internal energy between the final and initial states of the system. *See* ADIABATIC PROCESS; ENERGY; HEAT.

The down-to-earth implication of this slightly casuistical argument is that there are two modes of

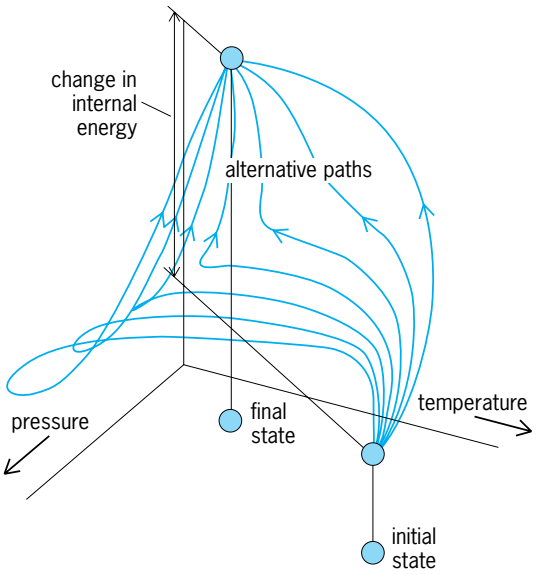


Fig. 1. The observation that the same work is done whatever the adiabatic path between two specified states of a system implies that each state has a characteristic property such that the difference in its values is equal to the adiabatic work required to pass between them. This characteristic property, a so-called state property, is called the internal energy of the system.

transferring energy between a system and its surroundings. One is by doing work; the other is by heating the system. Work and heat are modes of transferring energy. They are not forms of energy in their own right. Work is a mode of transfer that is equivalent (if not the case in actuality) to raising a weight in the surroundings. Heat is a mode of transfer that arises from a difference in temperature between the system and its surroundings. What is commonly called heat is more correctly called the thermal motion of the molecules of a system.

The molecular interpretation of thermodynamics adds insight to the operational distinction between work and heat. Work is a transfer of energy that stimulates (or is caused by) organized molecular motion in the surroundings. Thus, the raising of a weight by a system corresponds to the organized, unidirectional motion of the atoms of the weight. In contrast, heat is the transfer of energy that stimulates (or is caused by) chaotic molecular motion in the surroundings. Thus, the emergence of energy as heat into the surroundings is the chaotic, tumbling-out of stored energy.

The first law of thermodynamics states that the internal energy of an isolated system is conserved. That is, for a system to which no energy can be transferred by the agency of work or of heat, the internal energy remains constant. This law is a cousin of the law of the conservation of energy in mechanics, but it is richer, for it implies the equivalence of heat and work for bringing about changes in the internal energy of a system (and heat is foreign to classical mechanics). The first law is a statement of permission: no change may occur in an isolated system unless that change corresponds to a constant internal energy.

It is common in thermodynamics to switch attention from the change in internal energy of a system to the change in enthalpy, ΔH, of the system. The change in internal energy and the change in enthalpy of a system subjected to constant pressure p are related by Eq. (2), where ΔV is the change in volume of

$$\Delta H = \Delta U + p\Delta V \tag{2}$$

the system that accompanies the change of interest. The interpretation of ΔH is that it is equal to the energy that may be obtained as heat when the process occurs. This interpretation follows from the fact that the term $p\Delta V$ takes into account the work of driving back the surroundings that must take place during the process, and that is therefore not available for providing heat. Enthalpy changes are widely used in thermochemistry, the branch of physical chemistry concerned with the heat productions and requirements of chemical reactions. *See* ENTHALPY; THERMOCHEMISTRY.

The enthalpy itself is defined in terms of the internal energy by Eq. (3). Two important quantities in

$$H = U + pV \tag{3}$$

thermodynamics are the heat capacities at constant volume C_V and constant pressure C_p. These quantities are defined as the slope of the variation of internal energy and enthalpy, respectively, with respect to temperature. The two quantities differ on account of the work that must be done to change the volume of the system when the constraint is that of constant pressure, and then less energy is available for raising the temperature of the system. For a perfect gas, they are related by Eq. (4), where n is the amount of

$$C_p - C_V = nR \tag{4}$$

substance and R is the gas constant. *See* GAS; HEAT CAPACITY.

Second law of thermodynamics. The second law of thermodynamics deals with the distinction between spontaneous and nonspontaneous processes. A process is spontaneous if it occurs without needing to be driven. In other words, spontaneous changes are natural changes, like the cooling of hot metal and the free expansion of a gas. Many conceivable changes occur with the conservation of energy globally, and hence are not in conflict with the first law; but many of those changes turn out to be nonspontaneous, and hence occur only if they are driven.

The second law was formulated by Lord Kelvin and by R. Clausius in a manner relating to observation: "no cyclic engine operates without a heat sink" and "heat does not transfer spontaneously from a cool to a hotter body," respectively (**Fig. 2**). The two statements are logically equivalent in the sense that failure of one implies failure of the other. However, both may be absorbed into a single statement: the entropy of an isolated system increases when a spontaneous change occurs. The property of entropy is introduced to formulate the law quantitatively in exactly the same way that the properties

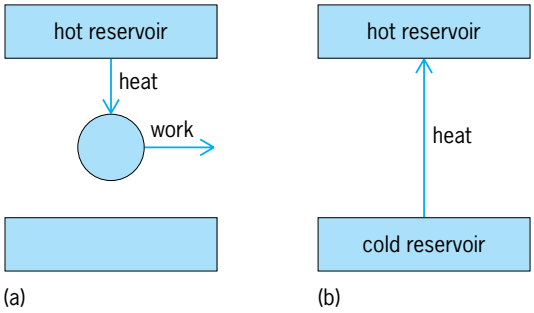


Fig. 2. Representation of the statements of the second law of thermodynamics by (*a*) Lord Kelvin and (*b*) R. Clausius. In each case, the law states that the device cannot operate as shown.

of temperature and internal energy are introduced to render the zeroth and first laws quantitative and precise.

The entropy, S, of a system is a measure of the quality of the energy that it stores. The formal definition is based on Eq. (5), where dS is the change in

$$dS = \frac{dq_{\text{reversible}}}{T} \tag{5}$$

entropy of a system, dq is the energy transferred to the system as heat, T is the temperature, and the subscript "reversible" signifies that the transfer must be carried out reversibly (without entropy production other than in the system). When a given quantity of energy is transferred as heat, the change in entropy is large if the transfer occurs at a low temperature and small if the temperature is high.

This somewhat austere definition of entropy is greatly illuminated by L. Boltzmann's interpretation of entropy as a measure of the disorder of a system. The connection can be appreciated qualitatively at least by noting that if the temperature is high, the transfer of a given quantity of energy as heat stimulates a relatively small additional disorder in the thermal motion of the molecules of a system; in contrast, if the temperature is low, the same transfer could stimulate a relatively large additional disorder. This connection between entropy and disorder is justified by a more detailed analysis, and in general it is safe to interpret the entropy of a system or its surroundings as a measure of the disorder present.

The formal statement of the second law of thermodynamics is that the entropy of an isolated system increases in the course of a spontaneous change. The illumination of the law brought about by the association of entropy and disorder is that in an isolated system (into which technology cannot penetrate) the only changes that may occur are those in which there is no increase in order. Thus, energy and matter tend to disperse in disorder (that is, entropy tends to increase), and this dispersal is the driving force of spontaneous change. *See* TIME, ARROW OF.

This collapse into chaos need not be uniform. The isolated system need not be homogeneous, and there may be an increase in order in one part so long as there is a compensating increase in disorder in another part. Thus, in thermodynamics, collapse into disorder in one region of the universe can result in the emergence of order in another region. The criterion for the emergence of order is that the decrease in entropy associated with it is canceled by a greater increase in entropy elsewhere. *See* ENTROPY.

Third law of thermodynamics. The practical significance of the second law is that it limits the extent to which the internal energy may be extracted from a system as work. In short, in order for a process to generate work, it must be spontaneous. (There is no point in using a process to produce work if that process itself needs to be driven; it is then no more than a gear wheel.) Therefore, any work-producing process must be accompanied by an increase in entropy. If a quantity of energy were to be withdrawn from a hot source and converted entirely into work, there would be a decrease in the entropy of the hot source, and no compensating increase elsewhere. Therefore, such a process is not spontaneous and cannot be used to generate work. For the process to be spontaneous, it is necessary to discard some energy as heat in a sink of lower temperature. In other words, nature in effect exacts a tax on the extraction of energy as work. There is therefore a fundamental limit on the efficiency of engines that convert heat into work (**Fig. 3**).

The quantitative limit on the efficiency, ϵ, which is defined as the work produced divided by the heat absorbed from the hot source, was first derived by S. Carnot. He found that, regardless of the details of the construction of the engine, the maximum efficiency (that is, the work obtained after payment of the minimum allowable tax to ensure spontaneity) is given by Eq. (6), where T_{hot} is the temperature of the

$$\epsilon = 1 - \frac{T_{\text{cold}}}{T_{\text{hot}}} \tag{6}$$

hot source and T_{cold} is the temperature of the cold sink. The greatest efficiencies are obtained with the coldest sinks and the hottest sources, and these are

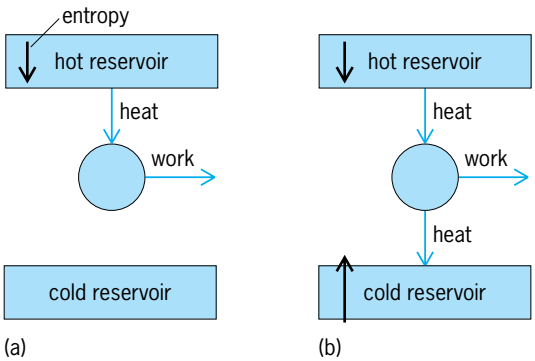


Fig. 3. Criterion for ability of a device to produce work. (*a*) This device cannot produce work because there is a decrease in entropy of the hot reservoir as energy leaves it as heat, but no compensating increase in entropy of the cold sink; overall, there is a decrease in entropy. (*b*) So long as a certain quantity of energy is discarded as heat into the cold reservoir, the overall change in entropy may be positive, and the engine can produce work spontaneously.

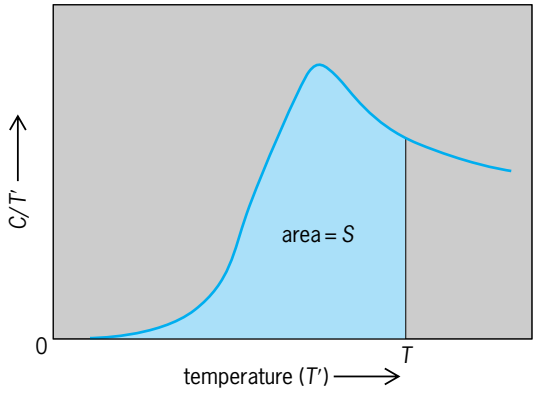


Fig. 4. Representation of the experimental determination of entropy. Measurements are made of the heat capacity, *C*, down to as low a temperature, *T′*, as possible, and *C*/*T′* is plotted against *T′*. The area under the curve up to the temperature of interest, *T*, is equal to the entropy, *S*, of the system; this procedure supposes that $S = 0$ at $T = 0$.

the design requirements of modern power plants. *See* CARNOT CYCLE.

Perfect efficiency ($\epsilon = 1$) would be obtained if the cold sink were at absolute zero ($T_{\text{cold}} = 0$). However, the third law of thermodynamics, which is another summary of observations, asserts that absolute zero is unattainable in a finite number of steps for any process. Therefore, heat can never be completely converted into work in a heat engine. The implication of the third law in this form is that the entropy change accompanying any process approaches zero as the temperature approaches zero. That implication in turn implies that all substances tend toward the same entropy as the temperature is reduced to zero. It is therefore sensible to take the entropy of all perfect crystalline substances (substances in which there is no residual disorder arising from the location of atoms) as equal to zero. A common short statement of the third law is therefore that all perfect crystalline substances have zero entropy at absolute zero ($T = 0$). This statement is consistent with the interpretation of entropy as a measure of disorder, since at absolute zero all thermal motion has been quenched. *See* ABSOLUTE ZERO.

In practice, the entropy of a sample of a substance is measured by determining its heat capacity, C, at all temperatures between zero and the temperature of interest, T, and evaluating the integral given in Eq. (7). Graphically, C/T' is plotted against T', and

$$S = \int_0^T \frac{C}{T'}\, dT' \tag{7}$$

the area under the curve up to the temperature of interest is equal to the entropy (**Fig. 4**). In practice, measurements of the heat capacity are made down to as low a temperature as possible, and certain approximations are generally carried out in order to extrapolate these measurements down to absolute zero. A polynomial is fitted to the data, and the integration in Eq. (7) is performed analytically. If there are phase transitions below the temperature of interest, a contribution from each such transition is added, equal to the enthalpy change of the transition divided by the temperature at which it occurs. Such determinations show that the entropy of a substance increases as it changes from a solid to a liquid to a gas. *See* PHASE TRANSITIONS.

Gibbs free energy. One of the most important derived quantities in thermodynamics is the Gibbs energy, G, which is widely called the free energy. It is defined by Eq. (8), where H is the enthalpy of the

$$G = H - TS \tag{8}$$

system, T is its thermodynamic temperature, and S is its entropy. The principal significance of G is that a change in G is a measure of the energy of the system that is free to do work other than simply driving back the walls of the system as the process occurs. For instance, it is a measure of the electrical work that may be extracted when a chemical reaction takes place, or the work of constructing a protein that a biochemical process may achieve.

The Gibbs energy can be developed in two different ways. First, it is quite easy to show from formal thermodynamics that Eq. (9) is valid. That is, the

$$\Delta G = -T\Delta S(\text{total}) \tag{9}$$

change, ΔG, in the Gibbs energy is proportional to the total change, ΔS(total), in the entropy of the system and its surroundings. The negative sign in Eq. (9) indicates that an increase in the total entropy corresponds to a decrease in the Gibbs energy (**Fig. 5**). Because a spontaneous change occurs in the direction of the increase in total entropy, it follows that another way of expressing the signpost of spontaneous change is that it occurs in the direction of decreasing Gibbs energy. To this extent, the Gibbs energy is no more than a disguised version of the total entropy.

However, the Gibbs energy is much more than that, for (as discussed above) it shows how much

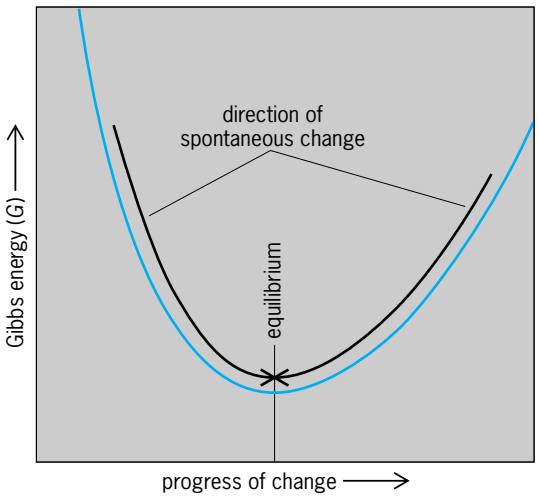


Fig. 5. Gibbs energy of a system. A spontaneous change (under conditions of constant temperature and pressure) corresponds to a decrease in Gibbs energy. This is a disguised form of the identification of the spontaneous direction of change with an increase in the total entropy of the system and its surroundings.

nonexpansion work may be extracted from a process. Broadly speaking, because a change in Gibbs energy at constant temperature can be expressed as $\Delta G = \Delta H - T\Delta S$, the latter term represents the tax exacted by nature to ensure that overall a process is spontaneous. Whereas ΔH measures the energy that may be extracted as heat, some energy may need to be discarded into the surroundings to ensure that overall there is an increase in entropy when the process occurs, and that quantity ($T\Delta S$) is then no longer available for doing work. This is the origin of the name free energy for the Gibbs energy, for it represents that energy stored by the system that is free to be extracted as work.

For a chemical reaction, the standard reaction Gibbs energy is calculated from the differences of the standard Gibbs energies of formation of each substance, the change in Gibbs energy accompanying the formation of the substance from its elements under standard conditions (a pressure of 1 bar or 10^5 pascals). The standard reaction Gibbs energy is the principal thermodynamic function for applications in chemistry.

The Gibbs energy is at the center of attention in chemical thermodynamics. It is deployed by introducing a related quantity called the chemical potential. The chemical potential is defined as the slope of the graph showing how the total Gibbs energy of a system varies as the composition of one of its components is increased. The slope of the graph varies with composition, so the chemical potential also varies with composition. Broadly speaking, the chemical potential can be interpreted as a measure of the potential of a substance to undergo chemical change: if its chemical potential is high, then it has a high potential for bringing about chemical change. Thus, a gas at high pressure has a higher chemical potential than one at low pressure, and a substance at high temperature has a higher chemical potential than the same substance at a lower temperature.

An implication of the second law is that the chemical potential of a substance must be the same throughout any phase in which it occurs and the same in all phases that are at equilibrium in a system. These requirements lead, by a subtle argument, to one of the most celebrated conclusions in chemical thermodynamics, the Gibbs phase rule, Eq. (10).

$$F = C - P + 2 \qquad (10)$$

In this expression, C is the number of components in the system (essentially the number of chemically distinct species), P is the number of phases that are in equilibrium with one another, and F is the variance, the number of variables that may be changed without changing the number of phases that are in equilibrium with one another. (In some formulations, C is denoted $S - M$, where S is the number of substances and M is the number of reactions which relate them and which are at equilibrium.) The phase rule is particularly important for discussing the structure of phase diagrams, which are charts showing the range of temperature and composition over which various phases of a system are in equilibrium (**Fig. 6**). *See* PHASE EQUILIBRIUM; PHASE RULE.

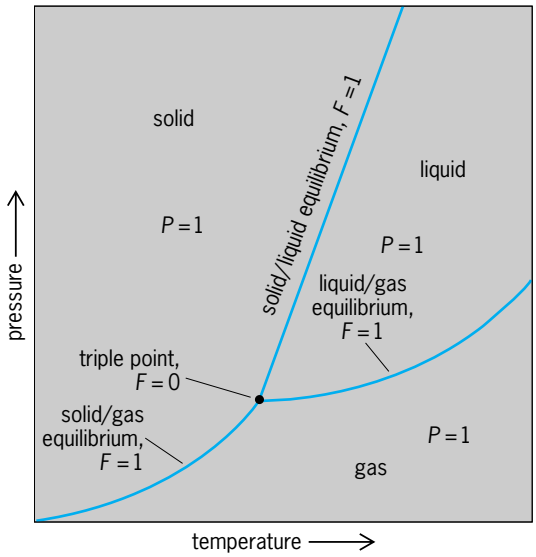


Fig. 6. Simple phase diagram for a one-component system ($C = 1$) and its interpretation in terms of the phase rule. The lines ($F = 1$) show the conditions under which two phases ($P = 2$) are in mutual equilibrium; the triple point ($F = 0$) is the unique set of conditions under which three phases ($P = 3$) can mutually coexist in equilibrium. The regions separated by the phase boundaries show the conditions where the specified phase is thermodynamically the most stable.

In systems in which chemical reactions can occur, the chemical potentials of the reactants and products can be used to determine the composition at which the reaction mixture has reached a state of dynamic equilibrium, with no remaining tendency for spontaneous change. In chemistry, the state of chemical equilibrium is normally expressed in terms of the equilibrium constant, K, which is defined in terms of the concentrations or partial pressures of the participating species. In general, if the products dominate in the reaction mixture at equilibrium, then the equilibrium constant is greater than 1, and if the reactants dominate, then it is less than 1. Manipulations of standard thermodynamic relations show that the standard reaction Gibbs energy of any reaction is proportional to the negative of the logarithm of the equilibrium constant. *See* CHEMICAL EQUILIBRIUM; CHEMICAL THERMODYNAMICS; FREE ENERGY.

Thermodynamics of irreversible processes. The thermodynamic study of irreversible processes centers on the rate of entropy production and the relation between the fluxes and the forces that give rise to them. These fluxes include diffusion (the flux of matter), thermal conduction (the flux of energy of thermal motion), and electrical conduction (the flux of electric charge). In each case, the flux arises from a generalized potential difference of some kind. Thus, diffusion is related to a concentration gradient, and thermal conduction is related to a temperature gradient. In each case, the rate of change of entropy arising from the flux is proportional to both the flux and

the gradient giving rise to the flux. Thus, a high flux of matter down a steep concentration gradient results in a rapid change in entropy. *See* CONDUCTION (ELECTRICITY); CONDUCTION (HEAT); DIFFUSION.

An important observation is that the fluxes, J_i, and potentials, X_j, are not independent of one another. Thus, a temperature gradient can result in diffusion, and a concentration gradient can result in a flux of energy. The general relation between flux and potential is therefore given by Eq. (11), where the L_{ij}

$$J_i = \sum_j L_{ij} X_j \tag{11}$$

are called the phenomenological coefficients. It was shown by L. Onsager that for conditions not far from equilibrium, $L_{ij} = L_{ji}$. This Onsager reciprocity relation implies that there is a symmetry in the ability of a potential X_j to give to a flux J_i and of a potential X_j to give rise to a flux J_j. *See* THERMODYNAMIC PROCESSES; THERMOELECTRICITY. P. W. Atkins

Bibliography. P. W. Atkins, *Physical Chemistry*, 6th ed., 1998; W. T. Grandy, *Foundations of Statistical Mechanics*, 2 vols., 1988; J. Keizer, *Statistical Thermodynamics of Nonequilibrium Processes*, 1987; I. Müller, *Thermodynamics*, 1985; R. E. Sonntag, C. Borgnakke, and G. J. Van Wylen, *Fundamentals of Thermodynamics*, 5th ed., 1997; K. Wark, *Thermodynamics*, 6th ed., 2000; S. E. Wood and R. Battino, *Thermodynamics of Chemical Systems*, 1990.

Thermodynamic processes

Changes of any property of an aggregation of matter and energy, accompanied by thermal effects. The participants in a process are first identified as a system to be studied; the boundaries of the system are established; the initial state of the system is determined; the path of the changing states is laid out; and, finally, supplementary data are stated to establish the thermodynamic process. These steps will be explained in the following paragraphs. At all times it must be remembered that the only processes which are allowed are those compatible with the first and second laws of thermodynamics: Energy is neither created nor destroyed and the entropy of the system plus its surroundings always increases.

A system and its boundaries. To evaluate the results of a process, it is necessary to know the participants that undergo the process, and their mass and energy. A region, or a system, is selected for study, and its contents determined. This region may have both mass and energy entering or leaving during a particular change of conditions, and these mass and energy transfers may result in changes both within the system and within the surroundings which envelop the system.

As the system undergoes a particular change of condition, such as a balloon collapsing due to the escape of gas or a liquid solution brought to a boil in a nuclear reactor, the transfers of mass and energy which occur can be evaluated at the boundaries of the arbitrarily defined system under analysis.

A question that immediately arises is whether a system such as a tank of compressed air should have boundaries which include or exclude the metal walls of the tank. The answer depends upon the aim of the analysis. If its aim is to establish a relationship among the physical properties of the gas, such as to determine how the pressure of the gas varies with the gas temperature at a constant volume, then only the behavior of the gas is involved; the metal walls do not belong within the system. However, if the problem is to determine how much externally applied heat would be required to raise the temperature of the enclosed gas a given amount, then the specific heat of the metal walls, as well as that of the gas, must be considered, and the system boundaries should include the walls through which the heat flows to reach the gaseous contents. In the laboratory, regardless of where the system boundaries are taken, the walls will always play a role and must be reckoned with.

State of a system. To establish the exact path of a process, the initial state of the system must be determined, specifying the values of variables such as temperature, pressure, volume, and quantity of material. If a number of chemicals are present in the system, the number of variables needed is usually equal to the number of independently variable substances present plus two such as temperature and pressure; exceptions to this rule occur in variable electric or magnetic fields and in some other well-defined cases. Thus, the number of properties required to specify the state of a system depends upon the complexity of the system. Whenever a system changes from one state to another, a process occurs.

Whenever an unbalance occurs in an intensive property such as temperature, pressure, or density, either within the system or between the system and its surroundings, the force of the unbalance can initiate a process that causes a change of state. Examples are the unequal molecular concentration of different gases within a single rigid enclosure, a difference of temperature across the system boundary, a difference of pressure normal to a nonrigid system boundary, or a difference of electrical potential across an electrically conducting system boundary. The direction of the change of state caused by the unbalanced force is such as to reduce the unbalanced driving potential. Rates of changes of state tend to decelerate as this driving potential is decreased.

Equilibrium. The decelerating rate of change implies that all states move toward new conditions of equilibrium. When there are no longer any balanced forces acting within the boundaries of a system or between the system and its surroundings, then no mechanical changes can take place, and the system is said to be in mechanical equilibrium. A system in mechanical equilibrium, such as a mixture of hydrogen and oxygen, under certain conditions might undergo a chemical change. However, if there is no net change in the chemical constituents, then the

mixture is said to be in chemical as well as in mechanical equilibrium.

If all parts of a system in chemical and mechanical equilibrium attain a uniform temperature and if, in addition, the system and its surroundings either are at the same temperature or are separated by a thermally nonconducting boundary, then the system has also reached a condition of thermal equilibrium.

Whenever a system is in mechanical, chemical, and thermal equilibrium, so that no mechanical, chemical, or thermal changes can occur, the system is in thermodynamic equilibrium. The state of equilibrium is at a point where the tendency of the system to minimize its energy is balanced by the tendency toward a condition of maximum randomness. In thermodynamics, the state of a system can be defined only when it is in equilibrium. The static state on a macroscopic level is nevertheless underlaid by rapid molecular changes; thermodynamic equilibrium is a condition where the forward and reverse rates of the various changes are all equal to one another. In general, those systems considered in thermodynamics can include not only mixtures of material substances but also mixtures of matter and all forms of energy. For example, one could consider the equilibrium between a gas of charged particles and electromagnetic radiation contained in an oven.

Process path. If under the influence of an unbalanced intensive factor the state of a system is altered, then the change of state of the system is described in terms of the end states or difference between the initial and final properties.

The path of a change of state is the locus of the whole series of states through which the system passes when going from an initial to a final state. For example, suppose a gas expands to twice its volume and that its initial and final temperatures are the same. Various paths connect these initial and final states: isothermal expansion, with temperature held constant at all times, or adiabatic expansion which results in cooling followed by heating back to the initial temperature while holding volume fixed.

Each of these paths can be altered by making the gas do varying amounts of work by pushing out a piston during the expansion, so that an extremely large number of paths can be followed even for such a simple example. The detailed path must be specified if the heat or work is to be a known quantity; however, changes in the thermodynamic properties depend only on the initial and final states and not upon the path.

There are several corollaries from the above descriptions of systems, boundaries, states, and processes. First, all thermodynamic properties are identical for identical states. Second, the change in a property between initial and final states is independent of path or processes. The third corollary is that a quantity whose change is fixed by the end states and is independent of the path is a point function or a property. However, it must be remembered that by the second law of thermodynamics not all states are available (possible final states) from a given initial state and not all conceivable paths are possible in going toward an available state.

Pressure-volume-temperature diagram. Whereas the state of a system is a point function, the change of state of a system, or a process, is a path function. Various processes or methods of change of a system from one state to another may be depicted graphically as a path on a plot using thermodynamic properties as coordinates.

The variable properties most frequently and conveniently measured are pressure, volume, and temperature. If any two of these are held fixed (independent variables), the third is determined (dependent variable). To depict the relationship among these physical properties of the particular working substance, these three variables may be used as the coordinates of a three-dimensional space. The resulting surface is a graphic presentation of the equation of state for this working substance, and all possible equilibrium states of the substance lie on this *P-V-T* surface. The *P-V-T* surface may be extensive enough to include all three phases of the working substance: solid, liquid, and vapor.

Because a *P-V-T* surface represents all equilibrium conditions of the working substance, any line on the surface represents a possible reversible process, or a succession of equilibrium states.

The portion of the *P-V-T* surface shown in **Fig. 1** typifies most real substances; it is characterized by contraction of the substance on freezing. Going from the liquid surface to the liquid-solid surface onto the solid surface involves a decrease in both temperature and volume. Water is one of the few exceptions to this condition; it expands upon freezing, and its resultant *P-V-T* surface is somewhat modified where the solid and liquid phases abut.

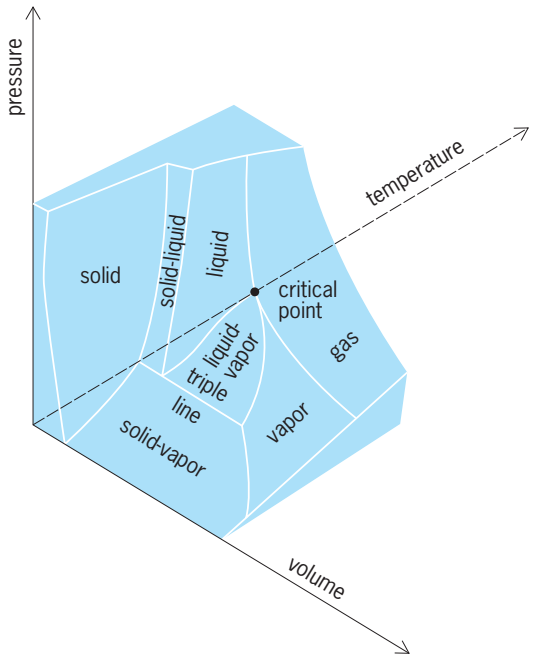


Fig. 1. Portion of pressure-volume-temperature (*P-V-T*) surface for a typical substance.

Gibbs' phase rule is defined in Eq. (1). Here f is the

$$f = c - p + 2 \tag{1}$$

degree of freedom; this integer states the number of intensive properties (such as temperature, pressure, and mole fractions or chemical potentials of the components) which can be varied independently of each other and thereby fix the particular equilibrium state of the system (see discussion under "Temperature-entropy diagram" below). Also, p indicates the number of phases (gas, liquid, or solid) and c the number of component substances in the system. Consider a one-component system (a pure substance) which is either in the liquid, gaseous, or solid phase. In equilibrium the system has two degrees of freedom; that is, two independent thermodynamic properties must be chosen to specify the state. Among the thermodynamic properties of a substance which can be quantitatively evaluated are the pressure, temperature, specific volume, internal energy, enthalpy, and entropy. From among these properties, any two may be selected. If these two prove to be independent of each other, when the values of these two properties are fixed, the state is determined and the values of all the other properties are also fixed. A one-component system with two phases in equilibrium (such as liquid in equilibrium with its vapor in a closed vessel) has $f = 1$; that is, only one intensive property can be independently specified. Also, a one-component system with three phases in equilibrium has no degree of freedom. Examination of Fig. 1 shows that the three surfaces (solid-liquid, solid-vapor, and liquid-vapor) are generated by lines parallel to the volume axis. Moving the system along such lines (constant pressure and temperature) involves a heat exchange and a change in the relative proportion of the two phases. Note that there is an entropy increment associated with this change.

One can project the three-dimensional surface onto the *P-T* plane as in **Fig. 2.** The triple point is

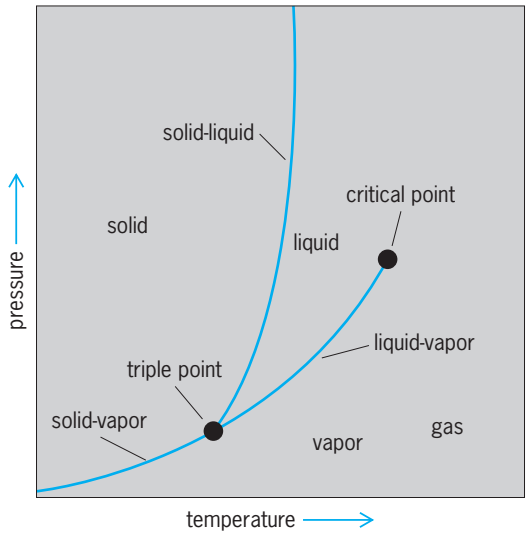


Fig. 2. Portion of equilibrium surface projected on pressure temperature (*P-T*) plane.

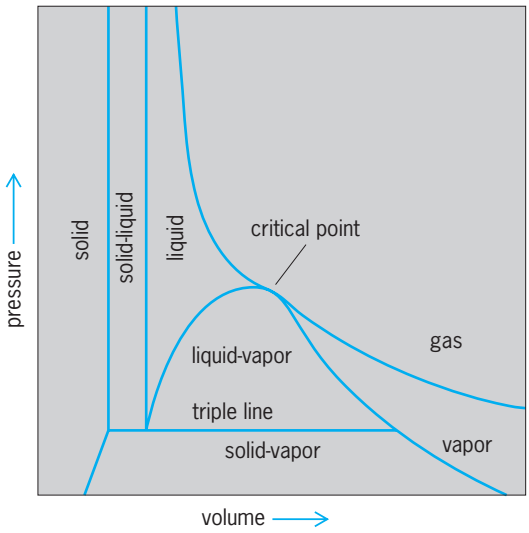


Fig. 3. Portion of equilibrium surface projected on pressure-volume (*P-V*) plane.

the point where the three phases are in equilibrium. When the temperature exceeds the critical temperature (at the critical point), only the gaseous phase is possible. The gas is called a vapor when it can coexist with another phase (at temperatures below the critical point). The *P-T* diagram for water would have the solid-liquid curve going upward from the triple point to the left (contrary to the ordinary substance pictured in Fig. 2). Then the property so well known to ice skaters would be evident. As the solid-liquid line is crossed from the low-pressure side to the high-pressure side, the water changes from solid to liquid: ice melts upon application of pressure.

Work of a process. The three-dimensional surface can also be projected onto the *P-V* plane to get **Fig. 3**. This plot has a special significance. The area under any reversible path on this plane represents the work done during the process. The fact that this *P-V* area represents useful work can be demonstrated by the following example.

Let a gas undergo an infinitesimal expansion in a cylinder equipped with a frictionless piston, and let this expansion perform useful work on the surroundings. The work done during this infinitesimal expansion is the force multiplied by the distance through which it acts, as in Eq. (2), wherein dW is

$$dW = F\,dl \tag{2}$$

an infinitesimally small work quantity, F is the force, and dl is the infinitesimal distance through which F acts.

But force F is equal to the pressure P of the fluid times the area A of the piston, or PA. However, the product of the area of the piston times the infinitesimal displacement is really the infinitesimal volume swept by the piston, or $A\,dl = dV$, with dV equal to an infinitesimal volume. Thus Eq. (3) is valid. The

$$dW = PA\,dl = P\,dV \tag{3}$$

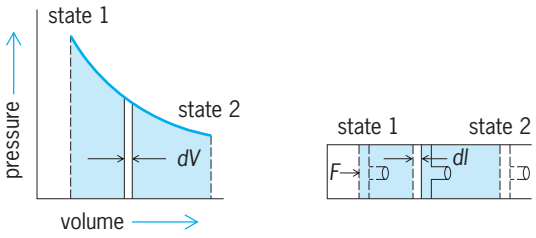


Fig. 4. Area under path in *P-V* plane is work done by expanding gas against piston.

work term is found by integration, as in Eq. (4).

$$ {}_1W_2 = \int_1^2 P\,dV \tag{4} $$

Figure 4 shows that the integral represents the area under the path described by the expansion from state 1 to state 2 on the *P-V* plane. Thus, the area on the *P-V* plane represents work done during this expansion process.

Temperature-entropy diagram. Energy quantities may be depicted as the product of two factors: an intensive property and an extensive one. Examples of intensive properties are pressure, temperature, and magnetic field; extensive ones are volume, magnetization, and mass. Thus, in differential form, work has been presented as the product of a pressure exerted against an area which sweeps through an infinitesimal volume, as in Eq. (5). Note that as a gas

$$ dW = P\,dV \tag{5} $$

expands, it is doing work on its environment. However, a number of different kinds of work are known. For example, one could have work on polarization of a dielectric, of magnetization, of stretching a wire, or of making new surface area. In all cases, the infinitesimal work is given by Eq. (6), where X is a generalized

$$ dW = X\,dx \tag{6} $$

applied force which is an intensive quantity such as voltage, magnetic field, or surface tension; and dx is a generalized displacement of the system and is thus extensive. Examples of dx include changes in electric polarization, magnetization, length of a stretched wire, or surface area.

By extending this approach, one can depict transferred heat as the product of an intensive property, temperature, and a distributed or extensive property, defined as entropy, for which the symbol is S. *See* ENTROPY.

If an infinitesimal quantity of heat dQ is transferred during a reversible process, this process may be expressed mathematically as in Eq. (7), with T being

$$ dQ = T\,dS \tag{7} $$

the absolute temperature and dS the infinitesimal entropy quantity.

Furthermore, a plot of the change of state of the system undergoing this reversible heat transfer can be drawn on a plane in which the coordinates are absolute temperature and entropy (**Fig. 5**). The total heat transferred during this process equals the area between this plotted line and the horizontal axis.

Reversible processes. Not all energy contained in or associated with a mass can be converted into useful work. Under ideal conditions only a fraction of the total energy present can be converted into work. The ideal conversions which retain the maximum available useful energy are reversible processes.

Characteristics of a reversible process are that the working substance is always in thermodynamic equilibrium and the process involves no dissipative effects such as viscosity, friction, inelasticity, electrical resistance, or magnetic hysteresis. Thus, reversible processes proceed quasistatically so that the system passes through a series of states of thermodynamic equilibrium, both internally and with its surroundings. This series of states may be traversed just as well in one direction as in the other.

If there are no dissipative effects, all useful work done by the system during a process in one direction can be returned to the system during the reverse process. When such a process is reversed so that the system returns to its starting state, it must leave an effect on the surroundings since, by the second law of thermodynamics, in energy conversion processes the form of energy is always degraded. Part of the energy of the system (including heat source) is transferred as heat from a higher temperature to a lower temperature. The energy rejected to a lower-temperature heat sink cannot be recovered. To return the system (including heat source and sink) to its original state, then, requires more energy than the useful work done by the system during a process in one direction. Of course, if the process were purely a mechanical one with no thermal effects, then both the surroundings and system could be returned to their initial states. *See* CARNOT CYCLE; THERMODYNAMIC CYCLE.

It is impossible to satisfy the conditions of a quasistatic process with no dissipative effects; a reversible process is an ideal abstraction which is not realizable in practice but is useful for theoretical calculations. An ideal reversible engine operating between hotter and cooler bodies at the temperatures

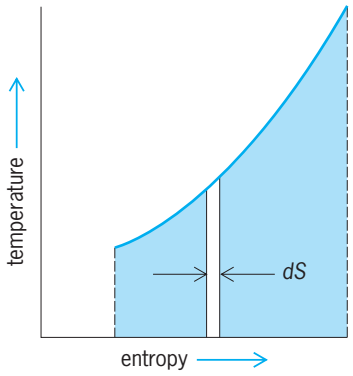


Fig. 5. Heat transferred during a reversible process is area under path in temperature-entropy (*T-S*) plane.

T_1 and T_2, respectively, can put out $(T_1 - T)/T_1$ of the transferred heat energy as useful work.

There are four reversible processes wherein one of the common thermodynamic parameters is kept constant. The general reversible process for a closed or nonflow system is described as a polytropic process. *See* ISENTROPIC PROCESS; ISOBARIC PROCESS; ISOMETRIC PROCESS; ISOTHERMAL PROCESS; POLYTROPIC PROCESS.

Irreversible processes. Actual changes of a system deviate from the idealized situation of a quasistatic process devoid of dissipative effects. The extent of the deviation from ideality is correspondingly the extent of the irreversibility of the process.

Real expansions take place in finite time, not infinitely slowly, and these expansions occur with friction of rubbing parts, turbulence of the fluid, pressure waves sweeping across and rebounding through the cylinder, and finite temperature gradients driving the transferred heat. These dissipative effects, the kind of effects that make a pendulum or yo-yo slow down and stop, also make the work output of actual irreversible expansions less than the maximum ideal work of a corresponding reversible process. For a reversible process, as stated earlier, the entropy change is given by $dS = dQ/T$. For an irreversible process even more entropy is produced (turbulence and loss of information) and there is the inequality $dS > dQ/T$. Philip E. Bloomfield; William A. Steele

Bibliography. H. A. Bent, *The Second Law*, 1965; Y. A. Cengel and M. A. Boles, *Thermodynamics: An Engineering Approach*, 3d ed., 1998; J. P. Holman, *Thermodynamics,* 4th ed., 1988; M. Mott-Smith, *The Concept of Energy Simply Explained*, 1934; F. W. Sears and G. L. Salinger, *Thermodynamics, the Kinetic Theory of Gases and Statistical Mechanics,* 3d ed., 1975; K. Wark, *Thermodynamics,* 6th ed., 1999.

Thermoelectric power generator

A solid-state heat engine which employs the electron gas as a working fluid. It directly converts heat energy into electrical energy using the Seebeck effect. This phenomenon can be demonstrated using a thermocouple which comprises two legs (thermoelements) of dissimilar conducting materials joined at one end to form a junction. If this junction is maintained at a temperature which differs from ambient, a voltage is generated across the open ends of the thermoelements. When the circuit is completed with a load, a current flows in the circuit and power is generated. In practice the thermocouples are fabricated generally from *n*- and *p*-type semiconductors, and several hundred are connected electrically in series to form a module which is the active component of a thermoelectric generator. Provided a temperature difference is maintained across the device, it will generate electrical power. Heat is provided from a variety of sources depending on the application, and they include burning fossil fuels in terrestrial and military applications, decaying long-life isotopes in medical and deep-space applications, and waste heat. The performance of the thermoelectric generator, in terms of efficiency, output power, and economic viability, depends upon its temperature regime of operation; the materials used in the module construction; its electrical, thermal, and geometrical design; and the generator load. The power output spectrum of thermoelectric generators spans 14 orders of magnitude and ranges from nanowatt generators fabricated using integrated circuit technology to the nuclear reactor–powered 100-kW SP-100 generator intended to provide electrical power to orbiting space stations. *See* NUCLEAR BATTERY; RADIOACTIVITY AND RADIATION APPLICATIONS; SEEBECK EFFECT; SPACE POWER SYSTEMS; SPACE PROBE; THERMOCOUPLE; THERMOELECTRICITY.

Generating parameters. A thermoelectric generator is a heat engine, and like all heat engines it obeys the laws of thermodynamics. The efficiency of an ideal thermoelectric generator, one in which there are no heat losses, is defined as the ratio of the electrical power delivered to the load to the heat absorbed at the hot junction. *See* EFFICIENCY; THERMODYNAMIC PRINCIPLES; THERMODYNAMIC PROCESSES.

Expressions for the important parameters in thermoelectric generation can readily be derived by considering the simplest generator consisting of a single thermocouple with thermoelements fabricated from *n*- and *p*-type semiconductors respectively (**Fig. 1**). The efficiency of the generator is given by Eq. (1). If it

$$\phi = \frac{\text{energy supplied to the load}}{\text{heat energy absorbed at hot junction}} \quad (1)$$

is assumed that the electrical conductivities, thermal conductivities, and Seebeck coefficients of the thermoelements *a* and *b* are constant within an arm, and that the contact resistances at the hot and cold junctions are negligible compared with the sum of the arm resistances, then the efficiency can be expressed

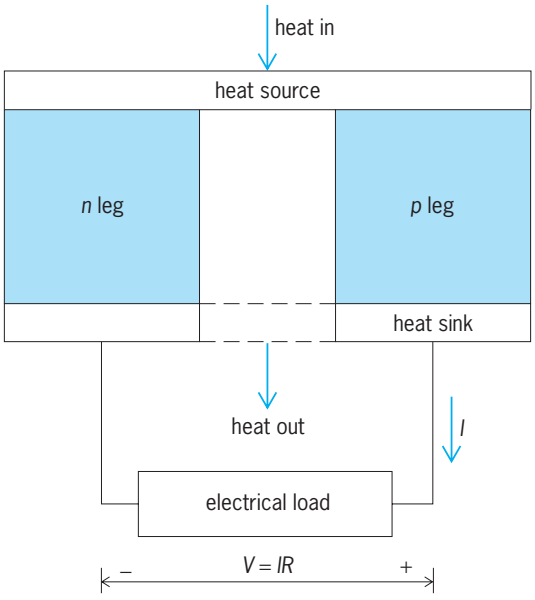


Fig. 1. Diagram of single-couple thermoelectric generator.

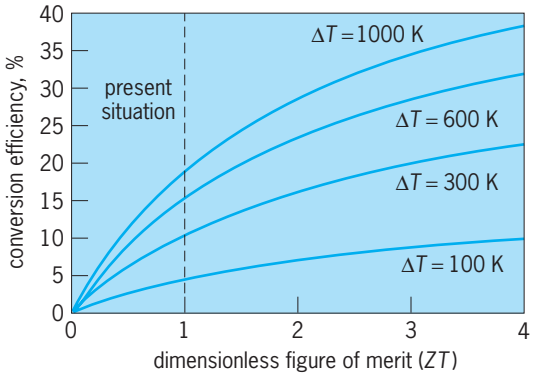


Fig. 2. Conversion efficiency as a function of the dimensionless figure of merit, *ZT*, for temperature differences of operation, ΔT, with cold junction Z_2 at 300 K (80°F). (For temperature differences, 1 K = 1.8°F.)

as Eq. (2), where I is the current, α_{ab} is the total See-

$$\phi = \frac{I^2 R_L}{\alpha_{ab} I T_1 + \lambda'(T_1 - T_2) - \frac{1}{2} I^2 R} \tag{2}$$

beck coefficient of a and b, λ' is the thermal conductance of a and b in parallel, R is the series resistance of a and b, R_L is the load resistance, and T_1 and T_2 are the absolute temperatures of the hot and cold junctions. In thermoelectric materials, the electrical conductivity, thermal conductivity, and seebeck coefficient change with temperature, and in actual application this is taken into account. Appropriate allowances can also be made for contact resistances. However, the simple expression for the efficiency, Eq. (2), can still be employed with an acceptable degree of accuracy. Assuming average values for these parameters provides results which are within 10%

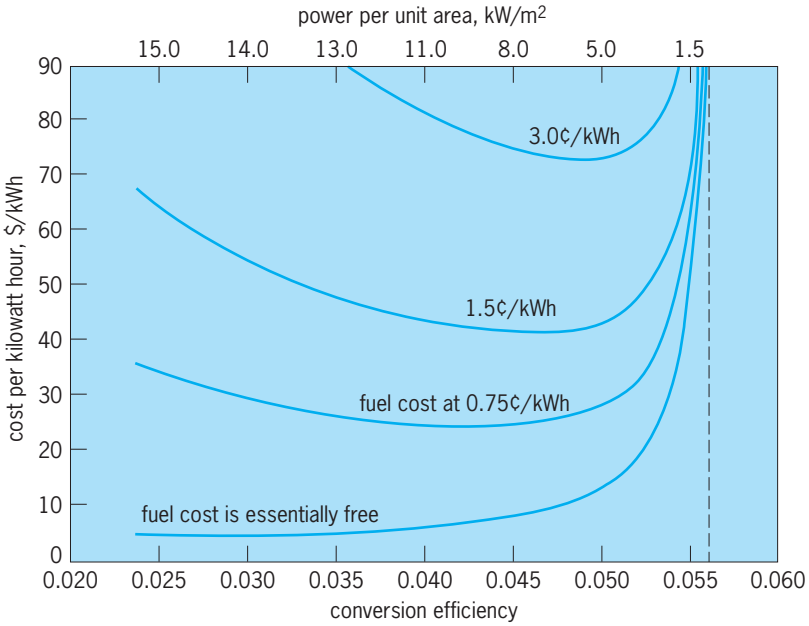


Fig. 3. Cost per killowatthour as a function of conversion efficiency and power per unit area for different fuel costs. The calculation is based on a typical 127-pair thermocouple module operating at ΔT = 120 K (216°F) for 15 years.

of the true value. *See* CONDUCTION (ELECTRICITY); CONDUCTION (HEAT); SEMICONDUCTOR.

Efficiency is clearly a function of the ratio of the load resistance to the sum of the generator arm resistances. At maximum power output, it can be shown that the efficiency is given in Eq. (3), where Z_c is the figure of merit of the couple, given by Eq. (4). The maximum efficiency is given by Eq. (5), where n_c,

$$\phi_p = \frac{T_1 - T_2}{\frac{3T_1}{2} + \frac{T_2}{2} + \frac{4}{Z_c}} \tag{3}$$

$$Z_c = \frac{\alpha_{ab}^2}{R\lambda'} \tag{4}$$

$$\phi_{\max} = \eta_c \gamma \tag{5}$$

the Carnot efficiency, is given by Eq. (6), and γ is given by Eqs. (7) and (8). The maximum efficiency

$$\eta_c = \frac{T_1 - T_2}{T_1} \tag{6}$$

$$\gamma = \frac{\sqrt{1 + Z_c \overline{T}} - 1}{\sqrt{1 + Z_c \overline{T}} + \frac{T_2}{T_1}} \tag{7}$$

$$\overline{T} = \frac{T_1 + T_2}{2} \tag{8}$$

is thus the product of the Carnot efficiency, which is clearly less than unity, and γ, which embodies the parameters of the materials. *See* CARNOT CYCLE.

If the geometries of a and b are matched to minimize heat absorption, then the figure of merit is given by Eq. (9), where σ_a and σ_b are the electrical con-

$$Z_c = \frac{\alpha_{ab}^2}{\left[\left(\frac{\lambda_a}{\sigma_a}\right)^{1/2} + \left(\frac{\lambda_b}{\sigma_b}\right)^{1/2}\right]^2} \tag{9}$$

ductivities a and b, and λ_a and λ_b are their thermal conductivities. In practice the two arms of the junction have similar material constants, in which case the concept of a figure of merit for a material is employed and is given by Eq. (10). The quantity Z varies

$$Z = \frac{\alpha^2 \sigma}{\lambda} \tag{10}$$

with temperature, and a more meaningful parameter is the dimensionless figure of merit ZT, where T is the absolute temperature at which the performance of the material is considered; $\alpha^2\sigma$ is referred to as the electrical power factor.

Improving conversion efficiency. Historically the use of thermoelectric generators has been restricted to specialized applications where combinations of their desirable properties, such as the absence of moving parts, reliability, and silent operation, has outweighed their low overall conversion efficiency, typically 5%. In these applications, fuel cost or weight is a major consideration, and improving the conversion efficiency is the main research target. The effect of ZT on the generating conversion efficiency for different temperature differences of operation is

displayed in **Fig. 2**. Understandably, improving the figure of merit has been regarded as the most important factor in increasing the conversion efficiency of a thermoelectric generator.

Cost reduction. Consideration has been given to applications when the fuel cost is low or essentially free as with waste heat. The economic viability is measured by the cost per kilowatthour of electricity C. In the case of a thermoelectric generator, it can be shown that C is related to the generators fabrication cost C_g and the running cost (such as fuel) C_r by Eq. (11), where p and Δt are the power and operating period. The relative importance of C_g and C_r for the fuel cost is shown in **Fig. 3**. When the fuel is relatively expensive, the module should be optimized to achieve high conversion efficiency. However, when the fuel cost is low or essentially free as in waste heat recovery, then the cost per watt is mainly determined by the power per unit area and the operating period.

$$C = \frac{C_g}{p\Delta t} + \frac{C_r}{\phi} \tag{11}$$

Materials. Selection of thermocouple material depends upon the generator's temperature regime of operation. The figures of merit of established thermoelectric materials reach maxima at different temperatures, and semiconductor compounds or alloys based on bismuth telluride, lead telluride, and silicon germanium cover the temperature ranges up to 150°C (300°F), 650°C (1200°F), and 1000°C (1830°F) respectively, with the best materials capable of generating electrical power with an efficiency of around 20%. Material research is focused on improving the figure of merit and to a lesser, though an increasing, extent, the electrical power factor.

Improving the figure of merit. Although no theoretical upper boundary to the dimensionless figure of merit ZT exists, phenomenological evidence points to the existence of a barrier around $ZT = 2$. Nevertheless, improvement in the figure of merit is being sought primarily by reducing the lattice thermal conductivity. The bulk materials being researched include glasslike crystals such as the filled skutterudites within whose atomic cage weakly bound atoms or molecules rattle and should conduct heat like a glass but electricity like a crystal. The quest for improving the figure of merit by reducing the thermal conductivity has moved to exotic low-dimensional structures such as superlattices, quantum wires, and quantum dots, where in theory the reduced dimensions and the presence of interfaces give rise to a thermal conductivity which is lower than the bulk counterpart. A ZT close to 2 has been reported but as yet has not been confirmed independently. *See* ARTIFICIALLY LAYERED STRUCTURES; MESOSCOPIC PHYSICS; QUANTIZED ELECTRONIC STRUCTURE (QUEST); THERMAL CONDUCTION IN SOLIDS.

Improving the power factor. The emergence of thermoelectrics as a technology for application in waste heat recovery has resulted in a successful search for materials with high electrical power factors and cheap materials. The rare-earth ytterbium-aluminum compound $YbAl_3$ has a power factor almost three times that of bismuth telluride, the established material for low-temperature application, while magnesium tin (MgSn) has almost the same performance as lead telluride but is available at less than a quarter of the cost.

Applications. Thermoelectric generators continue to find novel applications, such as miniature thermocouple arrays to power a wristwatch using body heat and the utilization of automobile exhaust heat to thermoelectrically supplement the electrical power supply. Thermoelectric generation is an established technology and in some applications has become economically competitive, particularly in waste heat recovery.

D. M. Rowe

Bibliography. D. M. Rowe (ed.), *CRC Handbook of Thermoelectrics*, CRC Press, 1995; D. M. Rowe and C. M. Bhandari (eds.), *Modern Thermoelectrics*, Prentice Hall, 1983.

Thermoelectricity

The direct conversion of heat into electrical energy, or the reverse, in solid or liquid conductors by means of three interrelated phenomena—the Seebeck effect, the Peltier effect, and the Thomson effect—including the influence of magnetic fields upon each. The Seebeck effect concerns the electromotive force (emf) generated in a circuit composed of two different conductors whose junctions are maintained at different temperatures. The Peltier effect refers to the reversible heat generated at the junction between two different conductors when a current passes through the junction. The Thomson effect involves the reversible generation of heat in a single current-carrying conductor along which a temperature gradient is maintained. Specifically excluded from the definition of thermoelectricity are the phenomena of Joule heating and thermionic emission. *See* ELECTROMOTIVE FORCE (EMF); JOULE'S LAW; THERMIONIC EMISSION.

The three thermoelectric effects are described in terms of three coefficients: the absolute thermoelectric power (or thermopower) S, the Peltier coefficient Π, and the Thomson coefficient μ, each of which is defined for a homogeneous conductor at a given temperature. These coefficients are connected by the Kelvin relations, which convert complete knowledge of one into complete knowledge of all three. It is therefore necessary to measure only one of the three coefficients; usually the thermopower S is chosen. The combination of electrical resistivity, thermal conductivity, and thermopower (more generally, these quantities as tensors) is sufficient to provide a complete description of the electronic transport properties of conductors for which the electric current and heat current are linear functions of both the applied electric field and the temperature gradient. *See* ELECTRICAL CONDUCTIVITY OF METALS.

Thermoelectric effects have significant applications in both science and technology and show

promise of more importance in the future. Studies of thermoelectricity in metals and semiconductors yield information about electronic structure and about the interactions between electrons and both lattice vibrations and impurities. Practical applications include the measurement of temperature, generation of power, cooling, and heating. Thermocouples are widely used for temperature measurement, providing both accuracy and sensitivity. Research has been undertaken concerning the direct thermoelectric generation of electricity using the heat produced by nuclear reactions or generated at automobile exhausts. Cooling units using the Peltier effect have been constructed in sizes up to those of home refrigerators. Development of thermoelectric heating has also been undertaken.

Seebeck Effect

In 1821, T. J. Seebeck discovered that when two different conductors are joined into a loop and a temperature difference is maintained between the two junctions, an emf is generated. Such a loop is called a thermocouple, and the emf generated is called a thermoelectric (or Seebeck) emf.

Measurements. The magnitude of the emf generated by a thermocouple is standardly measured using the system shown in **Fig. 1**. Here the contact points between conductors A and B are called junctions. Each junction is maintained at a well-controlled temperature (either T_1 or T_0) by immersion in a bath or connection to a heat reservoir. These baths or reservoirs are indicated by the white squares. From each junction, conductor A is brought to a measuring device M, usually a potentiometer. When the potentiometer is balanced, no current flows, thereby allowing direct measurement of the open-circuit emf, undiminished by resistive losses and unperturbed by spurious effects arising from Joule heating or from Peltier heating and cooling at the junctions. This open-circuit emf is the thermoelectric emf. *See* POTENTIOMETER.

Equations. According to the experimentally established law of Magnus, for homogeneous conductors A and B the thermoelectric emf depends only upon the temperatures of the two junctions and not upon either the shapes of the samples or the detailed forms of the temperature distributions along them. This emf can thus be symbolized $E_{AB}(T_0,T_1)$. According to both theory and experiment, if one of the conductors, say B, is a superconductor in its superconducting state, it makes no contribution to E_{AB} (except for very small effects near the superconducting transition temperature T_c, discussed below). That is, when B is superconducting, $E_{AB}(T_0,T_1)$ is determined solely by conductor A, and can be written as $E_A(T_0,T_1)$. *See* SUPERCONDUCTIVITY.

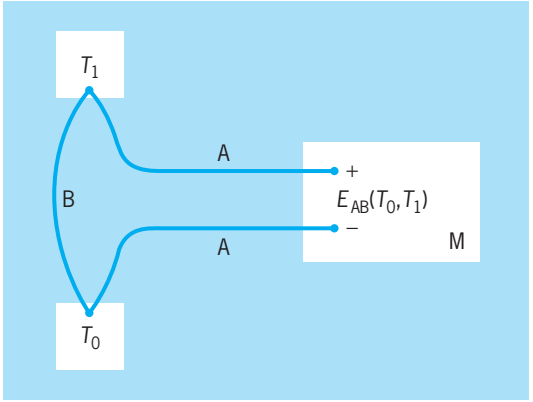


Fig. 1. Diagram of apparatus usually used for measuring thermoelectric (Seebeck) emf $E_{AB}(T_0,T_1)$. M is an instrument for measuring potential.

It is convenient to express this emf in terms of a property that depends only upon a single temperature. Such a property is the absolute thermoelectric power (or, simply, thermopower) $S_A(T)$, defined so that Eq. (1) is valid.

$$E_A(T_0, T_1) = \int_{T_0}^{T_1} S_A(T)\, dT \tag{1}$$

If $E_A(T, T+\Delta T)$ is known—for example, from measurements involving a superconductor—$S_A(T)$ can be determined from Eq. (2). If Eq. (1) is valid for any

$$S_A(T) = \lim_{\Delta T \to 0} \frac{E_A(T, T+\Delta T)}{\Delta T} \tag{2}$$

homogeneous conductor, then it ought to apply to both sides of the thermocouple shown in Fig. 1. Indeed, it has been verified experimentally that the emf $E_{AB}(T_0,T_1)$ produced by a thermocouple is just the difference between the emf's, calculated using Eq. (1), produced by its two arms. This result can be derived as follows. Employing the usual sign convention, to calculate $E_{AB}(T_0,T_1)$, begin at the cooler bath, T_0, integrate $S_A(T)\, dT$ along conductor A up to the warmer bath, T_1, and then return to T_0 along conductor B by integrating $S_B(T)\, dT$. This circular excursion produces $E_{AB}(T_0,T_1)$, given by Eq. (3). Inverting the last integral in Eq. (3) gives Eq. (4), which from Eq. (1) can be rewritten as Eq. (5). Alternatively, combining the two integrals in Eq. (4) gives Eq. (6). Defining S_{AB} according to Eq. (7) then yields Eq. (8).

$$E_{AB}(T_0, T_1) = \int_{T_0}^{T_1} S_A(T)\, dT + \int_{T_1}^{T_0} S_B(T)\, dT \tag{3}$$

$$E_{AB}(T_0, T_1) = \int_{T_0}^{T_1} S_A(T)\, dT - \int_{T_0}^{T_1} S_B(T)\, dT \tag{4}$$

$$E_{AB}(T_0, T_1) = E_A(T_0, T_1) - E_B(T_0, T_1) \tag{5}$$

$$E_{AB}(T_0, T_1) = \int_{T_0}^{T_1} [S_A(T)] - S_B(T)]\, dT \tag{6}$$

$$S_{AB}(T) = S_A(T) - S_B(T) \tag{7}$$

$$E_{AB}(T_0, T_1) = \int_{T_0}^{T_1} S_{AB}(T)\, dT \tag{8}$$

Equation (6) shows that $E_{AB}(T_0,T_1)$ can be calculated for a given thermocouple whenever the thermopowers $S_A(T)$ and $S_B(T)$ are known for its two constitutents over the temperature range T_0 to T_1.

By convention, the signs of $S_A(T)$ and $S_B(T)$ are chosen so that, if the temperature difference $T_1 - T_0$ is taken small enough so that $S_A(T)$ and $S_B(T)$ can be presumed constant, then $S_A(T) > S_B(T)$ when the emf $E_{AB}(T_0,T_1)$ has the polarity indicated in Fig. 1.

Results of equations. These equations lead directly to the following experimentally and theoretically verified results.

Uniform temperature. In a circuit kept at a uniform temperature throughout, $E = 0$, even though the circuit may consist of a number of different conductors. This result follows directly from Eq. (8), since $dT = 0$ everywhere throughout the circuit. It follows also from thermodynamic reasoning. If E did not equal 0, the circuit could drive an electrical motor and make it do work. But the only source of energy would be heat from the surroundings which, by assumption, are at the same uniform temperature as the circuit. Thus, a contradiction with the second law of thermodynamics would result. *See* CHEMICAL THERMODYNAMICS.

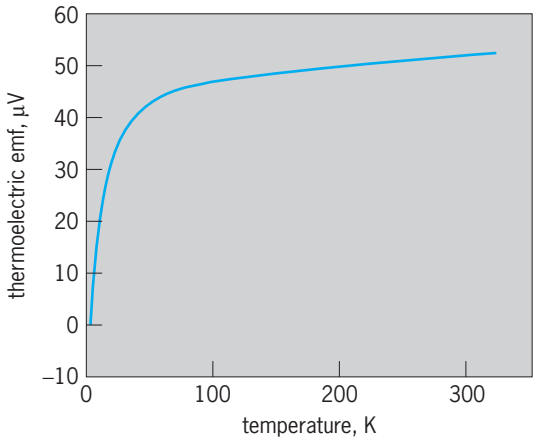


Fig. 2. Thermoelectric emf of a thermocouple formed from pure annealed and pure cold-worked copper. The cold junction reference temperature is 4.2 K (−452°F). °F = (K × 1.8) − 459.67. (*After R. H. Kropschot and F. J. Blatt, Thermoelectric power of cold-rolled pure copper, Phys. Rev., 116:617–620, 1959*)

Homogeneous conductor. A circuit composed of a single, homogeneous conductor cannot produce a thermoelectric emf. This follows from Eq. (6) when $S_B(T)$ is set equal to $S_A(T)$. It is important to emphasize that in this context "homogeneous" means perfectly uniform throughout. A sample made of an isotropic material can be inhomogeneous either because of small variations in chemical composition or because of strain. **Figure 2** shows the thermoelectric emf generated by a thermocouple in which one arm is a cold-rolled copper (Cu) sample, and the other arm is the same material after annealing at an elevated temperature to remove the effects of the strain introduced by the cold-rolling. **Figure 3** shows how the addition of impurities can change the thermopower of a pure metal. An additional effect can occur in a noncubic material. As illustrated in **Fig. 4**, a thermocouple formed from two samples cut in different

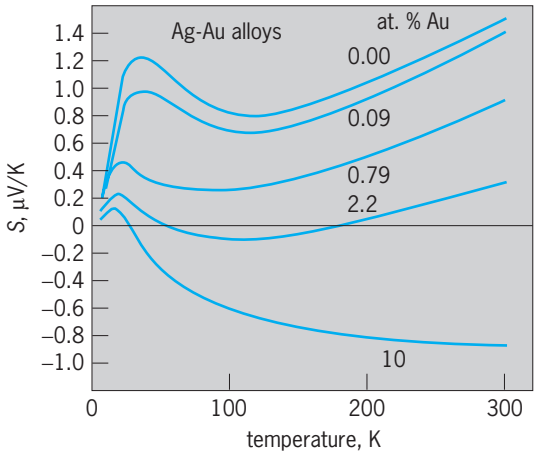


Fig. 3. Thermopower *S* from 0 to 300 K for pure silver (Ag) and a series of dilute silver-gold (Ag-Au) alloys, °F = (K × 1.8) − 459.67. (*After R. S. Crisp and J. Rungis, Thermoelectric power and thermal conductivity in the silver-gold alloy system from 3–300°K, Phil. Mag., 22:217–236, 1970*)

orientations from a noncubic single crystal can generate a thermoelectric emf even if each sample is highly homogeneous. *See* CRYSTAL STRUCTURE.

If material B is superconducting, so that $S_B = 0$, Eq. (5) reduces to $E_{AB}(T_0,T_1) = E_A(T_0,T_1)$, as assumed above.

Source of emf. Finally, Eq. (6) makes clear that the source of the thermoelectric emf in a thermocouple lies in the bodies of the two materials of which it is composed, rather than at the junctions. This serves to emphasize that thermoelectric emf's are not related to the contact potential or Volta effect, which is a potential difference across the junction between two different metals arising from the difference between their Fermi energies. The contact potential is present even in the absence of temperature gradients or electric currents. *See* CONTACT POTENTIAL DIFFERENCE.

Peltier Effect

In 1834, C. A. Peltier discovered that when an electric current passes through two different conductors

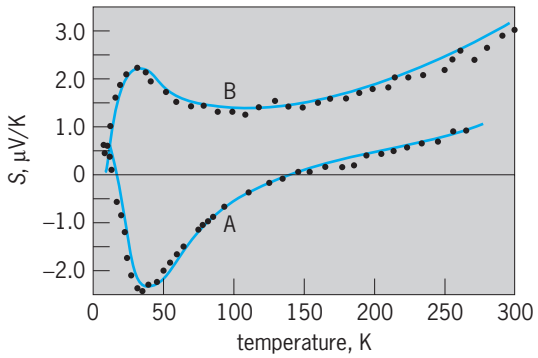


Fig. 4. Thermopower *S* of zinc parallel (A) and perpendicular (B) to the hexagonal axis. °F = (K × 1.8) − 459.67. (*After V. A. Rowe and P. A. Schroeder, Thermopower of Mg, Cd and Zn between 1.2° and 300°K, J. Phys. Chem. Sol., 31:1–8, 1970*)

connected in a loop, one of the two junctions between the conductors cools and the other warms. If the direction of the current is reversed, the effect also reverses: the first junction warms and the second cools. In 1853, Quintus Icilius showed that the rate of heat output or intake at each junction is directly proportional to the current i. The Peltier coefficient Π_{AB} is defined as the heat generated per second per unit current flow through the junction between materials A and B. By convention, Π_{AB} is taken to be positive when cooling occurs at the junction through which current flows from conductor A to conductor B. Quintus Icilius's result guarantees that the Peltier coefficient is independent of the magnitude of the current i. Additional experiments have shown that it is also independent of the shapes of the conductors. It therefore depends only upon the two materials and the temperature of the junction, and can be written as $\Pi_{AB}(T)$ or, alternatively, as $\Pi_A(T) - \Pi_B(T)$, where $\Pi_A(T)$ and $\Pi_A(T)$ are the Peltier coefficients for materials A and B, respectively. The second form emphasizes that the Peltier coefficient is a bulk property which can be defined for a single conductor.

Because of the difficulty of measuring heat input or output from a junction, as well as complications resulting from the simultaneous presence of Joule heating and the Thomson effect, $\Pi_{AB}(T)$ has rarely been quantitatively measured. Rather its value is usually determined from the Kelvin relations, using experimental values for S_{AB}. The Peltier effect does, however, form the basis for thermoelectric heating and cooling.

Thomson Effect and Kelvin Relations

When an electric current passes through a conductor that is maintained at a constant temperature, heat is generated at a rate proportional to the square of the current. This phenomenon is called Joule heat, and its magnitude for any given material is determined by the electrical resistivity of the material. In 1854, William Thomson (Lord Kelvin), in an attempt to explain discrepancies between experimental results and a relationship between Π_{AB} and S_{AB} that he had derived from thermodynamic analysis of a thermocouple, postulated the existence of an additional reversible generation of heat when a temperature gradient is applied to a current-carrying conductor. This heat, called Thomson heat, is proportional to the product of the current and the temperature gradient. It is reversible in the sense that the conductor changes from a generator of Thomson heat to an absorber of Thomson heat when the direction of either the current or the temperature gradient (but not both at once) is reversed. By contrast, Joule heating is irreversible in that heat is generated for both directions of current flow.

The magnitude of Thomson heat generated (or absorbed) is determined by the Thomson coefficient μ. Using reasoning based upon equilibrium thermodynamics, Thomson derived results equivalent to Eqs. (9) and (10), called the Kelvin (or Kelvin-Onsager) relations.

$$\frac{\Pi A}{T} = S_A \tag{9}$$

$$\frac{\mu A}{T} = \frac{dS_A}{dT} \tag{10}$$

Here μ_A is the Thomson coefficient, defined as the heat generated per second per unit current flow per unit temperature gradient when current flows through conductor A in the presence of a temperature gradient. Equation (10) can be integrated to give Eq. (11), in which the third law of thermodynamics

$$S_A = \int_0^T \frac{\mu_A(T')}{T'}\, dT \tag{11}$$

has been evoked to set $S_A(0) = 0$. Using Eq. (11), $S_A(T)$ can be determined from measurements on a single conductor. In practice, however, accurate measurements of μ_A are very difficult to make; therefore, they have been carried out for only a few metals—notably lead (Pb), platinum (Pt), and tungsten (W)—which then serve as standards for determining $S_B(T)$ by using measurements of $S_{AB}(T)$ in conjunction with Eq. (7).

Long after the Thomson heat was observed and the Kelvin relations were verified experimentally, debate raged over the validity of the derivation employed by Thomson. However, the theory of irreversible processes, developed by L. Onsager in 1931, and by others, yields the same equations and thus provides them with a firm foundation.

Thermopowers of Metals and Semiconductors

Since the Kelvin relations provide recipes for calculating any two of the thermoelectric coefficients, S, Π, and μ, from the third, only one of the three coefficients need be measured to determine the thermoelectric properties of any given material. Although there are circumstances under which one of the other two coefficients may be preferred, because of ease and accuracy of measurement, it is usually the thermopower S that is measured.

Reference materials. Because S must be measured using a thermocouple, the quantity measured experimentally is $S_A - S_B$, the difference between the thermopowers of the two conductors constituting the thermocouple. Only when one of the arms of the thermocouple is superconducting and therefore has zero thermopower can the absolute thermopower of the other be directly measured. At temperatures up to about 18 K (-427°F) superconducting niobium-tin (Nb_3Sn) wire can be used for conductor B, thereby permitting direct determination of S_A. Direct measurements have been extended to just above 90 K (-298°F) using the $YBa_2Cu_3O_7$ high-temperature superconductors, and should be extendable to above 120 K (-244°F) using still higher-temperature superconductors. For even higher temperatures, a standard thermoelectric material is needed. For historical reasons, the reference material up to about room temperature has been chosen to be Pb. Until the mid-1970s, the thermopower

TABLE 1. Absolute thermoelectric power of *S* of pure lead between 0 and 350 K*

T, K†	S_{Pb}, μV/K
0.000	0.000
5.000	0.000
7.500	−0.220
8.000	−0.256
8.500	−0.298
9.000	−0.343
10.000	−0.433
11.000	−0.517
12.000	−0.593
14.000	−0.707
16.000	−0.770
18.000	−0.786
20.000	−0.779
30.000	−0.657
40.000	−0.575
50.000	−0.537
60.000	−0.527
70.000	−0.531
80.000	−0.544
90.000	−0.562
100.000	−0.583
110.000	−0.606
130.000	−0.656
150.000	−0.708
170.000	−0.760
190.000	−0.810
210.000	−0.858
230.000	−0.904
250.000	−0.948
270.000	−0.989
290.000	−1.028
310.000	−1.065
330.000	−1.101
350.000	−1.136

*After R. B. Roberts, The absolute scale of thermoelectricity, *Phil. Mag.*, 36:91–107, 1977.
†°C = K − 273.15; °F = (K × 1.8) − 459.67.

of Pb was calculated from Thomson coefficient measurements made in the early 1930s, and all published values of *S* were ultimately traceable to those measurements. In 1977, new measurements of the Thomson coefficient of Pb were made up to 350 K (170°F). The revised values of S_{Pb} are listed in **Table 1**. Above 50 K (−370°F) they differ from the old values by 0.25–0.3 microvolt/K; older values of *S* must be corrected for these differences if accuracy is important. Measurements in 1985 of the Thomson coefficients of Pt and W allow these two metals to be used as references up to 1600 K (2900°F) or 1800 K (3250°F), respectively.

Temperature variation. **Figure 5** shows the variation with temperature of the thermopowers of four different pure metals. The data for gold (Au), aluminum (Al), and Pt are typical of those for most simple metals and for some transition metals as well. The thermopower *S* consists of a slowly varying portion that increases approximately linearly with absolute temperature, upon which is superimposed a "hump" at lower temperatures. In analyzing these results, *S* is written as the sum of two terms, as in Eq. (12), where

$$S = S_d + S_g \tag{12}$$

S_d, called the electron-diffusion component, is the slowly varying portion, and S_g, called the phonon-drag component, is the hump. For some transition metals, on the other hand, the behavior of *S* is more complex as illustrated by the data for rhodium (Rh) in Fig. 5. **Figure 6** shows comparative data for a sample *p*-type semiconductor. The separation of *S* into S_d and S_g is still valid, but at high temperatures S_d now varies more weakly than linearly with temperature. Note also the different ordinate scales in Fig. 5 (μV/K) and Fig. 6 (mV/K)—the thermopower of a semiconductor can be a thousand times larger than that of a metal.

Theory. When a small temperature difference ΔT is established across a conductor, heat is carried from

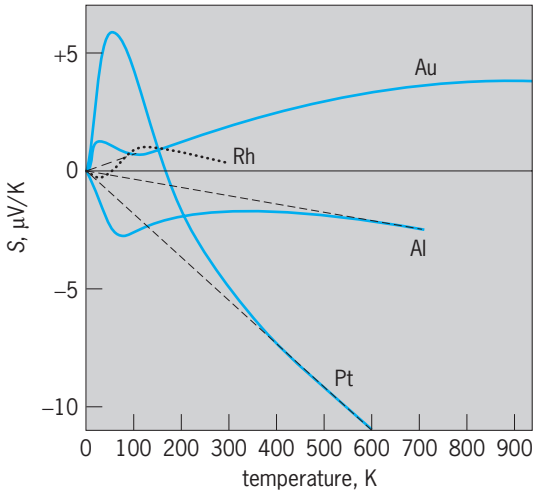


Fig. 5. Thermopower *S* of the metals gold (Au), aluminum (Al), platinum (Pt), and rhodium (Rh) as a function of temperature. The differences between the solid curves for Pt, Al, and Au and the broken lines indicate the magnitude of the phonon-drag component *S*. °F = (K × 1.8) − 459.67.

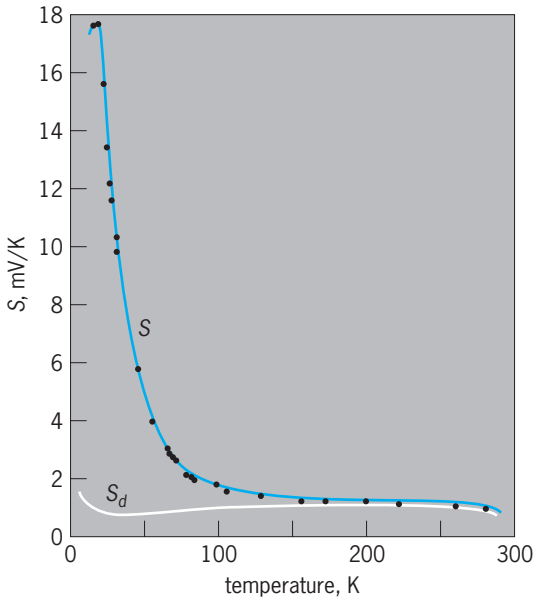


Fig. 6. Thermopower *S* of *p*-type germanium (1.5×10^{14} acceptors per cubic centimeter) and calculated value for the electron-diffusion thermopower S_d. (*After C. Harring, The role of low-frequency phonons in thermoelectricity and thermal conductivity, Proc. Int. Coll. 1956, Garmisch-Partenkirchen, Vieweg. Braunschweig, 1958*)

its hot end to its cold end by the flow of both electrons and phonons (quantized lattice vibrations). If the electron current is constrained to be zero—for example by the insertion of a high-resistance measuring device in series with the conductor—the electrons will redistribute themselves in space so as to produce an emf along the conductor. This is the thermoelectric emf. If the phonon flow could somehow be turned off, this emf would be just $S_d\Delta T$. However, the phonon flow cannot be turned off, and as the phonons move down the sample, they interact with the electrons and "drag" them along. This process produces an additional contribution to the emf, $S_g\Delta T$. *See* CONDUCTION (HEAT); LATTICE VIBRATIONS; PHONON; THERMAL CONDUCTION IN SOLIDS.

Source of S_d. The conduction electrons in a metal are those having energies near the Fermi energy η. Only these electrons are important for thermoelectricity. As illustrated in **Fig. 7**, in a metal, the energy distribution of these electrons varies with the temperature. At high temperatures, the metal has more high-energy electrons and fewer low-energy ones than when it is at low temperatures. This means that if a temperature gradient is established along a metal sample, the total number of electrons will remain constant, but the hot end will have more high-energy electrons than the cold end, and the cold end will have more low-energy electrons than the hot end. The high-energy electrons will diffuse toward the cold end, and the low-energy electrons will diffuse toward the hot end. However, in general, the diffusion rate is a function of electron energy, and thus a net electron current will result. This current will cause electrons to pile up at one end of the metal (usually the cold end) and thereby produce an emf that opposes the further flow of electrons. When the emf becomes large enough, the current will be reduced to zero. This equilibrium emf is the thermoelectric emf arising from electron diffusion. Essentially the same argument applies to semiconductors, except that in that case the electrons (or holes) are those just above (or just below) the band gap. *See* FREE-ELECTRON THEORY OF METALS; HOLE STATES IN SOLIDS; SEMICONDUCTOR.

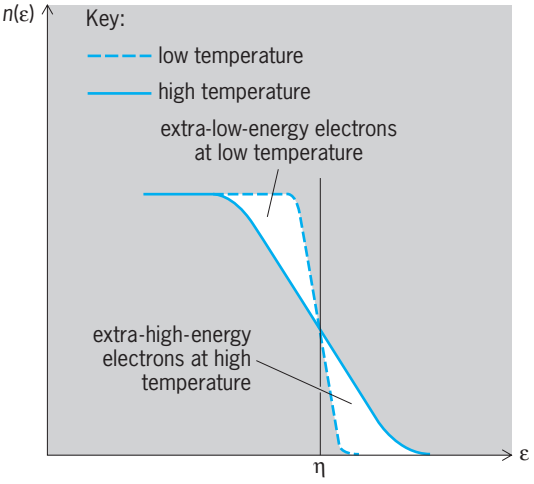


Fig. 7. Variation with energy ϵ of the number of conduction electrons $n(\varepsilon)$ in a metal in the vicinity of the Fermi energy η for two different temperatures. A small variation of η with temperature has been neglected.

TABLE 2. Comparision between theoretical values of *S* and experimental data

Metal	Thermopower (S), μV/K: Theoretical values at 0°C according to Eq. (13)	Thermopower (S), μV/K: Experimental data at approximately 0°C
Lithium (Li)	−2	+11
Sodium (Na)	−3	−6
Potassium (K)	−5	−12
Copper (Cu)	−1.5	+1.4
Gold (Au)	−2	+1.7
Aluminum (Al)	−0.7	−1.7

S_g for a metal. For a completely free-electron metal, S_d should be given by Eq. (13), where k is Boltzmann's constant, e is the charge on an electron, T is the absolute (Kelvin) temperature, and η is the Fermi energy of the metal. According to Eq. (13), S_d should be negative—since e is a negative quantity—and should increase linearly with T. In **Table 2** the predictions of Eq. (13) are compared with experiment for a number of the most free-electron-like metals. Equation (13) correctly predicts the general size of S_d, but often misses the actual value by a factor of 2 or more and in several cases predicts the wrong sign. To understand the thermopowers of real metals, it is necessary to use a more sophisticated model that takes into account interactions between the electrons in the metal and the crystal lattice, as well as scattering of the electrons by impurities and phonons. The proper generalization of Eq. (13) is Eq. (14), where

$$S_d = \frac{\pi^2 k}{2e}\left(\frac{kT}{\eta}\right) \tag{13}$$

$$S_d = \frac{\pi^2 k^2 T}{3e}\left[\frac{\delta \ln \sigma(\varepsilon)}{\delta \varepsilon}\right]_{\varepsilon=\eta} \tag{14}$$

$\sigma(\varepsilon)$ is a generalized energy-dependent conductivity defined so that $\sigma(\eta)$ is the experimental electrical conductivity of the metal, and the logarithmic derivative with respect to the energy ε is to be evaluated at $\varepsilon = \eta$. For free electrons, Eq. (14) reduces to Eq. (13). But more generally, Eq. (14) is able to account for all of the deviations of experiment from Eq. (13). If the logarithmic derivative is negative, S_d will be positive; S_d will differ in size from Eq. (13) if the logarithmic derivative does not have the free-electron value $(3/2)\eta^{-1}$; and $S_d(T)$ will deviate from a linear dependence on T if the logarithmic derivative is temperature-dependent.

In metals, research on S_d has attacked such diverse topics as understanding changes due to alloying with both magnetic and nonmagnetic impurities,

strain, application of pressure or magnetic fields, and changes near phase transitions. In some cases the changes can be dramatic. **Figure 8** shows that the addition of very small amounts of the magnetic impurity iron (Fe) can produce enormous changes in S_d for copper (Cu) at low temperatures. Sample 1 (in which the deviation of the thermopower from zero is too small to be seen with the chosen scale) is most representative of pure Cu because the Fe is present as an oxide and is thus not in "magnetic form." **Figure 9** shows that at low temperatures application of a magnetic field H to Al can cause S_d to change sign. [To obtain a temperature-independent quantity, S_d has been divided by the absolute temperature T. To remove the effects of varying impurity concentrations, H has been divided by $\rho(4.2\ \text{K})nec$, where $\rho(4.2\ \text{K})$ is the sample resistivity at 4.2 K, n is the number of electrons per unit volume in the metal, and c is the speed of light.] **Figure 10** illustrates the significant changes that can occur in S when a metal melts.

Substantial effort has been devoted to the study of thermoelectricity in liquid metals and liquid metal alloys. There has also been considerably interest in the thermopower of quasi-one-dimensional conductors (**Fig. 11**), in amorphous metals (also called metallic glasses or metglasses), in many-body effects in thermoelectricity, and in thermoelectric effects in inhomogeneous superconductors, such as loops consisting of two different superconducting materials. Thermoelectric effects in superconductors are much smaller than those in normal metals and are generally visible only very near T_c. Their study gives insight into nonequilibrium processes in superconductors. *See* METALLIC GLASSES.

S_d for a semiconductor. Equation (13) is appropriate for a free-electron gas that obeys Fermi-Dirac statistics. The conduction electrons in a metal obey these statistics. However, there are so few conduction electrons in a semiconductor that, to a good approximation, they can be treated as though they obey different statistics—Maxwell-Boltzmann statistics. For free electrons obeying these statistics, S_d is given by Eq. (15), which predicts that S_d should be

$$S_d = \frac{3k}{2e} \tag{15}$$

temperature-independent and have the value $S_d = -130$ times; 10^{-6} V/K. For a p-type extrinsic semiconductor, in which the carriers are approximated as free holes, the sign of S_d will be reversed to positive. Examination of the data of Fig. 6 shows that S_d is nearly independent of temperature but is considerably larger than predicted by Eq. (15). Again, a complete understanding of the thermopowers of semiconductors requires the generalization of Eq. (15). The appropriate generalizations are different for single-band and multiband semiconductors, the latter being considerably more complicated. For a single-band (extrinsic) semiconductor, the generalization is relatively straightforward and yields predictions for S_d which, in agreement with experiment,

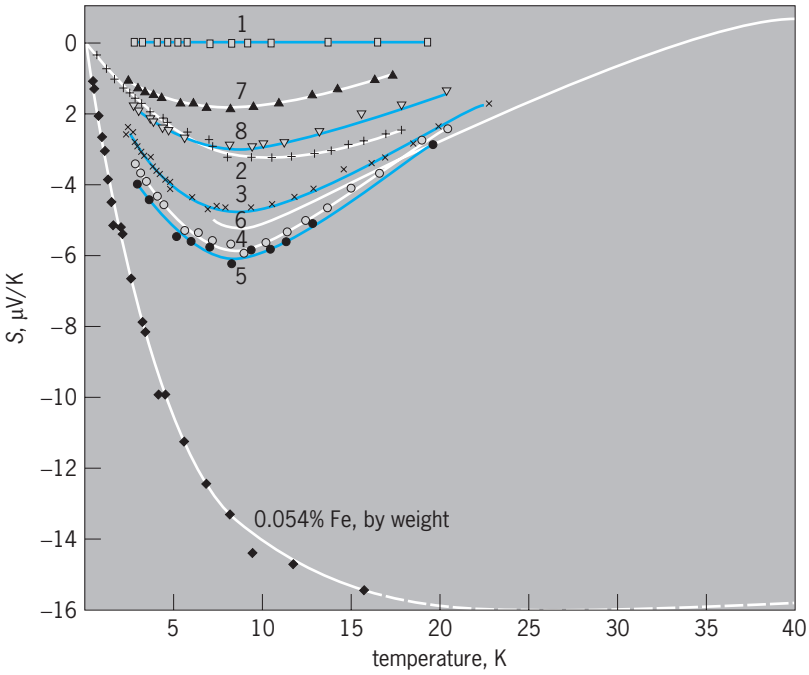


Fig. 8. Low-temperature thermopowers of various samples of copper containing very small concentrations of iron (Fe). Specific compositions of samples 1–8 are unknown. °F = (K × 1.8) − 459.67. (*After A. V. Gold et al., The thermoelectric power of pure copper, Phil. Mag., 5:765–786, 1960*)

vary slowly with temperature and are several times larger than the prediction of Eq. (15). [The white curve for S_d in Fig. 6 is calculated from this generalization.] Experimental interest in the thermopower of semiconductors concerns topics similar to those for metals. In addition, the large magnitudes of the thermopowers of semiconductors continue to spur

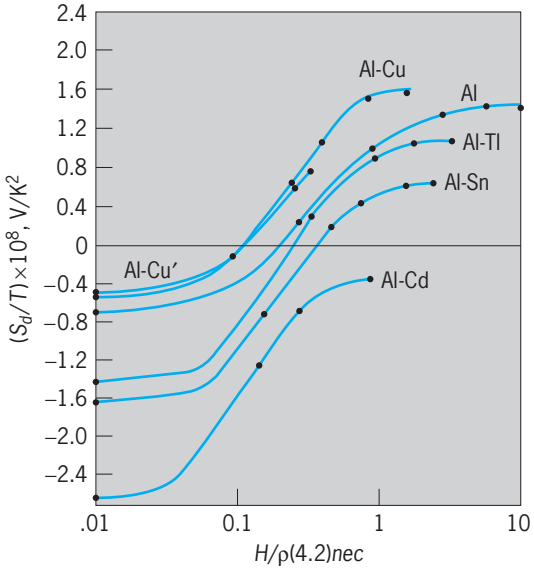


Fig. 9. Variation with magnetic field H of the low-temperature electron-diffusion thermopower S_d of aluminum (Al) and various dilute aluminum-based alloys. Sample labeled Al-Cu is a second sample of Al-Cu. (*After R. S. Averback, C. H. Stephan, and J. Bass, Magnetic field dependence of the thermopower of dilute aluminum alloys, J. Low Temp. Phys., 12:319–346, 1973*)

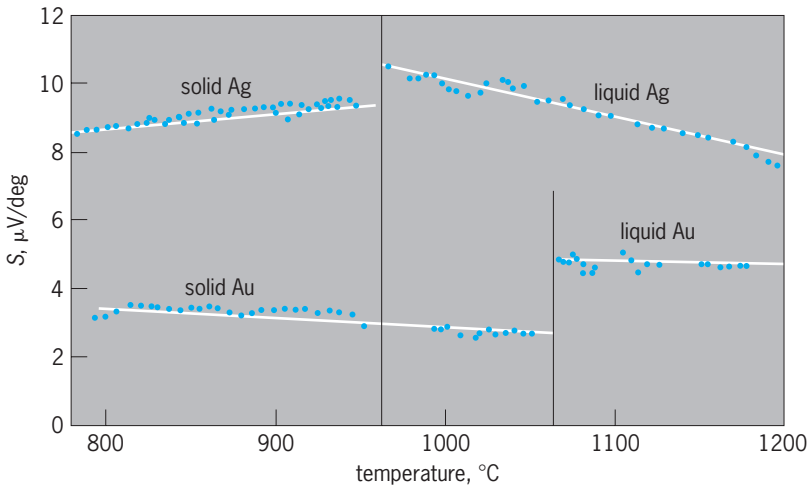


Fig. 10. Changes in the thermopowers of gold (Au) and silver (Ag) upon melting. °F = (°C × 1.8) + 32. (*After R. A. Howe and J. E. Enderby, The thermoelectric power of liquid Ag-Au, Phil. Mag., 16:467–476, 1967*)

efforts to develop materials better suited for electric power generation and thermoelectric cooling. *See* BAND THEORY OF SOLIDS; BOLTZMANN STATISTICS; FERMI-DIRAC STATISTICS.

Source of S_g. Unlike the behavior of S_d, which is determined in both metals and semiconductors primarily by the properties of the charge carriers, the behavior of S_g is determined in both cases primarily by the properties of the phonons. At low temperatures, phonons scatter mainly from electrons or impurities rather than from other phonons. The initial increase in S_g with increasing temperature at the very lowest temperatures in Figs. 5 and 6 results from an increasing number of phonons becoming available to drag the electrons along. However, as the temperature increases, the phonons begin to scatter more frequently from each other. Eventually, phonon-phonon scattering becomes dominant, the electrons are no longer dragged along, and S_g falls off in magnitude with increasing temperature. Interest in phonon drag is associated with such questions as whether it is the sole source of the humps in Figs. 5 and 6, how it changes as impurities are added, and how it is affected by a magnetic field.

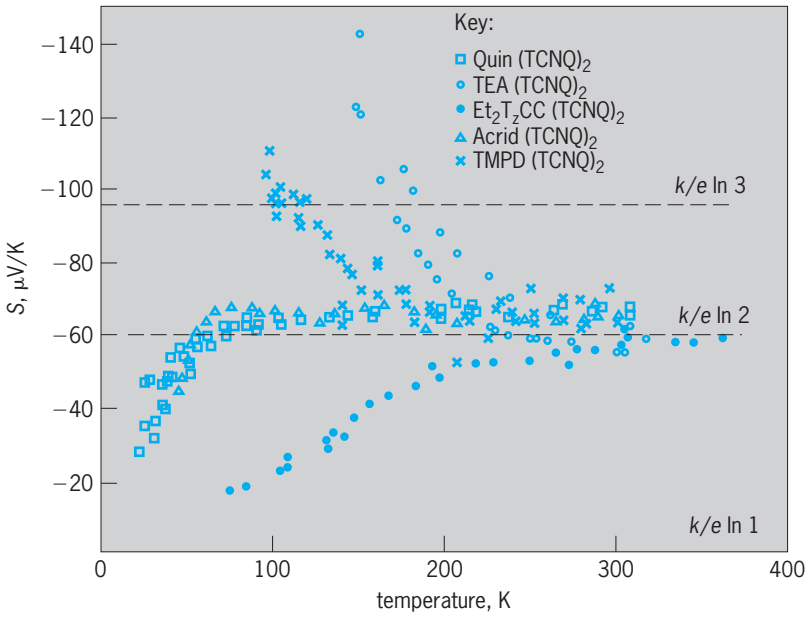


Fig. 11. Thermopower of highly conducting salts of the form (Donor)$^+$ $(TCNQ)_2^-$. By 300 K (80°F) all of the thermopowers are very close to the "entropy per carrier" of (k/e) ln 2 = −60 μV/K, where k is Boltzmann's constant and e is the electron charge. °F = (°K × 1.8) − 459.67. (*After F. J. Blatt and P. A. Schroeder, eds., Thermoelectricity in Metallic Conductors, Plenum Press, 1978*)

Applications

The most important practical application of thermoelectric phenomena to date is in the accurate measurement of temperature. The phenomenon involved is the Seebeck effect. Of less importance so far are the direct generation of electrical power by application of heat (also involving the Seebeck effect) and thermoelectric cooling and heating (involving the Peltier effect).

A basic system suitable for all four applications is illustrated schematically in **Fig. 12**. Several thermocouples are connected together in series to form a thermopile, a device with increased output (for power generation or cooling and heating) or sensitivity (for temperature measurement) relative to a single thermocouple. The junctions forming one end of the thermopile are all at the same low temperature T_L, and the junctions forming the other end are at the high temperature T_H. The thermopile is connected to a device D that differs for each application. For temperature measurement, the temperature T_L is fixed, for example, by means of a bath; the temperature T_H becomes the running temperature T that is to be measured; and the device is a potentiometer for measuring the thermoelectric emf generated by the thermopile. For power generation, the temperature T_L is fixed by connection to a heat sink; the temperature T_H is fixed at a value determined by the output of the heat source and the thermal conductance of the thermopile; and the device is whatever is to be run by the electricity that is generated. For heating or cooling, the device is a current generator that passes current through the thermopile. If the current flows in the proper direction, the junctions at T_H will heat up and those at T_L will cool down. If T_H is fixed by connection to a heat sink, thermoelectric cooling will be provided at T_L. Alternatively, if T_L is fixed, thermoelectric heating will be provided at T_H. If the heat sink is room-temperature, such a system has the advantage that at any given location it can be converted from a cooler to a heater merely by reversing the direction of the current.

Temperature measurement. In principle, any material property that varies with temperature can serve as the basis for a thermometer. In practice, the two properties most often used for precision thermometry are electrical resistance and thermoelectric emf. Thermocouples are widely employed to measure temperature in both scientific research and industrial processes. In the United States alone,

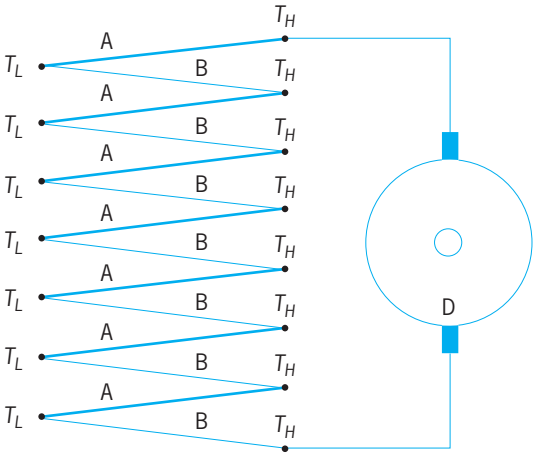


Fig. 12. Thermopile, a battery of thermocouples connected in series; *D* is a device appropriate to the particular application.

several hundred tons of thermocouple materials are produced annually. *See* TEMPERATURE MEASUREMENT; THERMOCOUPLE.

Construction of instruments. In spite of their smaller thermopowers, metals are usually preferred to semiconductors for precision temperature measurements because they are cheaper, are easier to fabricate into convenient forms such as thin wires, and have more reproducible thermoelectric properties. With modern potentiometric systems, standard metallic thermocouples provide temperature sensitivity adequate for most needs—small fractions of a degree Celsius are routinely obtained. If greater sensitivity is required, several thermocouples can be connected in series to form a thermopile (Fig. 12). A 10-element thermopile provides a temperature sensitivity 10 times as great as that of each of its constituent thermocouples. However, the effects of any inhomogeneities are also enhanced 10 times.

The thermocouple system standardly used to measure temperature is shown in **Fig. 13**. It consists of wires of three metals, A, B, and C, where C is usually the metal Cu. The junction between the wires of metals A and B is located at the temperature to be measured T. Each of these two wires is joined to a wire of metal C at the reference temperature T_0. The other ends of the two wires of metal C are connected to the potentiometer at room temperature T_r. Integrating the appropriate thermopowers around the circuit of Fig. 13 yields the total thermoelectric emf E in terms of the separate emf's generated by each of the four pieces of wire, as given in Eq. (16).

$$E = E_A(T_0, T) - E_B(T_0, T) + E_{C1}(T_0, T_r) - E_{C2}(T_0, T_r) \qquad (16)$$

If the two wires C1 and C2 have identical thermoelectric characteristics, the last two terms in Eq. (16) cancel, and, with the use of Eq. (5), Eq. (17) results.

$$E = E_A(T_0, T) - E_B(T_0, T) = E_{AB}(T_0, T) \qquad (17)$$

That is, two matched pieces of metal C produce no contribution to the thermoelectric emf of the circuit shown in Fig. 13, provided their ends are maintained at exactly the same two temperatures. This means that it is not necessary to use either of the sometimes expensive metals making up the thermocouple to go from the reference-temperature bath to the potentiometer. That part of the circuit can be constructed of any uniform, homogeneous metal. Copper is often used because it is inexpensive, is available in adequate purity to ensure uniform, homogeneous samples when handled with care, can be obtained in a wide variety of wire diameters, and can be either spotwelded or soldered to the ends of the thermocouple wires. Special low-thermal emf alloys are available for making solder connections to Cu in thermocouple circuits.

Choice of materials. Characteristics that make a thermocouple suitable as a general-purpose thermometer include adequate sensitivity over a wide temperature range, stability against physical and chemical change under various conditions of use and over extended periods of time, availability in a selection of wire diameters, and moderate cost. No single pair of thermocouple materials satisfies all needs. Platinum versus platinum–10% rhodium can be used up to 1700°C (3100°F). A thermocouple combining the two alloys chromel and alumel gives greater sensitivity and an emf that is closely linear with temperature, but cannot be used to as high a temperature. A combination of Cu versus the alloy constantan also has high sensitivity above room temperature and adequate sensitivity down to as low as 15 K (−433°F). For temperatures of 4 K (−452°F) or lower, special gold-cobalt alloys versus Cu or gold-iron alloys versus chromel are used.

Thermocouple tables. To use a thermocouple composed of metals A and B as a thermometer, it is necessary to know how $E_{AB}(T_0,T)$ varies with temperature T for some reference temperature T_0. According to Eq. (6), $E_{AB}(T_0,T_1)$ can be determined for

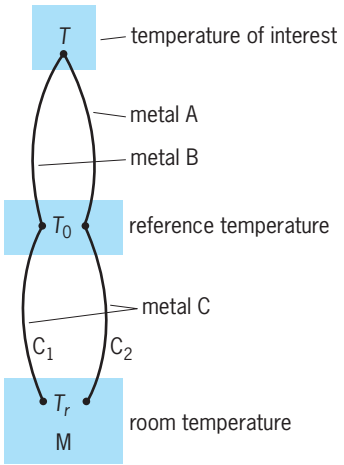


Fig. 13. Thermocouple system standardly used to measure temperature; M is a measuring device, usually a potentiometer, which is at room temperature.

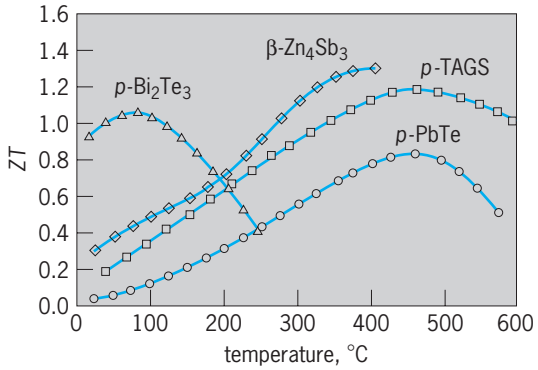


Fig. 14. Representative values of the dimensionless figure of merit *ZT* as a function of temperature for *p*-type β-Zn_4Sb_3 (diamonds). These values are to be compared with those for state-of-the-art *p*-type thermoelectric materials: PbTe- and Bi_2Te_3-based alloys and TAGS (Te-Ag-Ge-Sb) alloys. °F = (°C × 1.8) + 32. (*From T. Caillat, J.-P. Fleurial, and A. Borschevsky, A low thermal conductivity compound for thermoelectric applications: β-Zn_4Sb_3, 15th IEEE International Conference on Thermoelectrics, p. 151, 1996*)

any two temperatures T_0 and T_1 if both $S_A(T)$ and $S_B(T)$ are known for all temperatures between T_0 and T_1. Knowledge of $S_A(T)$ and $S_B(T)$ allows construction of a table of values for $E_{AB}(T_0,T)$ using any arbitrary reference temperature T_0. Such tables are available for the thermocouples mentioned above, and for some others as well, usually with a reference temperature of 0°C (32°F). A table of $E_{AB}(T_0,T)$ for one reference temperature T_0 can be converted into a table for any other reference temperature T_2 merely by subtracting a constant value $E_{AB}(T_0,T_2)$ from each entry in the table to give Eq. (18). Here $E_{AB}(T_0,T_2)$

$$E_{AB}(T_2, T) = E_{AB}(T_0, T) - E_{AB}(T_0, T_2) \qquad (18)$$

is a positive quantity when T_2 is greater than T_0 and when $S_{AB}(T)$ is positive between T_0 and T_2.

Other uses. Thermoelectric power generators, heaters, or coolers made from even the best presently available materials have the disadvantages of relatively low efficiencies and concomitant high cost per unit of output. Their use has therefore been largely restricted to situations in which these disadvantages are outweighed by such advantages as small size, low maintenance due to lack of moving parts, quiet and vibration-free performance, light weight, and long life.

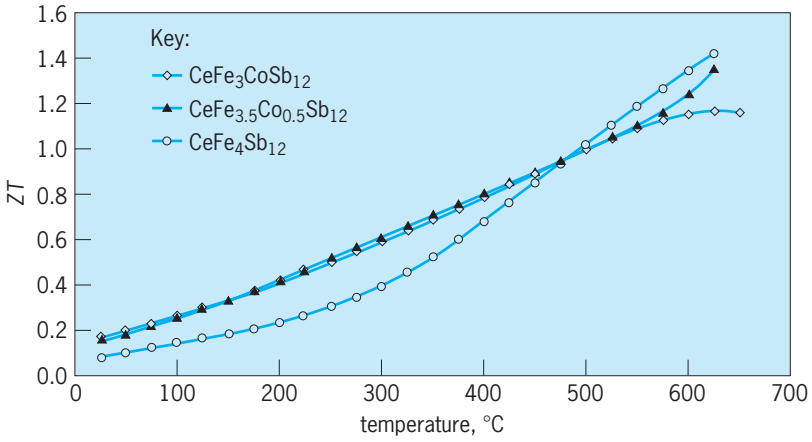


Fig. 15. Dimensionless figure of merit *ZT* as a function of temperature for cerium (Ce)-filled skutterudite samples with different compositions. °F = (°C × 1.8) + 32. (*From J.-P. Fleurial et al., High figure of merit in Ce-filled skutterudites, 15th IEEE International Conference on Thermoelectrics, p. 91, 1996*)

Figure of merit. A measure of the utility of a given thermoelectric material for power generation, cooling, or heating at an absolute temperature T is provided by the dimensionless figure of merit ZT given by Eq. (19). Here S is the thermopower of the material, σ is its electrical conductivity, and κ is its thermal conductivity. The largest values for ZT are obtained in semimetals and highly doped semiconductors, which are the materials normally used in practical thermoelectric devices. As illustrated in **Fig. 14**, for most materials ZT varies with temperature, going through maxima at different temperatures. Thus, combining available materials into thermocouples often results in values of ZT too small to be competitive over a wide enough temperature range to be useful. The best available values of $ZT \approx 1$ yield conversion efficiencies of only a few percent. Values of $ZT > 2$ over a wide enough temperature range could make thermoelectrics competitive for some uses described below, and values of $ZT \geq 4$ over wide temperature ranges in materials with high-temperature stability and affordable cost might revolutionize heating, cooling, and power generation. *See* THERMOELECTRIC POWER GENERATOR.

$$ZT = \frac{S^2 \sigma T}{\kappa} \qquad (19)$$

For a long time, little progress was made in increasing ZT beyond the values for the established BiTe- and PbTe-based thermoelectric materials shown in Fig. 14, although no rigorous theoretical limit on the value of ZT is known. However, research on thermoelectric materials has accelerated due to the discovery of values of $ZT \geq 1$ in β-Zn_4Sb_3 (Fig. 14) and ternary filled skutterudites (**Fig. 15**) of the form LnT_4Pn_{12} (Ln = rare earth or Th; T = Fe, Ru, Os, Co, Rh, Ir; Pn = P, As, Sb), as well as new ideas that might lead to larger ZT.

The quantity of importance in a thermoelectric device is the figure of merit of the thermocouple rather than the separate figures of merit of its constituents. Although at least one constituent should have a high figure of merit, two constituents with high figures of merit do not guarantee that the figure of merit of the thermocouple will be high. For example, if the thermopowers of the two constituents are identical, the figure of merit of the couple will be zero.

Just as the figures of merit for single materials vary with temperature, so do the figures of merit for thermocouples formed from two such materials. This means that one thermocouple can be better than another in one temperature range but less efficient in another. To take maximum advantage of the different properties of different couples, thermocouples are often cascaded (**Fig. 16**). Cascading produces power generation in stages, the higher temperature

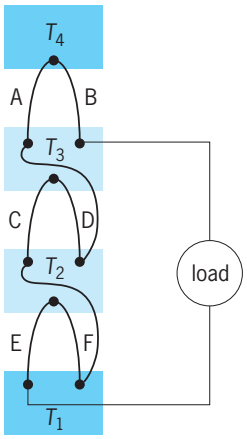


Fig. 16. Three-level cascade consisting of three different thermocouples (A versus B, C versus D, and E versus F) at four temperatures (*T*).

of each stage being determined by the heat rejected from the stage above. Thus, the highest and lowest temperatures T_4 and T_1 are fixed by connection to external reservoirs, whereas the middle temperatures T_3 and T_2 are determined by the properties of the materials. By cascading, a series of thermocouples can be used simultaneously in the temperature ranges where their figures of merit are highest. Cascaded thermocouple systems have achieved conversion efficiencies as high as 10–15%.

Thermoelectric generators. A thermoelectric generator requires a heat source and a thermocouple. Kerosine lamps and firewood have been used as heat sources to produce a few watts of electricity in remote locations, and systems using motor fuel burners have produced 100 W. Test systems using sunlight have also been constructed. Radioactive sources, especially strontium-90 (^{90}Sr) and plutonium-238 (^{238}Pu), have provided the heat to activate small rugged thermoelectric batteries for use in lighthouses, in navigation buoys, at isolated weather stations or oil platforms, and in spaceships. Such small nuclear batteries operate pacemakers implanted in humans and data transmission systems for deep-space planetary probes. Thermopiles based on silicon-germanium (SiGe) alloys and powered by plutonium-238 supplied more than 100 W of power at an efficiency of over 6% to the *Voyager 1* and *Voyager 2* spacecraft for the 12 years (1977–1989) of their missions to the outer planets, and are expected to remain electrically active until at least the year 2015. Nuclear-powered batteries for medical use must be designed to remain intact following the maximum credible accident. Capabilities such as retention of integrity after crushing by 1 ton, or impact at 50 m/s (112 mi/h), or salt-water corrosion for centuries, or cremation at temperatures up to 1300°C (2400°F) for half an hour are required. Investigations have been made of the feasibility of thermoelectric generation using the copious heat generated by nuclear reactors, or the heat generated in the exhaust system of automobiles. Such uses would require the development of more efficient thermoelectric materials able to operate for a long time at the high temperatures that are encountered. *See* NUCLEAR BATTERY; NUCLEAR POWER; RADIOACTIVITY; SPACE POWER SYSTEMS.

Peltier cooling. With available values of *ZT*, thermoelectric refrigerators suitable for use in homes or automobiles are more expensive and less efficient than standard vapor-compression-cycle refrigerators. Their use is thus largely restricted to situations in which lower maintenance, increased life, absence of vibration, or quiet performance are essential, or to situations (as in space vehicles or artificial satellites) in which the compressor type of refrigerator is impractical. A number are in use in hotels and other large facilities. A typical unit of about 50-liter (13-gallon) capacity requires a dc power input of 40 W, has a refrigerative capacity of 20 kcal/h (23 W), and a time to cool of 4–5 h. Larger *ZT* would also make thermoelectricity competitive for cooling of high-power electronic devices. *See* REFRIGERATION.

For lower temperatures, the proper choice of thermoelectric materials and the use of cascading can result in a reduction in temperature at the coldest junctions of as much as 150°C (270°F). Temperature drops of 100°C (180°F) have been obtained in single crystals of the semimetal bismuth through use of the thermomagnetic Ettingshausen effect. A large enough *ZT* down to temperatures of −200°C (−328°F) could allow widespread use of new high-temperature superconductors in electronic devices. *See* THERMOMAGNETIC EFFECTS.

Small cooling units with capacities of 10 W or less have been developed for miscellaneous applications such as cold traps for vacuum systems, cooling controls for thermocouple reference junctions, cooling devices for scientific equipment such as infrared detectors, cold stages on microscopes or on microtomes used for sectioning cooled tissues, and electronic equipment coolers. However, major commercial success of thermoelectric refrigeration requires thermocouple materials with higher values of *ZT*.

Thermoelectric heating. A thermoelectric heater referenced to room temperature is nothing more than a thermoelectric refrigerator with the current reversed. No large heaters have been marketed. However, various small household convenience devices have been developed, such as a baby-bottle cooler-warmer that cools the bottle until just before feeding time and then automatically switches to a heating cycle to warm it, and a thermoelectric hostess cart. *See* ELECTRICITY.

Jack Bass

Bibliography. F. J. Blatt et al., *Thermoelectric Power of Metals*, Plenum Press, New York, 1976; F. J. Blatt and P. A. Schroeder (eds.), *Thermoelectricity in Metallic Conductors*, Plenum Press, New York, 1978; F. J. DiSalvo, Thermoelectric cooling and power generation, *Science*, 285:703–706, 1999; D. M. Rowe (ed.), *CRC Handbook of Thermoelectrics*, CRC Press, Boca Raton, FL, 1995; *Proceedings of the 15th IEEE International Conference on Thermoelectrics*, IEEE Catalog no. 96TH8169, 1996; T. M. Tritt et al. (eds.), Thermoelectric Materials 1998—The Next

Generation Materials for Small-Scale Refrigeration and Power Generation Applications, *MRS Soc. Proc.*, vol. 545, 1998.

Thermoluminescence

The emission of light when certain solids are warmed, generally to a temperature lower than that needed to provoke visible incandescence. Two characteristics of thermoluminescence distinguish it from incandescence. First, the intensity of thermoluminescent emission does not remain constant at constant temperature, but decreases with time and eventually ceases altogether. Second, the spectrum of the thermoluminescence is highly dependent on the composition of the material and is only slightly affected by the temperature of heating. If a thermoluminescent material emits both thermoluminescence and incandescent light at some temperature of observation, the transient light emission is the thermoluminescence and the remaining steady-state emission is the incandescence. The transient nature of the thermoluminescent emission suggests that heating merely triggers the release of stored energy previously imparted to the material. Supporting this interpretation is the fact that after the thermoluminescence has been reduced to zero by heating, the sample can be made thermoluminescent again by exposure to one of a number of energy sources: x-rays and gamma rays, electron beams, nuclear particles, ultraviolet light, and, in some cases, even shortwave visible light (violet and blue). A thermoluminescent material, therefore, has a memory of its earlier exposure to an energizing source, and this memory is utilized in a number of applications. Many natural minerals are thermoluminescent, but the most efficient materials of this type are specially formulated synthetic solids (phosphors). *See* LUMINESCENCE.

Mechanism. In addition to special sites capable of emitting light (luminescent centers), thermoluminescent phosphors have centers that can trap electrons or holes when these are produced in the solid by ionizing radiation. The luminescent center itself is often the hole trap, and the electron is trapped at another center, although the reverse situation can also occur. In the former case, if the temperature is low and the energy required to release an electron from a trap (the trap depth) is large, electrons will remain trapped and no luminescence will occur. If, however, the temperature of the phosphor is progressively raised, electrons will receive increasing amounts of thermal energy and will have an increased probability of escape from the traps. Freed electrons may then go over to luminescent centers and recombine with holes trapped at or near these centers. The energy liberated by the recombination can excite the luminescent centers, causing them to emit light. *See* HOLE STATES IN SOLIDS; TRAPS IN SOLIDS.

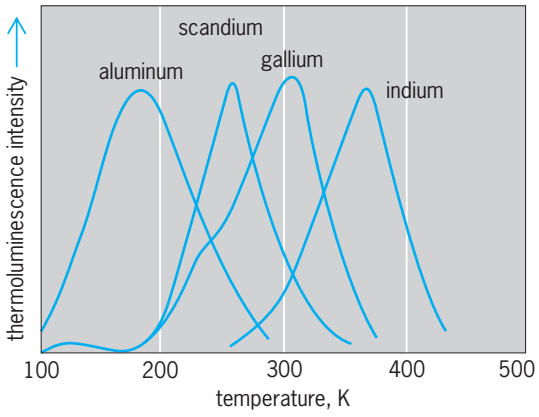
Glow curves. A plot of luminescence intensity versus temperature (or time, if a constant heating rate is employed) is called a glow curve. The initial rise of intensity with temperature is due to the increased number of electrons escaping from traps as the temperature is raised. When all traps are of the same depth, the intensity peaks at some temperature due to the decrease in population of the trapped electrons, and it finally drops to zero when all the traps are emptied. The exact shape of the glow curve depends on the heating rate. The higher the heating rate, the brighter is the thermoluminescent glow, but the shorter its duration. The total number of light quanta emitted (the light sum) is generally independent of the heating rate, however.

When the glow curve consists of a single peak, corresponding to a single species of trap, the trap depth E (in joules or other energy units) is given to a good approximation by the equation below, where

$$E = \frac{1.51\, kT^*T'}{(T^* - T')}$$

k is Boltzmann's constant, T^* is the temperature (in kelvins) of the phosphor at the peak of the curve, and T' is the temperature on the low-temperature side at which the emission is one-half its peak value. The **illustration** shows single-peaked glow curves of this type for several zinc sulfide phosphors, all of which thermoluminesce green due to the presence of copper (the activator) in the luminescent center. Traps of different depths are produced by different trivalent impurities (coactivators). In thermoluminescent materials containing more than one type of trapping center, the glow curve comprises a corresponding number of peaks, which often may be resolved and analyzed as described above. Thermoluminescence may thus be used to obtain information about the properties of traps in solids.

Applications. Radiation dosimeters based on thermoluminescence are widely used for monitoring integrated radiation exposure in nuclear power plants, hospitals, and other installations where high-energy radiations are likely to be encountered. The key elements of the dosimeters, thermoluminescent phosphors with deep traps, can store some of the energy



Glow curves for the green luminescence of several zinc sulfide phosphors, each of which contains traces of copper and different trivalent ions. Luminescent center is due to presence of copper, the activator in each case. Traps are put in by various trivalent coactivators as shown. °F = (K × 1.8) − 459.67.

absorbed from these radiations for very long periods of time at normal temperatures and release it as luminescence on demand when appropriately heated. The brightness (or light sum) of the luminescence is a measure of the original radiation dose. *See* DOSIMETER.

Thermoluminescence induced by ionizing radiation also offers a method for archeological and possibly geological dating. Archeological dating is based on the principle that the high-temperature firing during the preparation of a ceramic object empties all the electron and hole traps in the material. It is further assumed that during the burial of a potsherd it has been exposed to a constant rate of irradiation from its self-contained isotopes, radioactive minerals in the ground, and cosmic rays. The dose accrued by the archeological specimen is determined by measuring its glow curve and calibrating the observed thermoluminescent emission against a portion of the sample that has been given a known dose of appropriate radiation. The annual dose to the potsherd from its own constituents and the surrounding environment must be estimated by separate measurements of their radioactivities. From these data the specimen age can be derived. *See* DATING METHODS.

Although the basic principle is thus quite simple, thermoluminescent dating involves a number of complex considerations, in particular, the determination of the annual dose. A number of different procedures have been developed, however, to deal with the various complications.

Thermoluminescence has also been used to evaluate radiation doses received by survivors of atomic bomb radiation in Hiroshima and Nagasaki, Japan, decades after the events, by using bricks and tiles from buildings that had been exposed to the radiation.

James H. Schulman

Bibliography. M. J. Aitken, *Thermoluminescence Dating*, 1985; G. Lalos (ed.), *Calibration Handbook*: *Ionizing Radiation Measuring Instruments*, 1984; S. W. S. McKeever, *Thermoluminescence of Solids*, 1988; W. C. Roesch (ed.), *Final Report on U.S.-Japan Reassessment of Atomic Bomb Radiation Dosimetry in Hiroshima and Nagasaki*, vol. 1, 1986; D. R. Vij, *Thermoluminescent Materials*, 1993.

Thermomagnetic effects

Electrical or thermal phenomena occurring when a conductor or semiconductor which is carrying a thermal current (that is, is in a temperature gradient) is placed in a magnetic field. *See* SEMICONDUCTOR.

Let the temperature gradient be transverse to the magnetic field H_z, for example, along x. Then the following transverse-transverse effects are observed:

1. Ettingshausen-Nernst effect, an electric field along y, as in Eq. (1), where Q is known as the Ettingshausen-Nernst coefficient. This coefficient is related to the Ettingshausen coefficient P by Eq. (2),

$$E_y = Q\frac{\partial T}{\partial x}H_z \tag{1}$$

$$P = QT\sigma \tag{2}$$

where σ is the thermal conductivity in a transverse magnetic field. This relation was discovered by P. W. Bridgman; it has been shown to be an example of the Onsager reciprocity relations of irreversible thermodynamics. *See* CHEMICAL THERMODYNAMICS; GALVANOMAGNETIC EFFECTS; THERMOELECTRICITY.

2. Righi-Leduc effect, a temperature gradient along y, as in Eq. (3), where S is known as the Righi-Leduc coefficient.

$$\frac{\partial T}{\partial y} = S\frac{\partial T}{\partial x}H_z \tag{3}$$

Also, the following transverse-longitudinal effects are observed:

3. An electric potential change along x, amounting to a change of thermoelectric power.

4. A temperature gradient change along x, amounting to a change of thermal resistance.

Let the temperature gradient be along H. Then changes in thermoelectric power and in thermal conductivity are observed in the direction of H.

For related phenomena *see* HALL EFFECT; MAGNETORESISTANCE.

Elihu Abrahams; Frederic Keffer

Thermometer

An instrument that measures temperature. Although this broad definition includes all temperature-measuring devices, they are not all called thermometers. Other names have been generally adopted. For a discussion of two such devices *see* PYROMETER; THERMOCOUPLE. For a general discussion of temperature measurement *see* TEMPERATURE MEASUREMENT.

A variety of techniques and designs are used in instruments known as thermometers. Some of these depend on the expansion of a liquid or metal for the indicating means. Others employ the change in pressure of a gas to detect the temperature. Still others use the change in electrical resistance which occurs with temperature changes.

Liquid-in-glass thermometer. This thermometer consists of a liquid-filled glass bulb and a connecting partially filled capillary tube. When the temperature of the thermometer increases, the differential expansion between the glass and the liquid causes the liquid to rise in the capillary. In **Fig. 1***a* the graduations are etched on the glass stem. The thermometer in Fig. 1*b* has a separate graduated scale similar to that of the common household thermometer. A variety of liquids, such as mercury, alcohol, toluene, and pentane, and a number of different glasses are used in thermometer construction, so that various designs cover diverse ranges between about −300 and +1200°F (−185 and +650°C). The range covered by the ordinary mercury-in-glass thermometers is from about −38 to +500°F (−39 to +260°C), and with a special glass and an inert gas in the capillary tube the

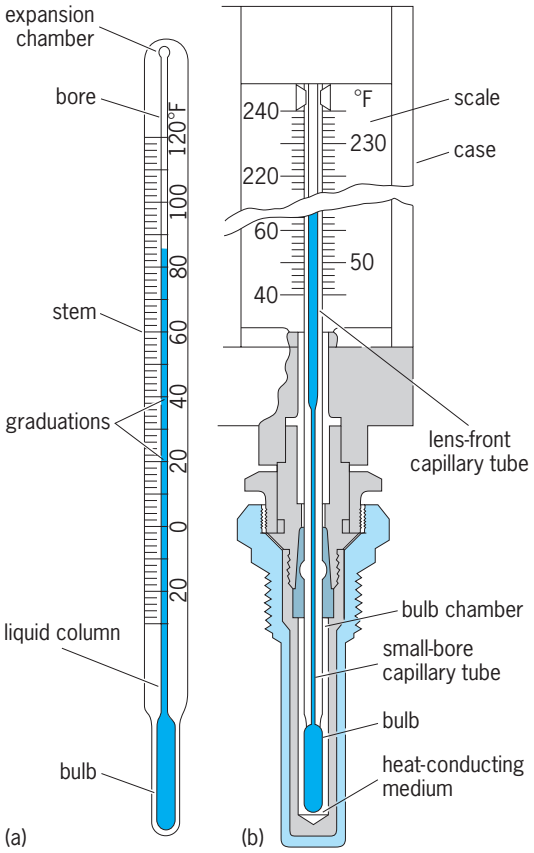


Fig. 1. Liquid-in-glass thermometers. (*a*) Etched-stem clinical thermometer. (*b*) Graduated-scale industrial thermometer. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

upper limit may be raised to about 1100°F (600°C). The usual ranges covered with the other liquids are: ethyl alcohol, −110 to +160°F (−79 to +71°C); and pentane, −330 to +85°F (−201 to +29°C).

Expansion and contraction chambers are sometimes provided at each end of the capillary to permit overrange and underrange use of the thermometer without loss of accuracy. When the entire thermometer is not subjected to the same temperature, an error occurs unless the thermometer is calibrated for these conditions. Many thermometers have an emergent stem and are calibrated for this type of service.

Maximum-registration thermometers, such as a fever thermometer, allow expansion of the liquid with increasing temperature but maintain the liquid column as the temperature decreases, thereby indicating the maximum value attained. Minimum registering thermometers are also available. If the liquid filling is metallic, electric contacts can be mounted in the stem wall to complete a circuit when temperature reaches a specified value. These are used in alarm and control systems. Some thermometers destined for rugged service are enclosed in a metal sleeve or armor.

Accuracies and speeds of response vary widely with the designs, ranges, and installations. For example, the Beckmann thermometer can be easily read to an accuracy (stated in terms of maximum error) of 0.002°F (0.001°C) and the fever thermometer to 0.1°F (0.05°C). Industrial thermometers and armored thermometers are seldom more accurate than 1°F (0.5°C) as they are used. Mercury and glass thermometers have time constants as low as 0.1 s in well-stirred water, but industrial thermometers and all thermometers installed in wells may have time constants as long as 1 min.

Bimetallic thermometer. In this thermometer the differential expansion of thin dissimilar metals, bonded together into a narrow strip and coiled into the shape of a helix or spiral, is used to actuate a pointer (**Fig. 2**). In some designs the pointer is replaced with low-voltage contacts to control, through relays, operations which depend upon temperature, such as furnace controls. The designs of cases range from those for light laboratory service to those for heavy industrial use. Range spans are seldom shorter than 50°F (28°C) or longer than 400°F (220°C), with a maximum upper temperature limit for continuous service of 800°F (425°C) and a minimum of −300°F (−185°C). The shorter range spans are used near room temperatures, and accuracies in the neighborhood of 1°F can be achieved. At high and low temperatures the accuracy is seldom better than 5°F. The time constants of these thermometers are greater but of the same order of magnitude as the liquid-in-glass thermometer.

Filled-system thermometer. This type of thermometer, shown schematically in **Fig. 3**, has a bourdon

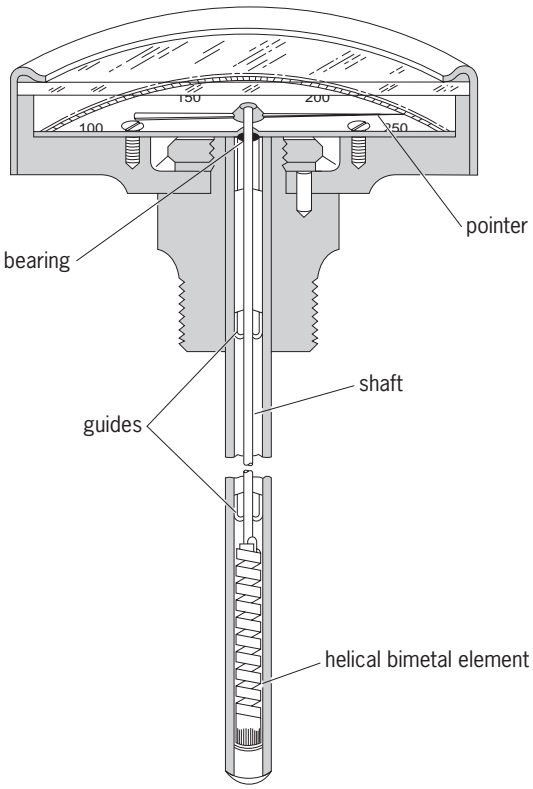


Fig. 2. Bimetallic thermometer. (*Weston Instruments, Division of Daystrom, Inc.*)

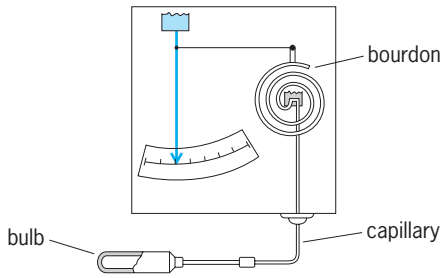


Fig. 3. Filled-system thermometer. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

tube connected by a capillary tube to a hollow bulb. When the system is designed for and filled with a gas (usually nitrogen or helium) the pressure in the system substantially follows the gas law, and a temperature indication is obtained from the bourdon tube. The temperature-pressure-motion relationship is nearly linear. Atmospheric pressure effects are minimized by filling the system to a high pressure. When the system is designed for and filled with a liquid, the volume change of the liquid actuates the bourdon tube. When mercury or its alloys are used as a filling medium, the temperature-volume-motion relationship is substantially linear. When hydrocarbon liquids are used, the liquid compressibility is appreciable, and the temperature-motion relationship is not so linear.

Since the fluids (liquid or gas) are homogeneous and extend to the bourdon tube, temperature changes on the capillary and on the bourdon tube will cause errors. These are made small by minimizing the volume in the capillary and bourdon tube and by providing ambient-temperature compensation. This compensation can be a duplicate system without a bulb to subtract the effect of the error; it can be a bimetallic compensator on the bourdon tube alone; or in the case of the mercury system a special capillary may be threaded with an invar wire compensator. The gas system has a relatively large bulb and a long range span (about 200°F minimum at room temperatures, 400°F near 1000°F; about 110°C minimum at room temperatures, 240°C near 540°C); the span may extend to a lower limit of about −400°F (−240°C) and an upper limit of about 1200°F (650°C). Hydrocarbon liquid systems have small bulbs and short range spans (as low as 25°F or 14°C); the span may extend to a lower limit of about −125°F (−87°C) and an upper limit of about 600°F (315°C). Mercury systems have somewhat larger bulbs (because of mercury's low temperature coefficient of expansion) and longer range spans and are used at temperatures between −40°F (−40°C) and 1200°F (650°C). Normally, accuracies of 1% of the range span are obtained from these instruments, but this is achieved only by proper selection with full knowledge of application conditions.

Vapor-pressure thermal system. This filled-system thermometer utilizes the vapor pressure of certain stable liquids to measure temperature, as shown in **Fig. 4**. The useful portion of any liquid-vapor pressure curve is between approximately 15 psia (100 kilopascals absolute) and the critical pressure, that is, the vapor pressure at the critical temperature, which is the highest temperature for a particular liquid-vapor system. Thus, when propane is used, the critical temperature is 206°F (97°C) and the vapor pressure is 617 psia (4.254 megapascals). A nonlinear relationship exists between the temperature and the vapor pressure, so the motion of the bourdon tube is greater at the upper end of the vapor-pressure curve. Therefore, these thermal systems are normally used near the upper end of their range, and an accuracy of 1% or better can be expected. Vapor-pressure systems are designed so that the active liquid-vapor interface occurs in the bulb, and the effective temperature occurs at this interface. There is no error due to ambient temperature changes on the capillary, and only the temperature effect on the metal bourdon tube produces an error at this point. The bourdon-tube error is normally small and may be compensated (bimetallic) if it must be reduced.

When the bulb and the bourdon tube are not at the same level a hydrostatic error occurs, but this is easily removed by zero setting. The effect of atmospheric pressure variations is minimized by using only the elevated portion of the vapor-pressure curve of the various liquids. Range spans vary widely, but near room temperature the useful portion of the span is about 120°F (67°C), and at elevated temperatures it is 200°F (110°C). Few vapor-pressure systems are used below 0°F (−18°C) and above 650°F (343°C).

The greatest advantage of the filled-system thermometer is its ability to provide a good, low-cost, temperature indication or record at a convenient point reasonably remote (up to 200 ft or 60 m) from the temperature being measured. The bourdon tube is powerful enough to operate sensitive detectors, the output of which can be amplified pneumatically, electrically, or hydraulically for control purposes. The particular characteristics of each class of thermal

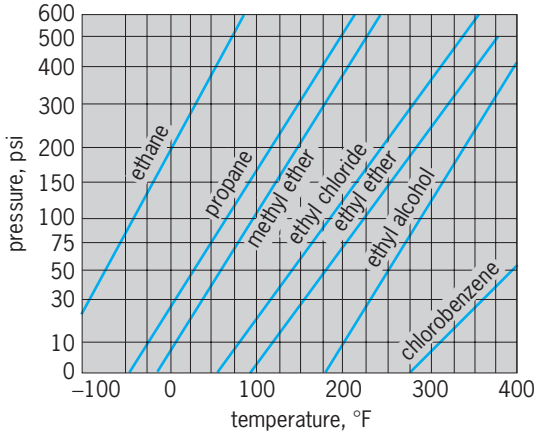


Fig. 4. Vapor-pressure versus temperature curves when various stable elements are used as thermal system fills. 1 psi = 6.895 kPa. °C = (°F − 32)/1.8. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

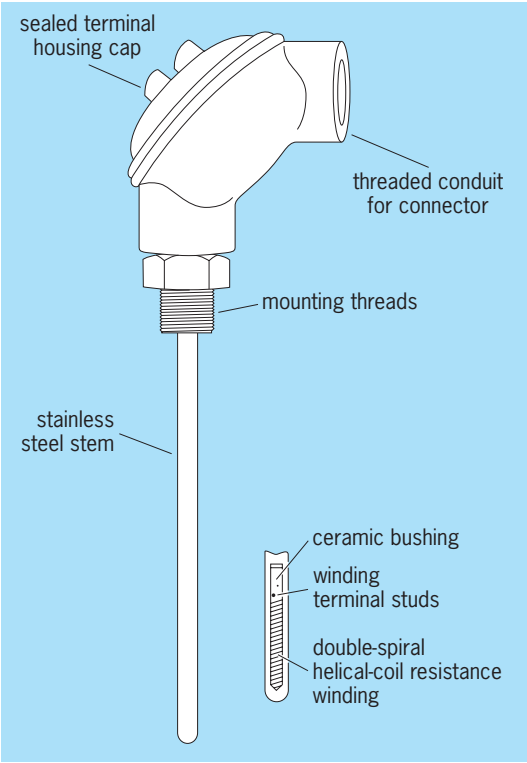


Fig. 5. Industrial-type resistance thermometer. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

system determine which will give the best service in various applications.

Resistance thermometer. In this type of thermometer the change in resistance of conductors or semiconductors with temperature change is used to measure temperature. Usually, the temperature-sensitive resistance element is incorporated in a bridge network which has a reasonably constant power supply.

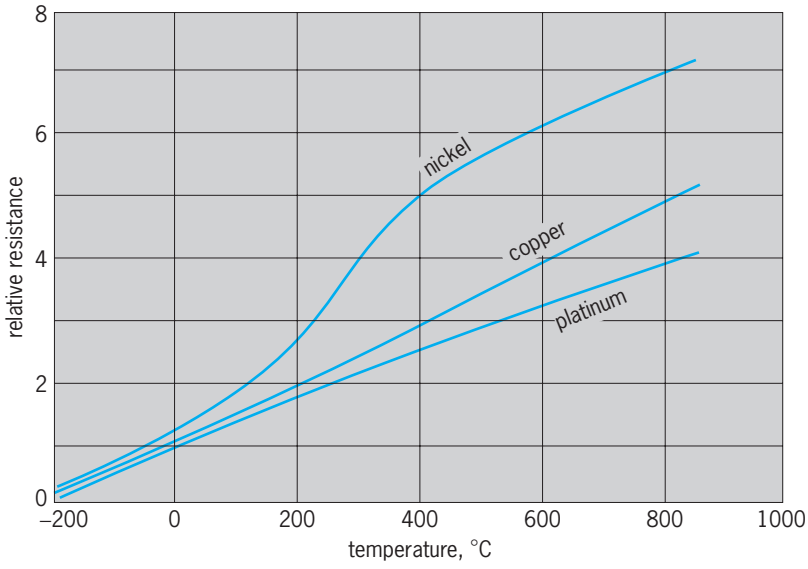


Fig. 6. Typical relative-resistance curves of several metals used in resistance thermometers. Relative resistance is the ratio of the resistance at the temperature of the metal to the resistance at 32°F (0°C). °F = (°C × 1.8) + 32. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

Although a deflection circuit is occasionally used, almost all instruments of this class use a null-balance system, in which the resistance change is balanced and measured by adjusting at least one other resistance in the bridge. All of the resistors in the bridge, except the measuring resistance, have low temperature coefficients, and the entire bridge circuit is designed to be insensitive to ambient temperature effects. The power supply to the resistance thermometer may be either direct or alternating current, the former preferred for precision measurements and the latter preferred when a servo system is used to rebalance the bridge. **Figure 5** shows an industrial resistance thermometer.

Metals commonly used as the sensitive element in resistance thermometers are platinum, nickel, and copper, and the change in resistance per degree Celsius is illustrated in **Fig. 6**. These are the most satisfactory metals, since they are stable, have a reasonable temperature coefficient of resistance, and can be drawn into fine homogeneous wires with a high resistance per unit length. For the resistance element, fine wire with a diameter of about 2.5 mils (64 micrometers) is usually wound into a small-diameter helix which is wound or otherwise mounted onto a mica support. Platinum wire elements can be used satisfactorily between −432 and 1650°F (−258 and 900°C), nickel between −238 and 572°F (−150 and 300°C), and copper between −328 and 248°F (−200 and 120°C).

A bare-wire element can be used in a clean, noncorrosive, nonconducting gas flowing at a low velocity, and its rate of response to temperature changes will be very rapid, not exceeding a few seconds. Usually, however, the element is mounted in a protecting well or tubing, from which it is electrically insulated, resulting in a time constant of a minute or more. The temperature-resistance relation of elements is determined by calibrations at the ice point, steam point, and sulfur point (832.28°F or 444.6°C). With calibrated industrial-grade resistance thermometers the uncertainty in temperature values may be about ±0.5°F (±0.3°C), while with a good, well-calibrated, and well-maintained laboratory instrument the uncertainty may be in the range of ±0.02°F (±0.01°C).

Since resistance thermometers carry a current, a self-heating error occurs. By keeping currents small and thermal conductivities high, this effect may be made negligible on most applications. In dc thermometry, thermal emfs must be carefully avoided in the circuitry. In ac thermometry, the circuitry must minimize inductive and capacitive disturbances.

Thermistor. This device is made of a solid semiconductor with a high temperature coefficient of resistance. The thermistor has a high resistance (**Fig. 7**), in comparison with metallic resistors, and is used as one element in a resistance bridge. Since thermistors are more sensitive to temperature changes than metallic resistors, accurate readings of small changes are possible. Thermistors are ceramic recrystallized mixtures of oxides of various metals and are usually in the form of small beads or disks with metallic leads.

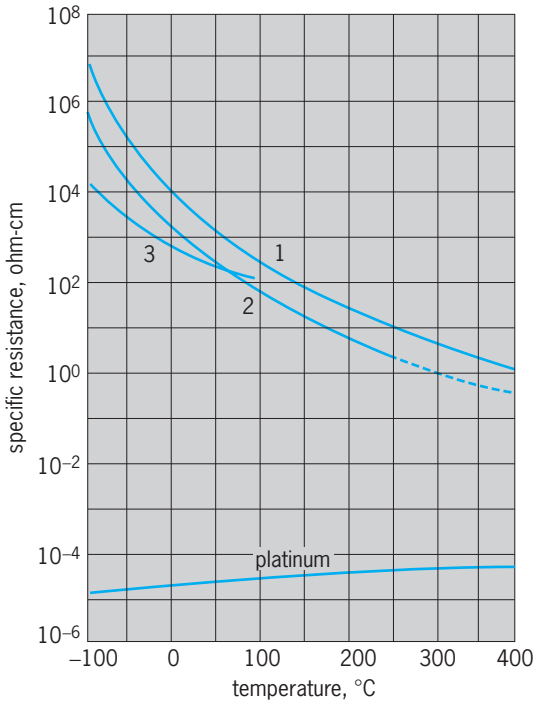


Fig. 7. Resistance-temperature properties characteristic of some typical thermometers. °F = (°C × 1.8) + 32. (*After D. M. Considine, ed., Process Instruments and Controls Handbook, McGraw-Hill, 1957*)

Thermistors are not as stable as metallic resistances, but certain compositions with good protection and care may change less than 1% per year. In general, thermistors are used between 212 and 750°F (100 and 400°C). They drift or deteriorate at higher temperatures, and at low temperatures their resistance tends to become excessive. *See* THERMISTOR.

Howard S. Bean

Bibliography. R. P. Benedict, *Fundamentals of Temperature, Pressure and Flow Measurements*, 3d ed., 1984; B. G. Liptak, *Temperature Measurement*, 1993; T. D. McGee, *Principles and Methods of Temperature Measurement*, 1988; G. K. McMillan and D. M. Considine (eds.), *Process/Industrial Instruments and Controls Handbook*, 5th ed., 1999; L. Michalski, K. Eckersdorf, and J. McGhee, *Temperature Measurement*, 2d ed., 2001; J. F. Schooley (ed.), *Temperature: Its Measurement and Control in Science and Industry*, vol. 5, 1982; J. F. Schooley, *Thermometry*, 1986.

Thermonuclear reaction

A nuclear fusion reaction which occurs between various nuclei of the light elements when they are constituents of a gas at very high temperatures. Thermonuclear reactions, the source of energy generation in the Sun and the stable stars, are utilized in the fusion bomb. *See* HYDROGEN BOMB; NUCLEAR FUSION; STELLAR EVOLUTION; SUN.

Thermonuclear reactions occur most readily between isotopes of hydrogen (deuterium and tritium) and less readily among a few other nuclei of higher atomic number. At the temperatures and densities required to produce an appreciable rate of thermonuclear reactions, all matter is completely ionized; that is, it exists only in the plasma state. Thermonuclear fusion reactions may then occur within such an ionized gas when the agitation energy of the stripped nuclei is sufficient to overcome their mutual electrostatic repulsions, allowing the colliding nuclei to approach each other closely enough to react. For this reason, reactions tend to occur much more readily between energy-rich nuclei of low atomic number (small charge) and particularly between those nuclei of the hot gas which have the greatest relative kinetic energy. This latter fact leads to the result that, at the lower fringe of temperatures where thermonuclear reactions may take place, the rate of reactions varies exceedingly rapidly with temperature. *See* PLASMA (PHYSICS).

The reaction rate may be calculated as follows: Consider a hot gas composed of a mixture of two energy-rich nuclei, for example, tritons and deuterons. The rate of reactions will be proportional to the rate of mutual collisions between the nuclei. This will in turn be proportional to the product of their individual particle densities. It will also be proportional to their mutual reaction cross section σ and relative velocity v. Thus Eq. (1) gives the rate of reaction. The quantity $(\sigma v)_{12}$ indicates an average value of σ and v obtained by integration of these quantities over the velocity distribution of the nuclei (usually assumed to be maxwellian). Since the total density $n = n_1 + n_2$, if the relative proportions of n_1 and n_2 are maintained, R_{12} varies as the square of the total particle density.

$$R_{12} = n_1 n_2 \langle \sigma v \rangle_{12} \qquad \text{reactions/(cm}^3\text{)(s)} \qquad (1)$$

The thermonuclear energy release per unit volume is proportional to the reaction rate and the energy release per reaction, as in Eq. (2). If this energy release, on the average, exceeds the energy losses from the system, the reaction can become self-perpetuating. *See* CARBON-NITROGEN-OXYGEN CYCLES; KINETIC THEORY OF MATTER; MAGNETOHYDRODYNAMICS; NUCLEAR REACTION; PINCH EFFECT; PROTON-PROTON CHAIN.

$$P_{12} = R_{12} W_{12} \qquad \text{ergs/(cm}^3\text{)(s)} \qquad (2)$$

Richard F. Post

Bibliography. R. O. Dendy, *Plasma Physics: An Introductory Course*, 2d ed., 1994; S. Glasstone and R. H. Lovberg, *Controlled Thermonuclear Reactions*, 1960, reprint 1975; K. Miyamoto, *Plasma Physics for Nuclear Fusion*, rev. ed., 1989; K. Nishikawa and M. Wakatani, *Plasma Physics: Basic Theory with Fusion Applications*, 3d ed., 2000.

Thermoregulation

The processes by which many animals actively maintain the temperature of part or all of their body within a specified range in order to stabilize or optimize temperature-sensitive physiological processes.

Thermoregulation is evident when the temperature of part or all of the body of a free-living animal is consistently different from, or less variable than, the temperature of a collection of inanimate objects of identical external properties scattered randomly around the same habitat. Body temperatures of normally active animals may range from 32 to 115°F (0 to 46°C) or more, but the tolerable range for any one species is much narrower.

Animals are commonly classified as warm-blooded or cold-blooded. When the temperature of the environment varies, the body temperature of a warm-blooded or homeothermic animal remains high and constant, while that of a cold-blooded or poikilothermic animal is low and variable. However, supposedly cold-blooded reptiles and insects, when active, may regulate body temperatures within 2–4°F (1–2°C) of a species-specific value in the 93–108°F (34–42°C) range typical of birds and mammals. Supposedly warm-blooded mammals and birds may allow their temperature to drop to 37–68°F (5–20°C) during hibernation or torpor. Further, optimal temperature varies with organ, time of day, and circumstance. In mammals, the testes are too hot to produce sperm at the same temperature required for normal brain function. A temporary high temperature (hyperthermia) improves muscle power output, digestion, and defense against many bacterial and some parasitic infections. Most animals have daily body temperature cycles of 2–9°F (1–5°C) or more. The cycles appear important; for example, desert iguanas will die in a few days if they are held continuously at the temperature preferred during activity. Animals may conserve energy with daily or seasonal periods of very low body temperature (hypothermia). Thus, this classification is often misleading.

A better classification is based on the principal source of heat used for thermoregulation. Endotherms (birds, mammals) use heat generated from food energy. Ectotherms (invertebrates, fish, amphibians, reptiles) use heat from environmental sources. Ecological roles are related to this classification. For example, terrestrial ectotherms have an advantage in warm climates, particularly where food is scarce, as 30–40% of energy intake is available for growth and reproduction. Endotherms can continue to be active when endotherms must seek shelter from cold, but only 1–3% of energy intake is available for growth and reproduction. Therefore, the high rate of heat loss in water means that only highly active animals that are large or well insulated can be aquatic endotherms. This classification also has limitations, however. Endotherms routinely use external heat sources to minimize the food cost of thermoregulation, and some ectotherms use food energy for thermoregulation. Social insects are ectotherms as individuals, but may be endotherms as colonies. Thus, it is best to focus on mechanisms widely used for thermoregulation. *See* PHYSIOLOGICAL ECOLOGY (ANIMAL).

Thermal environment. An organism and its thermal environment are a functional unit. A useful starting point for thermoregulation is operative temperature, indicated by a thermometer with external properties identical to those of the animal (for example, a taxidermic mount on a hollow metal core with a thermometer inside), located at the same point as the animal and with the same orientation to sun and wind. Such a thermometer combines air temperature and solar radiation in proportions determined by wind and external properties of the animal such as size, shape, color, fur, or feathers. At a fixed air temperature such as 68°F (20°C), operative temperature is increased by sunlight and decreased by wind. Sunlight usually heats darker-colored objects more than lighter ones. Wind cools small objects more effectively than large ones. Thus, small or light-colored animals are most sensitive to air temperature and wind, while large or dark-colored animals are relatively more sensitive to sun. Similarly, operative temperature may vary over the body. Appendages that are smaller than the torso are more sensitive to wind. The back and flanks may receive more solar radiation than the ventral side, which is usually shaded. In contrast, aquatic or burrowing animals are affected uniformly and almost exclusively by water or soil temperature. *See* CONVECTION (HEAT); HEAT RADIATION.

Passive thermoregulation. Variation in the body temperature of an animal may be noticeably reduced relative to typical animals when it or its nest has a particular color, shape, or size. The goal of thermoregulation is thus aided or achieved, but the means are passive because they do not involve a physiological or behavioral response to thermal conditions. For example, in an often-sunny environment, a light-colored animal may experience a cooler and less variable environment than would a darker one. The compass termites of Australia construct their nests as long, narrow, high structures with the long axis oriented within ±20° of a north-south line. Thus, a large area is heated by sunlight in the morning and evening when the air is cool and sunlight weak, while a much smaller area absorbs sunlight in the heat of the day. A fairly stable daytime colony temperature of 91–95°F (33–35°C) results. Large animals such as rhinoceroses and camels do not need to sweat much in the African sun. They passively store the solar heat in their massive bodies by allowing body temperature to increase 11–13°F (6–7°C). This heat is lost during the cool nights. A small increase in the temperature of a wet-skinned amphibian automatically results in a large increase in evaporation that prevents further rise in body temperature.

Behavior. Behavior is the most obviously active form of thermoregulation. Most animals are mobile, sensitive to their environment, and capable of complex behaviors. The simplest thermoregulatory behavior consists of moving to a favorable location. Fish can detect and respond to changes of 0.09°F (0.05°C) in water temperature. Terrestrial animals move short distances between sun and shade or wind and shelter to secure a favorable operative temperature. Therefore, some reptiles can regulate their temperature within 2–4°F (1–2°C) of a preferred value.

Because of heat storage, only the average operative temperature is important. For example, the antelope ground squirrel forages at lethal operative temperatures on the desert surface by allowing body temperature to rise 9°F (5°C). It then returns to a cool burrow, releases the stored heat, and repeats the cycle.

Operative temperature may be altered by changing posture in one place. Lizards face the sun to minimize the area exposed to solar heating, or orient broadside to maximize it. The sun also heats the surface that supports the animal, which then heats the boundary layer of relatively still air near the ground. Lizards basking in the sun in high mountains flatten their bodies onto the supporting surface to immerse themselves in warm, still, air. This behavior allows them to obtain an operative temperature up to 54°F (30°C) warmer than the temperature of free air. Conversely, gray gulls escape the stifling boundary layer on the Atacama desert in Chile, for example, by standing over their nests in the heat of the day. Some reptiles and amphibians also expand or contract pigmented cells in their skin to increase or decrease solar heating. Roadrunners and grebes increase solar heating by facing away from the sun and adjusting their rump feathers to expose patches of black skin. Many birds decrease solar heating by raising long feathers on their back to form a parasol that shades the body. Some ground squirrels use their tail as a similar sunshade.

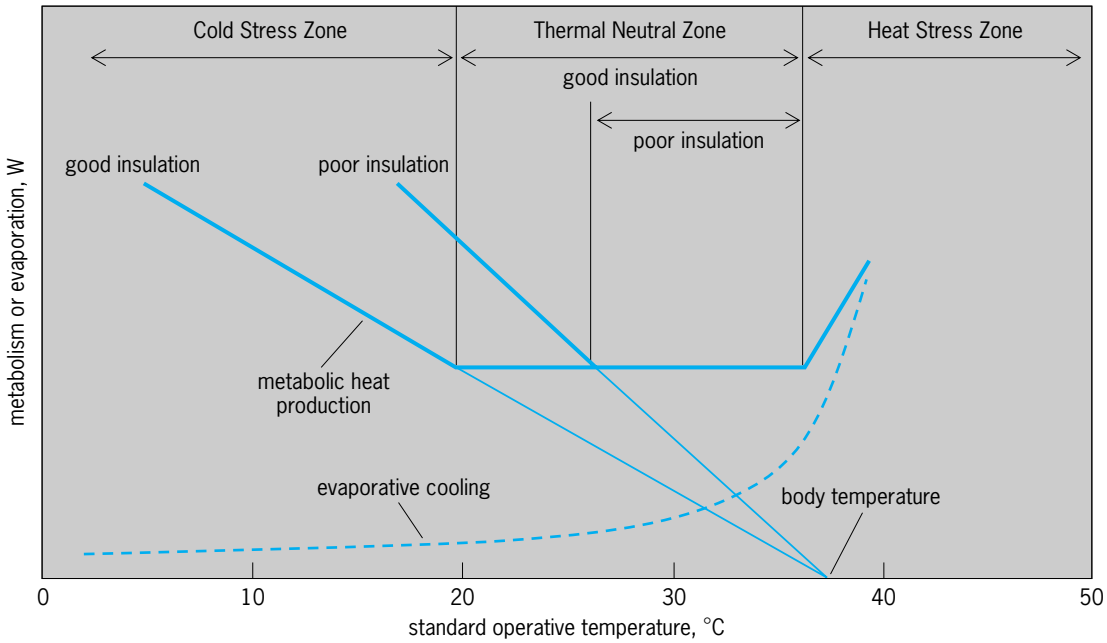
Evaporative cooling. Evaporation is an effective means of cooling the body. Evaporation from the respiratory mucous membranes is the most common mechanism. Nose breathing minimizes heat and water loss. Evaporation from the mucous membranes cools the nose during inhalation. During exhalation, water vapor condenses on the cool nasal membranes and is recovered. Evaporation can be greatly increased by exhaling from the mouth to prevent condensation. Additional increases in evaporation result from panting, that is, rapid breathing at the resonant frequency of the respiratory system. Evaporation from the eyes and the mucous membranes of the mouth and tongue is another source of cooling. Water is also evaporated from the skin of all animals, and can be varied for thermoregulation. Some desert frogs control evaporation by spreading an oily material on the skin. Reptiles, birds, and mammals have relatively impermeable skins, but evaporation can be increased by various means. The sweat glands in the skin of mammals are particularly effective and are one of the few purely thermoregulatory organs known. *See* SKIN.

Insulation. Changes in circulation can be used to regulate heat flow. Galápagos marine iguanas increase circulation to warm their bodies rapidly while basking on shore, and minimize circulation to slow cooling while foraging in the ocean. Countercurrent heat exchange is used to regulate heat flow to particular parts of the body while maintaining oxygen supply. Vessels carrying warm blood are in intimate contact with vessels carrying cool blood in the opposite direction. Large vessels may be divided into intermingled masses of small vessels to maximize heat exchange, forming an organ called a rete. However, retes can be bypassed by alternative circulation paths to regulate heat flow. Many animals living in warm environments have a rete that regulates brain temperature by cooling the arterial blood supply to the brain with blood draining from the nasal membranes, eyes, ears, or horns. Similar retes regulate heat loss from body to the legs in birds and mammals, from flight muscles to the abdomen of bumblebees, and from swimming muscles to the water in large fish such as tunas and sharks. *See* CARDIOVASCULAR SYSTEM; COUNTERCURRENT EXCHANGE (BIOLOGY).

Heat exchange with the environment is limited by the fur of mammals, feathers of birds, and furlike scales or setae of insects. Erection or compression of this insulation varies heat flow. Insulation thickness varies over the body to exploit variations in local operative temperature. Thermal windows are thinly insulated areas that are either shaded (abdomen of mammals, axilla of birds and mammals) or of small size (ears, face, legs) so that solar heating is minimized. Adjustments to circulation with retes and a change in posture vary heat flow to thermal windows. Conversely, the bison and vicuña present thickly furred areas to the wind in cold, windy conditions.

Metabolic heat. The oxidation of foodstuffs within the metabolic pathways of the body releases as much heat as if it were burned. Basal metabolism is the energy use rate of a fasting animal at rest. Activity, digestion, and thermoregulation increase metabolism above the basal rate. Endothermy is the utilization of metabolic heat for thermoregulation. Birds and mammals are typical examples, but significant endothermy also occurs in large salt-water fish, large reptiles, and large flying insects. Well-insulated animals can tolerate temperatures of −76°F (−60°C) indefinitely by using endothermy, but small or poorly insulated animals may use endothermy successfully only under mild conditions or for short periods.

The energy use of an idealized endotherm varies with standard operative temperature, that is, temperature which is adjusted for the effect of wind on the rate of metabolic heat loss (see **illus.**). In the thermal neutral zone, heat production is constant, and adjustments to insulation and evaporative cooling regulate body temperature. In the cold stress zone, insulation is at its maximum, and additional heat is produced for thermoregulation. The rate of increase of heat production as temperature falls is greater for poorly insulated animals than for well-insulated animals. Thus, a well-insulated animal has a relatively wider thermal neutral zone as well as an ability to tolerate lower temperatures. In the heat stress zone, evaporative cooling is increased by means such as panting that also increase heat production. Evaporative cooling increases nearly exponentially with temperature. If possible, endotherms use behavior to obtain temperatures in the thermal neutral zone unless the benefits of activity at stressful temperatures outweigh the costs.



Metabolic heat production and evaporative cooling of a hypothetical endotherm in response to standard operative temperature. The line relating heat production and temperature in the cold stress zone extrapolates to zero at body temperature.

Endotherms regulate only the temperature of the body core, that is, the brain, heart, and lungs. The heat production of these metabolically active organs is often supplemented with heat produced in muscles. Heat produced as a by-product of activity may substitute partially for thermoregulatory heat production, and imposes no thermoregulatory energy cost. Sharks and tuna use elaborate retes to conserve heat generated by swimming and so regulate body temperature at 86±9°F (30±5°C) in 9–86°F (5–30°C) water. In contrast, shivering produces heat only for thermoregulation and results in an extra cost. Large flying insects often cannot leave the ground until shivering of the flight muscles raises temperature to 95–104°F (35–40°C). Some animals have specialized heater organs for nonshivering thermogenesis, which is more efficient than shivering. Brown adipose tissue is a fatty tissue with a high density of mitochondria. It is found in the thorax of mammals, especially newborns and hibernators, and it warms the body core efficiently. Billfish and mackerel use modified eye muscles to generate large amounts of heat. Blood leaving these organs passes through a rete to warm the blood supply to the brain and the retina of the eye. However, many species, apparently including all birds, lack nonshivering thermogenesis. *See* METABOLISM.

Social insects. Massed colonies of some social insects thermoregulate as an endothermic superorganism. The combined metabolism of the million or more individuals and their fungus gardens in a colony of the African termite *Macrotermes bellicosus* heats air in the nest. The heated air rises up a system of chimneys and drives circulation in an elaborate system of ducts. The result is a colony temperature of 86±4°F (30±2°C) with outside air temperatures ranging 59–100°F (15–38°C). Bumblebees are adapted to colder climates and gather fibrous materials to insulate their nests. Colony metabolism can then regulate blood temperatures at 88±4°F (31±2°C) against Arctic summer air temperatures of 41–52°F (5–11°C). Honeybees have an even more complex thermoregulatory system.

Neural control. It was once believed that the preoptic area and hypothalamus of the brain was the master thermostat for vertebrates. However, the variety of mechanisms used in thermoregulation indicates a corresponding complexity in neural control. Temperature sensors distributed over the skin respond nearly immediately to changes in the environment and provide the major input. Nearly all parts of the central nervous system also respond to local thermal stimulation. These peripheral and central thermal inputs are integrated at a series of centers beginning in the spinal cord. This series clearly extends to the cerebral cortex, as a learning period is required before behavioral thermoregulation reaches maximum precision. Various components respond to the rate of temperature change as well as the difference between preferred and actual temperature. The neuroendocrine system then regulates metabolic heat production, the sympathetic nervous system controls blood flow, and the cerebral cortex controls behavioral thermoregulation. *See* ENDOCRINE SYSTEM (VERTEBRATE); HOMEOSTASIS; NERVOUS SYSTEM (VERTEBRATE).

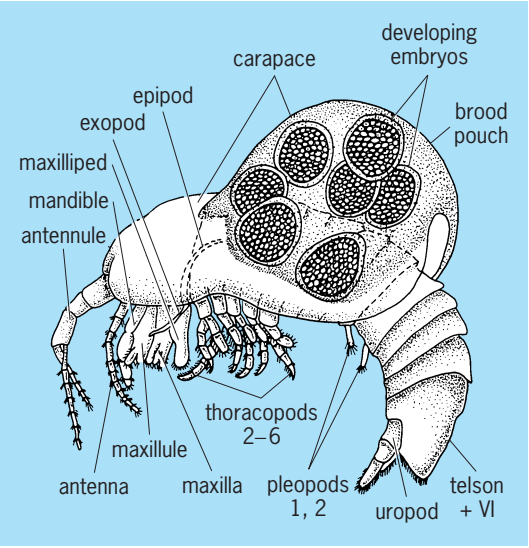
George Bakken

Bibliography. D. M. Gates, *Biophysical Ecology*, 1980; J. E. Heath (ed.), Thermoregulation in vertebrates, *Annu. Rev. Physiol.*, vol. 48, 1986;

B. Heinrich, *The Hot Blooded Insects*, 1993; P. C. Withers, *Comparative Animal Physiology*, 1992.

Thermosbaenacea

A small order of the crustacean superorder Peracarida. In thermosbaenaceans, the carapace, which may cover part of the cephalic region and one to several thoracic somites, is fused only to the first thoracic somite. The carapace of females is expanded to provide a dorsally positioned brood pouch where embryos hatch as subadults (see **illus.**). Eyes are reduced or absent. The abdomen consists of six somites and a telson; however, in *Thermosbaena*, at least, the telson and sixth somite are fused to form a pleotelson. The first pair of thoracic appendages are modified as maxillipeds, and may be sexually dimorphic; the remaining five to seven pairs provide the animals with locomotion. Sexes are separate.



An ovigerous female, *Thermosbaena* (Thermosbaenacea). (*After D. Barker, A study of Thermosbaena mirabilis (Malacostraca, Peracardia) and its reproduction, Quart. J. Micro. Sci., 103:261–286, 1962*)

Thermosbaenaceans are bottom dwellers; however, they can swim, and they are oriented ventral side up. They feed on detritus scraped from the substrate. Thermosbaenaceans have been found principally in thermal, but occasionally in cool, fresh- or brackish-water, lakes, springs, and interstitial coastal areas, and also in cave pools, in a geographic band stretching from the Mediterranean to the Caribbean and Gulf of Mexico. *See* PERACARIDA.

Patsy A. McLaughlin

Bibliography. T. E. Bowman and T. M. Iliffe, *Tulumella unidens*, a new genus and species of thermosbaenacean crustacean from the Yucatán Peninsula, Mexico, *Proc. Biol. Soc. Wash.*, 101(1):221–226, 1988; P. A. McLaughlin, *Comparative Morphology of Recent Crustacea*, 1980; F. R. Schram, *Crustacea*, 1986.

Thermosphere

A rarefied portion of the atmosphere, lying in a spherical shell between 50 and 300 mi (80 and 500 km) above the Earth's surface, where the temperature increases dramatically with altitude. Many satellites orbit in the upper thermosphere, and the drag on these satellites exerted by the atmosphere eventually brings them down to burn up in the lower atmosphere.

The thermosphere responds to the variable outputs of the Sun, the ultraviolet radiation at wavelengths less than 200 nanometers, and the solar wind plasma that flows outward from the Sun and interacts with the Earth's geomagnetic field. This interaction energizes the plasma, accelerates charged particles into the thermosphere, and produces the aurora borealis and aurora australis, nearly circular-shaped regions of luminosity that surround the magnetic north and south poles respectively. Embedded within the thermosphere is the ionosphere, a weakly ionized plasma. *See* IONOSPHERE; MAGNETOSPHERE; PLASMA (PHYSICS); SOLAR WIND.

Global mean structure. The atmosphere below about 50 mi (80 km) is almost entirely molecular: about 78% molecular nitrogen (N_2), 21% molecular oxygen (O_2), and 1% other minor gas constituents, such as ozone (O_3), carbon dioxide (CO_2), and argon (Ar). Above 50 mi (80 km), in the thermosphere, these molecular species are subjected to intense solar ultraviolet radiation and photodissociation that gradually turns the molecular species into the atomic species oxygen (O), nitrogen (N), and hydrogen (H). Solar radiation at wavelengths less than 130 nm can also ionize both molecular and atomic species and turn the neutral molecular and atomic species into a plasma at altitudes above 600 mi (1000 km). Thus there are various upper-atmosphere regimes (**Fig. 1**). Up to above 60 mi (100 km), atmospheric turbulence

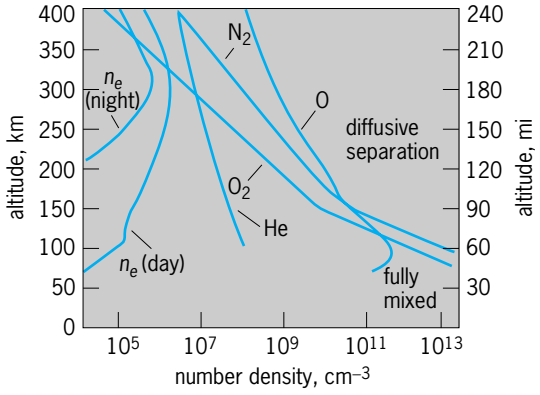


Fig. 1. Distribution of some important constituents in the thermosphere: atomic oxygen (O), molecular nitrogen (N_2), molecular oxygen (O_2), helium (He), and ionospheric electrons (n_e) for both daytime and nighttime conditions.

keeps the atmosphere well mixed, with the molecular concentrations dominating in the lower atmosphere. Above 60 mi (100 km), solar ultraviolet radiation most strongly dissociates molecular oxygen, and there is less mixing from atmospheric turbulence. The result is a transition area where molecular diffusion dominates and atmospheric species settle according to their molecular and atomic weights. Above 60 mi (100 km), atomic oxygen is the dominant species; above about 300 mi (500 km), helium (He) is dominant, and eventually atomic hydrogen, from the photodissociation of water vapor, becomes the dominant species at altitudes above 600 mi (1000 km).

The number of particles per unit volume (density values) of ionospheric electrons that peak near (180–240 mi) (300–400 km) depends upon the intensity of solar ultraviolet radiation (Fig. 1). The ionospheric densities are greatest during the day, and the lower ionosphere can chemically recombine at night. The dominant ion in the upper ionosphere is O^+, eventually supplemented by H^+ at high altitudes.

About 60% of the solar ultraviolet energy absorbed in the thermosphere and ionosphere heats the ambient neutral gas and ionospheric plasma; 20% is radiated out of the thermosphere as airglow from excited atoms and molecules; and 20% is stored as chemical energy of the dissociated oxygen and nitrogen molecules, which is released later when recombination of the atomic species occur. Most of the neutral gas heating that establishes the basic temperature structure of the thermosphere is derived from excess energy released by the products of ion-neutral and neutral chemical reactions that occur in the thermosphere and ionosphere. This heating is so strong in the rarefied upper atmosphere that heating rates range from about 50 K (90°F) per day near 60 mi (100 km), in the lower thermosphere, to 1000 K (1800°F) per day near 300 mi (500 km), in the upper thermosphere. *See* AIRGLOW; ULTRAVIOLET RADIATION.

Atomic and molecular gases in the thermosphere neither effectively radiate heat to space by infrared radiation nor provide the cooling rates needed to balance the intense heating rates. The average vertical temperature profile is, therefore, determined by a balance of local solar heating by the downward conduction of molecular thermal product to the region of minimum temperature near 50 mi (80 km; **Fig. 2**). There, energy is radiated to space by optically active infrared constituents such as carbon dioxide and ozone, which cannot exist at higher altitudes in the thermosphere because of intense solar photodissociation of these species. Infrared radiation from atomic species, such as atomic oxygen, is small in the thermosphere. *See* INFRARED RADIATION.

For heat to be conducted downward within the thermosphere, the temperature of the thermosphere must increase with altitude. In fact, the global mean temperature increases from about 200 K (−100°F) near 80 km (50 mi) to 700–1400 K (800–2100°F) above 180 mi (300 km), depending upon the intensity of solar ultraviolet radiation reaching the Earth. Above 180 mi (300 km), molecular thermal conduction occurs so fast that vertical temperature differences are largely eliminated; the isothermal temperature in the upper thermosphere is called the exosphere temperature. Above about 300 mi (500 km), in the exosphere, collisions between atmospheric gases are so rare that individual gas particles can either escape the Earth's gravitational pull or can obtain ballistic trajectories in their transit.

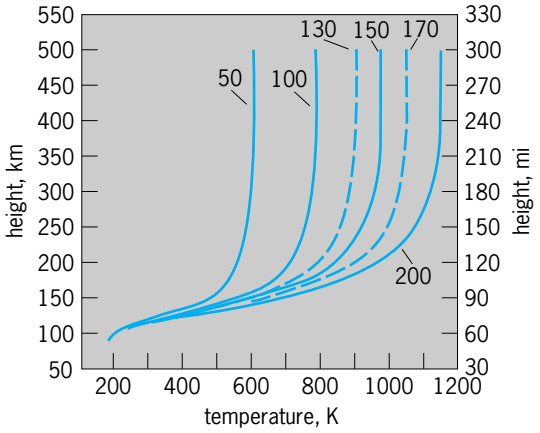


Fig. 2. Global mean-temperature profile in the thermosphere for various levels of solar activity. Labels on curves correspond to F10.7 values, which represent the level of the 10.7-cm radio emission from the Sun in units of 10^{-22} W m^{-2} Hz^{-1}. The value F10.7 = 50 is a hypothetical calculation, if the Sun should drop to this value. An F10.7 of about 70 represents a solar-cycle minimum, whereas an F10.7 of 200 represents solar cycle maximum activity. °F = (K × 1.8) − 459.67.

The temperature profile is strongly dependent upon the intensity of solar ultraviolet radiation, which varies with time and correlates with the sunspot cycle (approximately 11 years). During sunspot minimum the solar ultraviolet output produces an exospheric temperature of 500–700 K (440–800°F). At sunspot maximum the solar ultraviolet output can be as much as 10 times stronger at short wavelengths (10–30 nm) and 2–3 times stronger at longer ultraviolet wavelengths (100–150 nm), producing global exospheric temperatures of 1200–1500 K (1700–2200°F).

The chemical composition and densities of the thermosphere and ionosphere also vary with the solar ultraviolet output and vertical temperature profile. The intense solar radiation at sunspot maximum dissociates more molecular oxygen and nitrogen, thereby increasing the magnitude of the oxygen and nitrogen chemical cycles. It also produces larger electron and ion number densities within the ionosphere and drives other chemical cycles. In addition, the increased heating causes the atmosphere to expand outward from the Earth. Satellites orbiting at a given altitude thereby experience variations in atmospheric drag as the atmosphere expands and contracts in response to changing solar ultraviolet radiative output. *See* SUN.

Thermospheric circulation. As the Earth rotates, absorption of solar energy in the thermosphere

undergoes a daily variation. Dayside heating causes the atmosphere to expand and the loss of heat at night contracts it. This heating pattern creates pressure differences that drive a global circulation, transporting heat from the warm dayside to the cool nightside.

The meteorology of the thermosphere differs from the familiar weather experienced in the lower atmosphere. The fluid motions are governed by the same system of equations that is used by meteorologists to study the lower atmosphere; however, there are notable differences between some of the physical and chemical processes in the upper and lower atmosphere. In the troposphere temperature decreases with altitude, while in the thermosphere temperature increases with altitude. For this reason the thermosphere is more stable than the troposphere. Also, as the air in the thermosphere becomes rarefied, turbulence ceases; and the motions are strongly influenced by molecular diffusion, which provides an additional stability. Viscosity primarily transfers momentum between various altitudes, smoothing out vertical gradients in wind velocity, just as molecular thermal conductivity produces an isothermal vertical temperature profile. Above about 300 km (180 mi), the atmospheric layers are coupled so strongly by viscosity that the wind velocity becomes nearly constant with altitude.

Another important force that acts in the thermosphere is an ion drag force produced when the neutral gases in the thermosphere collide with the plasma of the ionosphere. Above about 75 mi (125 km), where the ion gyrofrequency (the frequency at which an ion spirals around a magnetic field line) is greater than the ion-neutral collision frequency, ions are essentially locked to the Earth's geomagnetic field lines and can move across them only in response to an electric field. A neutral-gas wind flowing through the relatively immobile ions experiences a collisional drag that is largest at the peak of the ionospheric layer, providing the main resistance that balances the driving pressure force within the upper thermosphere. In the lower thermosphere, however, the ion drag force is not as effective. The pressure force is primarily balanced by the Coriolis force, and winds flow perpendicular to pressure gradients, as they do in the lower atmosphere.

Thus, meteorology in the thermosphere is considerably different from that in the lower atmosphere. The flow is stable and viscous. Because of the stability and viscosity, there are no instabilities that lead to the development of high-pressure regions and low-pressure storms, such as those that exist in the troposphere. The thermosphere, however, can easily support the transmission of waves such as the diurnal and semi-diurnal tides, gravity waves generated by plasama processes, auroral particle inputs, and the Joule dissipation of ionospheric current systems. Joule dissipation is electrical resistive heating by a current flowing through, and electrons and ions colliding with, neutral particles, thereby transferring directed motion into heat. A crude visualization of the flow in the thermosphere is a wiggling bowl of viscous jelly responding to the variable forcings of solar ultraviolet radiation, auroral inputs, and complex neutral-gas-plasma interactions. *See* TROPOSPHERE.

Auroral influences. In addition to solar radiative heating driving the basic thermospheric circulation, there is an important heat and momentum source at high magnetic latitudes associated with auroral activity. The aurora is driven by the interactions of the solar-wind plasma with the Earth's geomagnetic field. Complex interactions between the solar wind and magnetosphere energize the plasma, causing some of it to be transferred into the ionosphere and thermosphere. Energetic electrons and ions bombard the atmosphere in the auroral ovals in both magnetic hemispheres. The electrons are sufficiently energetic to ionize atoms and dissociate molecules and produce copious amounts of emissions that can be seen from the ground and by satellites from space. The electrons flowing into the ionosphere and thermosphere carry an electric current of about a million amperes. During intense geomagnetic or auroral storms, the electric currents and energetic particles may deposit as much energy at high latitudes as that from the absorption of solar ultraviolet radiation. Typically, the global energy input during very quiet auroral activity is about 10^{10} W. This increases to 10^{11} W during moderate activity and up to 10^{12} W for short periods during intense auroral storms. *See* GEOMAGNETIC VARIATIONS.

The thermosphere responds to this energy input in a number of ways. The strong and rapid changes in energy input generate a variety of atmospheric waves that can transport energy from high to low latitudes. In addition, the more sustained energy inputs generate a mean circulation that flows from high to low latitudes; they can be large enough to compete with the circulation generated by solar radiative energy absorption that is basically from low to high latitudes. In the mean thermospheric circulation for equinox and solstice conditions, there is competition between the solar radiative forced circulation and auroral forced circulation for the three levels of geomagnetic activity (**Fig. 3**). During geomagnetic quiet conditions the circulation is from the Equator toward both poles during equinox and from the summer to winter pole for solstice. During moderate auroral activity the energy and momentum additions at magnetic conjugate high latitudes generate a pole-to-Equator component that is superimposed upon the solar-driven circulation. During an auroral storm the pole-to-Equator circulation driven by auroral heating dominates. Thus, it is evident that the mean motions in the thermosphere are in a constant state of agitation, depending upon the amount and duration of high-latitude heating from auroral processes. *See* AURORA; EQUINOX; MIDDLE-ATMOSPHERE DYNAMICS; SOLSTICE.

Global change. The Earth's atmosphere is subject to global change because of the release of trace gases, such as carbon dioxide and methane, by human activity or natural causes. These greenhouse gases have

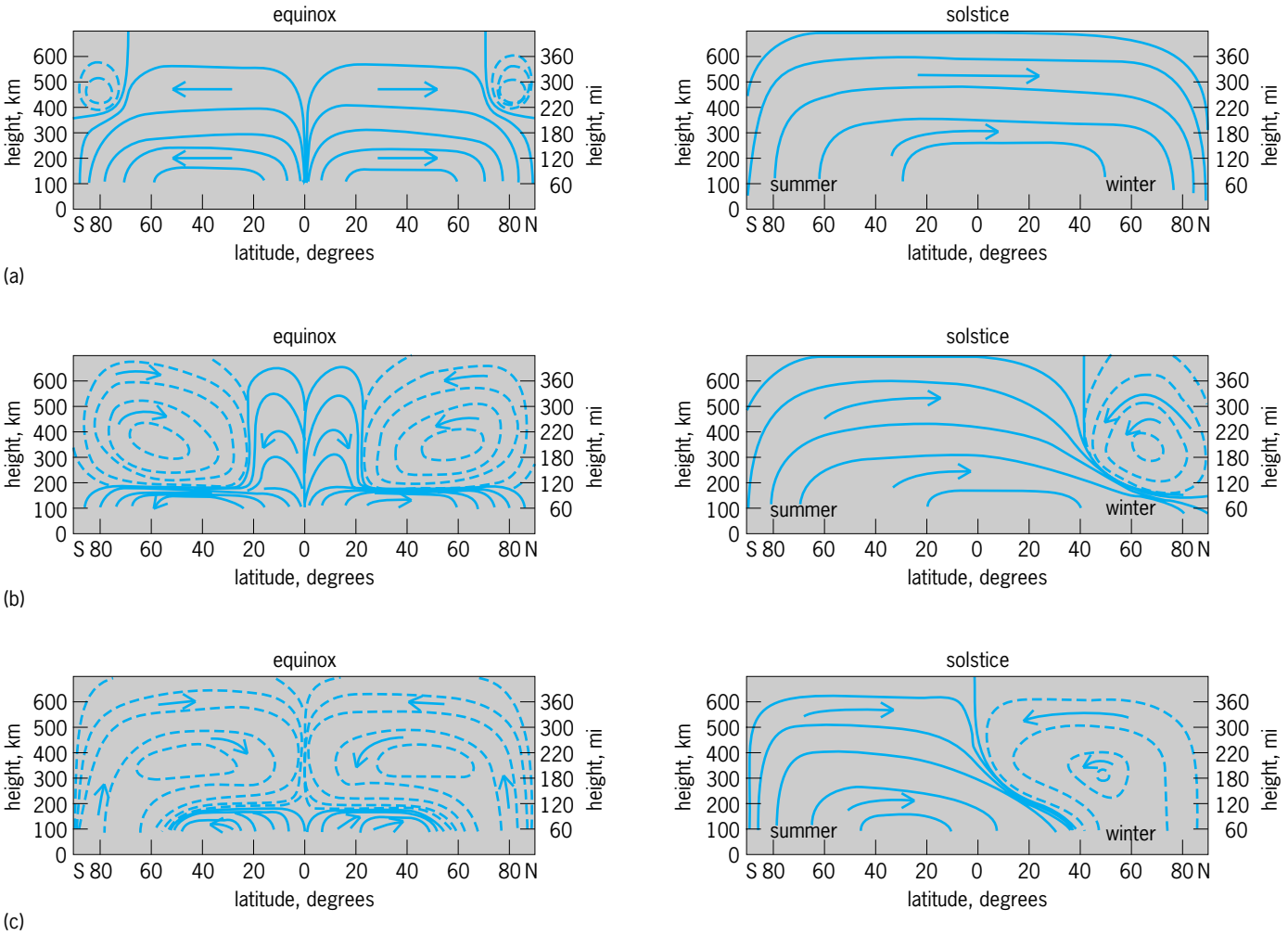


Fig. 3. Zonal mean meridional circulation in the Earth's thermosphere during equinox and solstice for various levels of auroral activity: (*a*) quiet auroral activity where circulation is primarily driven by solar ultraviolet heating; (*b*) average auroral activity where heating is 10^{11} W; and (*c*) during geomagnetic storms where the heating is about 10^{12} W. The contours schematically illustrate the mass flow, and the arrows indicate the direction of the motion.

been projected to cause changes in the lower atmosphere structure, where the troposphere is expected to warm a few degrees and the stratosphere to cool by 10–20 K (18–36°F) in response to a doubling of certain trace gases. Model calculations have been made that suggest that global change effects will also occur in the upper mesosphere of the atmosphere, primarily due to the increased carbon dioxide cooling to space. With a doubling of the concentrations of carbon dioxide and methane (from present-day conditions) that is expected to occur by the end of the twenty-first century, the mesosphere is predicted to cool by about 10 K (18°F) and the thermosphere to cool by about 50 K (90°F). In the lower atmosphere, increased carbon dioxide traps radiation and the atmosphere warms. But as the radiation transfers upward, it eventually finds space and it is no longer trapped. Therefore, increased concentrations of carbon dioxide allow more energy to escape to space. The Earth in steady state is in balance between solar ultraviolet radiation and visible energy absorbed and infrared carbon dioxide cooling to space. This cooling will cause the atmosphere to contract, reducing by 50% the drag on satellites at a given altitude and causing a redistribution of certain minor species. The peak of the ionospheric layer is predicted to drop by 12 mi (20 km), with a redistribution of the present-day plasma. The overall consequences of these effects is not known at the present time; however, it is likely that thermal tides, thermosphere-ionosphere interactions, and the thermosphere and ionosphere response to solar and auroral variability will all change considerably toward the end of the twenty-first century. *See* ATMOSPHERE; CLIMATE MODELING; MESOSPHERE; STRATOSPHERE.

R. G. Roble

Bibliography. S.-I. Akasofu and Y. Kamide (eds.), *The Solar Wind and the Earth*, 1987; Geophysics Study Committee, National Academy of Sciences, *The Upper Atmosphere and Magnetosphere: Studies in Geophysics*, 1977; S. Kato, *Dynamics of the Upper Atmosphere*, 1980; M. C. Kelly, *The Earth's Ionosphere: Plasma Physics and Electrodynamics*, 1989; M. H. Rees, *Physics and Chemistry of the Upper Atmosphere*, Cambridge, 1989.

Thermostat

An instrument which directly or indirectly controls one or more sources of heating and cooling to maintain a desired temperature. To perform this function a thermostat must have a sensing element and a transducer. The sensing element measures changes in the temperature and produces a desired effect

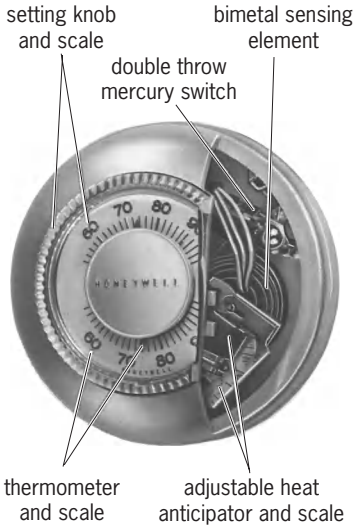


Fig. 1. Typical heat-cool thermostat. (*Honeywell Inc.*)

on the transducer. The transducer converts the effect produced by the sensing element into a suitable control of the device or devices which affect the temperature. The most commonly used principles for sensing changes in temperature are (1) unequal rate of expansion of two dissimilar metals bonded together (bimetals), (2) unequal expansion of two dissimilar metals (rod and tube), (3) liquid expansion (sealed diaphragm and remote bulb or sealed bellows with or without a remote bulb), (4) saturation pressure of a liquid-vapor system (bellows), and (5) temperature-sensitive resistance element.

The most commonly used transducers are a switch that makes or breaks an electric circuit, a potentiometer with a wiper that is moved by the sensing element, an electronic amplifier, and a pneumatic actuator.

The most common thermostat application is for room temperature control. **Figure 1** shows a typical on-off heating-cooling room thermostat. In a typical application the thermostat controls a gas valve, oil burner control, electric heat control, cooling compressor control, or damper actuator.

To reduce room temperature swings, high-performance on-off thermostats commonly include a means for heat anticipation. The temperature swing becomes excessive if thermostats without heat anticipation are used because of the switch differential (the temperature change required to go from the break to the make of the switch), the time lag of the sensing element (due to the mass of the thermostat) in sensing a change in room temperature, and the inability of the heating system to respond immediately to a signal from the thermostat.

To reduce this swing, a heater element (heat anticipator) is energized during the on period. This causes the thermostat to break prematurely. **Figure 2** shows a comparison of the room temperature variations when a thermostat with and without heat anticipation is used.

The same anticipation action can be obtained on cooling thermostats by energizing a heater (cool anticipator) during the off period of the thermostat. Room thermostats may be used to provide a variety of control functions, such as heat only; heat-cool; day-night, in which the night temperature is controlled at a lower level; and multistage, in which there may be one or more stages of heating, or one or more stages of cooling, or a combination of heating and cooling stages.

Thermostats are also used extensively in safety and limit applications. Thermostats are generally of the following types: insertion types that are mounted on ducts with the sensing element extending into a duct, immersion types that control a liquid in a pipe or tank with the sensing element extending into the liquid, and surface types in which the sensing element is mounted on a pipe or similar surface.

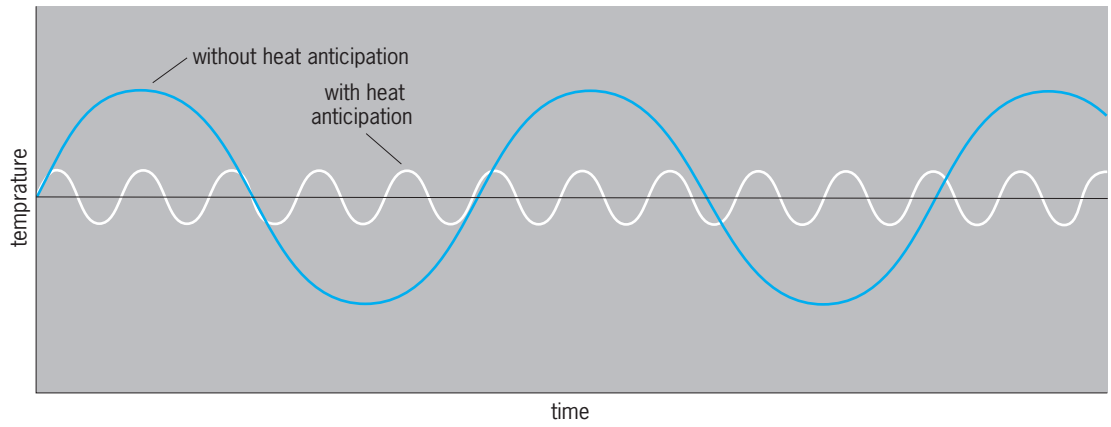


Fig. 2. Comparison of temperature variations using a timed on-off thermostat with and without heat anticipation.

See COMFORT HEATING; OIL BURNER.

Nathaniel Robbins, Jr.

Bibliography. V. C. Miles, *Thermostatic Control*, 2d ed., 1974; J. Olivieri, *How to Design Heating-Cooling Comfort Systems*, 1987; R. K. Schneider, *HVAC Control Systems*, 2d ed., 1988; J. Trost, *Efficient Buildings: Heating and Cooling*, 1987; G. Vacuumschmelze, *Thermostat Metals*, 1984.

Thermotherapy

The treatment of disease by the local or general application of heat. The following discussion is limited to the local application of heat as an adjunct to therapeutic management of musculoskeletal and joint diseases. The most commonly used methods for this form of treatment include hot packs, hydrotherapy, radiant heat, shortwave diathermy, microwave diathermy, ultrasound, and laser therapy.

Biophysics. The reason so many different methods are employed is that each modality heats selectively different anatomical structures, and thus the modality selected for a given treatment is based on the temperature distribution produced in the tissues. For vigorous heat application to a given site, the location of the peak temperature produced by the modality must coincide with the site so that maximally tolerated tissue temperatures can be obtained there without burning elsewhere. Customarily, the modalities are divided into those that heat superficial tissues and those that heat deep-seated tissues.

Heating superficial tissues. Hot packs, hydrotherapy, and radiant heat are used to heat superficial tissues. The heat is transferred from a hot pack or hot water, primarily through conduction and convection, whereas with application of radiant heat, the photons, or electromagnetic radiation, emitted from the lamp are converted into heat by absorption. However, since absorption occurs in the most superficial tissues, this form of therapy is classified as superficial heating. The photons used—from the yellow and red bands of the visible spectrum and the infrared band within the invisible range—have a relatively long wavelength and low energy content. *See* ELECTROMAGNETIC RADIATION.

Heating deep-seated tissues. Photons of ultraviolet radiation, x-rays, and radium penetrate deeper into the tissues and produce photochemical reactions long before the temperature increases significantly. Other forms of energy used for heating deep-seated tissues include shortwave, microwave, and ultrasound: shortwave diathermy induces a high-frequency (27-MHz) electromagnetic current in the tissues; microwaves operate at frequencies of 915 and 2456 MHz; and ultrasonic waves represent an acoustic vibration of 1-MHz frequency.

In the case of shortwaves, the electrical properties of the tissue determine the distribution and absorption of current and thus the heating pattern. Shortwaves may heat joints covered by very little soft tissue; superficial musculature; or, if applied with internal electrodes, the pelvic organs. Microwaves are reflected at tissue interfaces and are absorbed; microwave application generally heats selectively tissues with high water content. Ultrasound is reflected at interfaces between tissues of different acoustic impedance, which is the product of density times sound velocity. Ultrasound selectively heats these interfaces because (1) the reflected wave is superimposed upon the incoming wave and increases the available energy for absorption and (2) a significant part of the longitudinal compression wave of ultrasound is converted into shear waves at the interface. The shear waves are locally absorbed rapidly. Finally, the interface between a medium with low coefficient of absorption (such as soft tissue) and a medium with high coefficient (such as bone) is heated selectively, because most of the energy is absorbed in the superficial layers of the medium with high coefficient. By using these principles, ultrasound selectively heats deep-seated joints. *See* MICROWAVE; ULTRASONICS.

Therapeutic effects of heat. The therapeutic effects produced by heating selectively by ultrasound include an increase in the extensibility of collagen tissues. Disease or injury, such as arthritis, burns, scarring, or long-term immobilization in a cast, may cause shortening of collagen tissue producing severe limitation of the range of motion at a joint. This application of heat (mostly ultrasound) is often used in conjunction with physical therapy, such as stretching and other joint mobilization techniques. It also has been shown that the stiffness associated with rheumatoid arthritis can be alleviated by heat application.

Heat applied by using shortwaves and microwaves may reduce muscle spasms secondary to musculoskeletal pathology, such as when a slipped intervertebral disc impinges upon a nerve root. (Interestingly, cooling the area may have the same physiological effect.) Selective heating of muscle with microwave radiation has been used to accelerate absorption of hematomas (a collection or clot of blood in the muscle) and to prepare for stretching of the contracted and stiffened muscle. Unlike microwaves, ultrasound is not absorbed significantly in normal homogeneous muscle.

Heat therapy in the form of hyperthermia has been used as an effective adjunct to cancer therapy in combination with ionizing radiation in the form of x-rays or radium therapy.

Laser therapy. The most commonly used laser in physical therapy is the helium-neon laser, which produces photons of a wavelength of 632.3 nanometers at lower intensities than those applied in surgical procedures. At this wavelength the depth of penetration and the heating effect are similar to infrared light. However, the major difference between laser light and diffuse light of the same wavelength is that laser is a columnated beam of photons of the same frequency, with the wavelength in phase so that any desirable intensity can be easily produced. *See* LASER PHOTOBIOLOGY.

Side effects. All the therapies described are valuable adjuncts to therapy with drugs and other approaches. However, the energy output of all these

modalities is high enough to produce tissue destruction if not used with great care. Examples of other significant side effects are the malfunction of pacemakers as a result of shortwave or microwave application, the overheating of metal implants and the tissues surrounding them, and the production of destructive lesions in the eye when using diathermy. Ultrasound may produce cavitation (gas bubbles) in the fluid media of the eye, with possible destruction of the retina and blood vessels. High-intensity lasers may burn, destroy, or produce the type of destructive effects associated with ionizing radiation.

Justus F. Lehmann

Bibliography. F. J. Kottke and J. F. Lehmann (eds.), *Krusen's Handbook of Physical Medicine and Rehabilitation*, 4th ed., 1990; J. F. Lehmann (ed.), *Therapeutic Heat and Cold*, 4th ed., 1990.

Thévenin's theorem (electric networks)

A theorem from electric circuit theory. It is also known as the Helmholtz or Helmholtz-Thévenin theorem, since H. Helmholtz stated it in an earlier form prior to M. L. Thévenin. Closely related is the Norton theorem, which will also be discussed. Laplace transform notation will be used. *See* LAPLACE TRANSFORM.

Thévenin's theorem states that at a pair of terminals a network composed of lumped, linear circuit elements may, for purposes of analysis of external circuit or terminal behavior, be replaced by a voltage source $V(s)$ in series with a single impedance $Z(s)$. The source $V(s)$ is the Laplace transform of the voltage across the pair of terminals when they are open-circuited; $Z(s)$ is the transform impedance at the two terminals with all independent sources set to zero (**Fig. 1**). The Thévenin equivalent may also be found experimentally.

Norton's theorem states that a second equivalent network consists of a current source $I(s)$ in parallel with an impedance $Z(s)$. The impedance $Z(s)$ is identical with the Thévenin impedance, and $I(s)$ is the Laplace transform of the current between the two terminals when they are short-circuited (**Fig. 2**).

Thévenin's and Norton's equivalent networks are related by the equation $V(s) = Z(s) \cdot I(s)$. This may be seen by comparing Figs. 1*b* and 2*b*. In Fig. 1*b*, if terminals A and B are short-circuited, a current $I(s) = V(s)/Z(s)$ will flow; this is also true in Fig. 2*b*. Similarly the open-circuit voltage in Fig. 2*b* is $V(s) = Z(s) \cdot I(s)$. *See* ALTERNATING-CURRENT CIRCUIT THEORY.

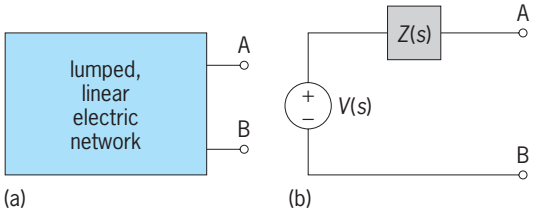


Fig. 1. Network and its Thévenin equivalent. (*a*) Original network. (*b*) Thévenin equivalent circuit.

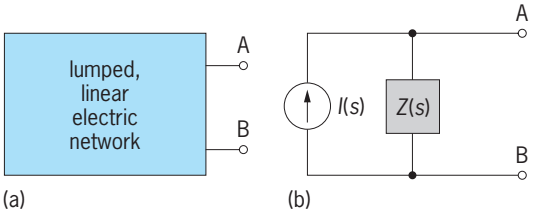


Fig. 2. Network and its Norton equivalent. (*a*) Original network. (*b*) Norton equivalent circuit.

These theorems are useful for the study of the behavior of a load connected to a (possibly complex) system that is supplying electric power to that load. The system may be a power distribution system, such as in a home or office, in which case the load may be lights or appliances. The system may be an electronic amplifier, in which case the load may be a loudspeaker. However, the theorem is of no value in studying the internal system behavior, because the behavior of the equivalent network is very different from that of the original.

Examples. Two examples will be used to show how Thévenin and Norton equivalent networks may be calculated from the original network and then used for some typical calculations.

Power distribution circuit. Suppose a simplified power distribution circuit contains the elements shown in **Fig. 3***a*. Sinusoidal steady-state operation is assumed. In this circuit the voltage across A, B is given by Eq. (1), and with the source set to zero, the impedance at A, B is given by Eq. (2). Thus the

$$\begin{aligned} V_{AB} &= \frac{Z_3}{Z_1 + Z_2 + Z_3} \cdot V_s \\ &= \frac{-j122.6}{0.55 + j3.5 - j122.6} \cdot 240/\underline{0^\circ} \\ &= 247.1/\underline{-0.26^\circ}\ V \end{aligned} \tag{1}$$

$$\begin{aligned} Z_{TH} &= Z_4 + \frac{Z_3(Z_1 + Z_2)}{Z_1 + Z_2 + Z_3} \\ &= 0.35 + \frac{(-j122.6)(0.55 + j3.5)}{0.55 + j3.5 - j122.6} \\ &= 3.72/\underline{75.47^\circ}\ \text{ohms} \end{aligned} \tag{2}$$

Thévenin equivalent is given in Fig. 3*b*, and the Norton equivalent in Fig. 3*c*.

When a load $Z_L = 18.3 + j2.1 = 18.42/\underline{6.55^\circ}$ is connected, at A, B, the current through the load is given by Eq. (3), and the power delivered to the load

$$\begin{aligned} I_L &= \frac{V_{TH}}{Z_{TH} + Z_L} = \frac{247.1/\underline{-0.26^\circ}}{3.72/\underline{75.47^\circ} + 18.42/\underline{6.55^\circ}} \\ &= (12.32)^2/\underline{-16.77^\circ} \end{aligned} \tag{3}$$

is $(12.32)^2(18.3) = 2.776$ kW. Other loads are handled in a similar fashion.

Amplifier. As a second example, suppose that the circuit of **Fig. 4***a* is a simplified model of an electronic amplifier, and that a load (loudspeaker) is to be connected at A, B. The circuit is driven by a current source $I(s)$. For analysis, the voltages $V_1(s)$ and

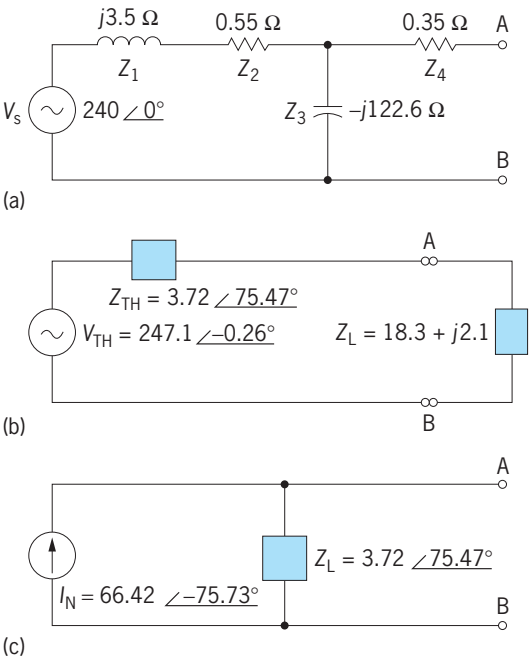


Fig. 3. Power distribution circuit and its Thévenin and Norton equivalents. (*a*) Original circuit. (*b*) Thévenin equivalent circuit with its load Z_L connected. (*c*) Norton equivalent circuit.

$V_2(s)$ are the transforms of the voltages on their respective nodes, and become the dependent variables in the analysis. The dependent or controlled source ($g_m V_1$) models the amplification. Two Kirchhoff current law equations, (4) and (5), may be used to find the Thévenin voltage, which is also V_2. Solution of this pair of equations gives Eq. (6).

$$I(s) = V_1\left(\frac{1}{R_1} + \frac{1}{R_2}\right) - V_2\left(\frac{1}{R_2}\right) \tag{4}$$

$$-g_m V_1 = -V_1\left(\frac{1}{R_2}\right) + V_2\left(\frac{1}{R_2} + \frac{1}{R_3}\right) \tag{5}$$

$$V_2(s) = \frac{(-g_m + 1/R_2)\cdot I(s)}{\dfrac{1}{R_1R_2} + \dfrac{1}{R_1R_3} + \dfrac{1}{R_2R_3} + \dfrac{g_m}{R_2}} \tag{6}$$

See KIRCHHOFF'S LAWS OF ELECTRIC CIRCUITS.

To find the Thévenin impedance, $I(s)$ must be set to zero, which leaves an infinite impedance in the branch, and an auxiliary current source $I_2(s)$ must be added between A and B (Fig. 4*b*). A new set of equations, (7) and (8), is written, and solved for

$$0 = V_1\left(\frac{1}{R_1} + \frac{1}{R_2}\right) - V_2\left(\frac{1}{R_2}\right) \tag{7}$$

$$-g_m V_1 + I_2(s) = -V_1\left(\frac{1}{R_2}\right) + V_2\left(\frac{1}{R_2} + \frac{1}{R_3}\right) \tag{8}$$

the ratio $V_2(s)/I_2(s)$, which is the desired Thévenin impedance; the controlled source must not be set to zero.

Solution of Eqs. (7) and (8) gives, after simplification, Eq. (9).

$$Z_{TH}(s) = \frac{V_2(s)}{I_2(s)} = \frac{R_3(R_1 + R_2)}{R_1 + R_2 + R_3 + g_m R_1 R_3} \tag{9}$$

To make a numerical example in a purely resistive circuit, let $R_1 = 2.0\ k\Omega$, $R_2 = 8.2$ kΩ, $R_3 = 400$ Ω, $g_m = 510 \times 10^3$ S (siemens), and $I = 4.0$ mA. Substitution of these into the equations gives a Thévenin voltage $V_{TH} = -31.96$ V, and a Thévenin impedance $Z_{TH} = 9.75$ Ω. This is shown in Fig. 4*c*, and the corresponding Norton equivalent is shown in Fig. 4*d*, where $I_N = -31.96/9.75 = -3.28$ A. In both cases the negative signs lead to a polarity reversal, which is reflected in Fig. 4*c* and *d*.

Suppose an 8.0-ohm speaker is connected at A, B. A current of 31.96/(8.0 + 9.75) = 1.80 A will flow, giving a power of $(1.80)^2(8.0) = 25.95$ W.

Proof. To prove this theorem, consider a general network with two accessible terminals, as in Figs. 1*a* and 2*a*, to which an auxiliary voltage source V and an impedance $Z_L(s)$ have been added (**Fig. 5**). Let this source be such as to cause $I_L = 0$ when switch S is closed. By superposition, Eq. (10) is valid. This

$$I_L = 0 = \frac{V_{TH}}{Z_{TH} + Z_L} - \frac{V}{Z_{TH} + Z_L} \tag{10}$$

shows that $V_{TH} = V$, and that the current that flows

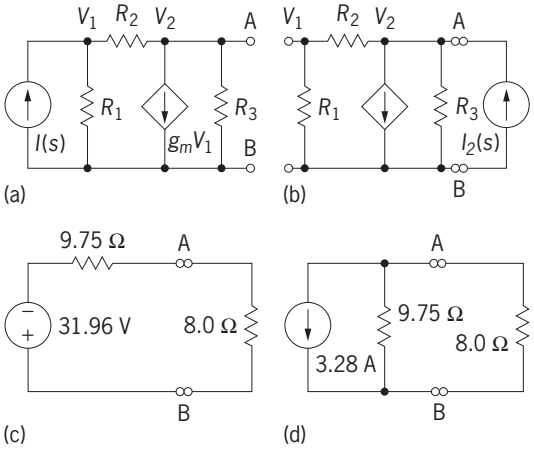


Fig. 4. Amplifier and its Thévenin and Norton equivalents. Numerical values are given in the text. (*a*) Original circuit. (*b*) Circuit constructed to find Thévenin impedance. (*c*) Thévenin equivalent circuit. (*d*) Norton equivalent circuit.

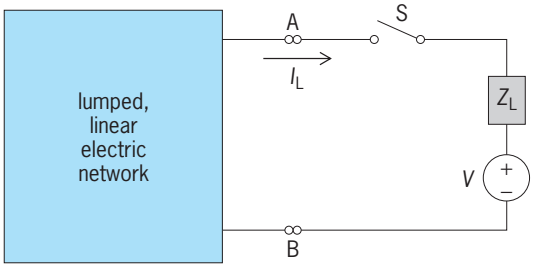


Fig. 5. Circuit constructed to demonstrate Thévenin's theorem.

when V is removed is given by Eq. (11).

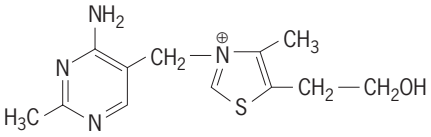
$$I_L = \frac{V_{TH}}{Z_{TH} + Z_L} \tag{11}$$

See NETWORK THEORY; SUPERPOSITION THEOREM (ELECTRIC NETWORKS). Edwin C. Jones, Jr.

Bibliography. R. DeCarlo and P. -M. Lin, *Linear Circuit Analysis*, 2d ed., 2001; J. D. Irwin and C.-H. Wu, *Basic Engineering Circuit Analysis*, 7th ed., 2001; M. Reed and R. Rohrer, *Applied Introductory Circuit Analysis*, 1999.

Thiamine

A water-soluble vitamin, also known as vitamin B_1 or aneurin, found in many foods; pork, liver, and whole grains are particularly good sources. Loss of the vitamin during production of flour and polished rice has led to the need for enrichment of these staples. The structural formula is shown below.



Chemistry. Thiamine is heat-labile, and considerable amounts are destroyed during cooking. It is unstable in alkaline solutions but relatively stable in acid solutions. It acts like a weak base and can be absorbed on basic ion-exchange materials such as decalso and fuller's earth, a property used to concentrate it so that it can be detected at the levels needed for analysis. Biological and microbiological methods for its estimation are available but are seldom used. A chemical assay based on the production of thiochrome, a fluorescent product obtained from alkaline oxidation of thiamine, is often used. The red blood cell transketolase assay, which is dependent upon the level of the coenzyme of thiamine, is the most commonly used functional assay for thiamine status.

Biochemistry. Thiamine functions as the coenzyme thiamine pyrophosphate (TPP) in enzyme systems that catalyze decarboxylations of α-keto acids and ketolations (condensation reactions of ketols) in sugar phosphates. Thiamine pyrophosphate–containing enzymes located in mitochondria catalyze the decarboxylation of pyruvate, α-ketoglutarate, and branched-chain α-ketoacids to eventually form acetyl-coenzyme A (CoA), succinyl-CoA, and amino acids (such as leucine, isoleucine, and valine), respectively (see **illustration**). These activities in higher organisms, including humans, are in association with other enzyme subunits that constitute so-called dehydrogenase complexes. The coenzyme TPP is also involved in the cytosol, where it associates with transketolase to effect interconversions of pentose to heptulose in the hexose monophosphate shunt, which is an alternative to the conventional glycolytic pathway of anaerobic glucose metabolism. Participation of thiamine as TPP in metabolic pathways is shown in the illustration.

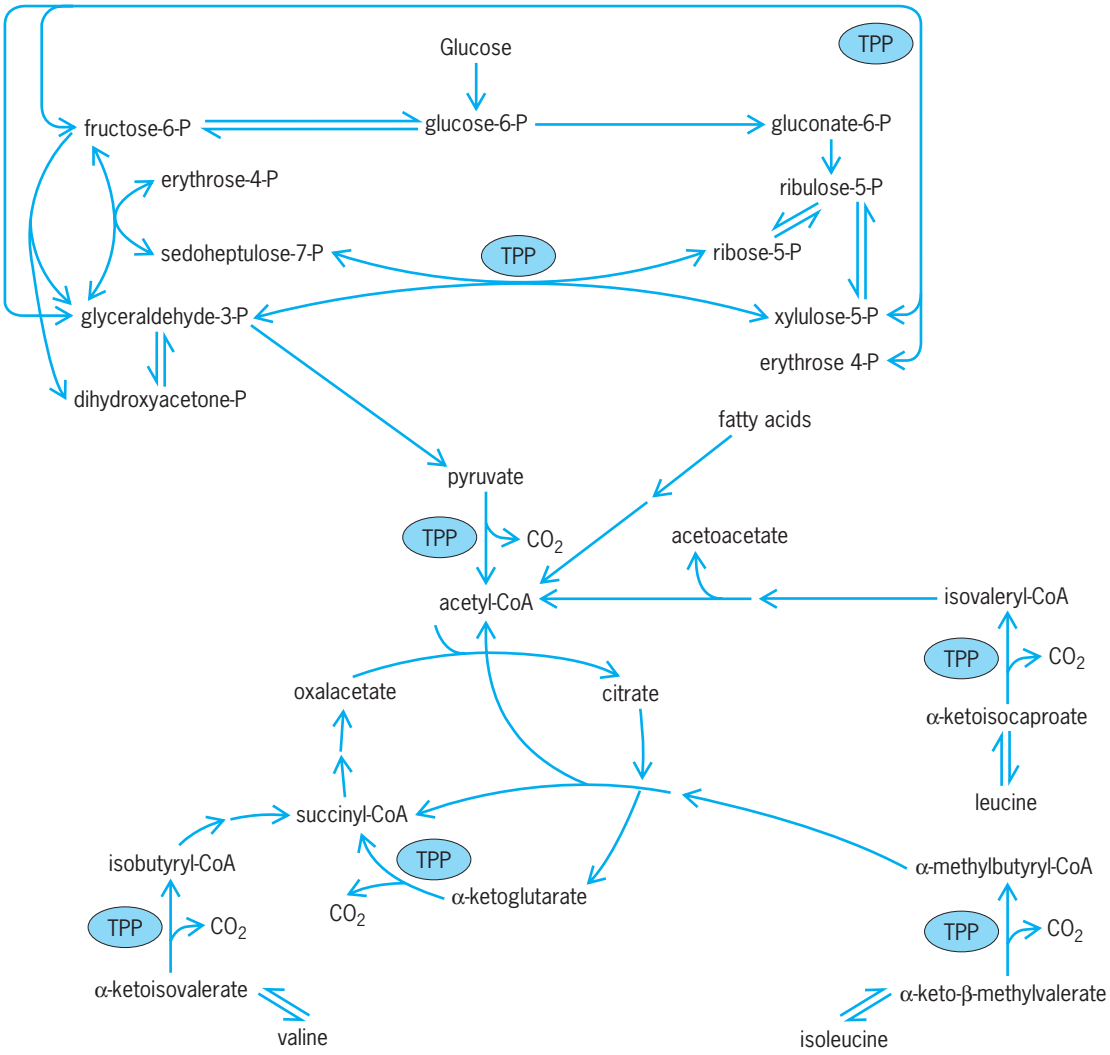
A recently characterized lyase that is required to shorten 3-methyl-branched fatty acids by α-oxidation also requires TPP. *See* CARBOHYDRATE METABOLISM; CITRIC ACID CYCLE; COENZYME; ENZYME.

Deficiency. Thiamine deficiency is known as beriberi in humans. Other species manifest the deficiency with polyneuritic conditions (which are characterized by degenerative or inflammatory lessions of several nerves simultaneously, usually symmetrical). Muscle and nerve tissues are affected by the deficiency, and poor growth is observed. People with beriberi are irritable, depressed, and weak; they often die of cardiac failure. Wernicke-Korsakoff syndrome, often observed in chronic alcoholics, is a cerebral beriberi characterized by brain lesions, liver disease, and partial paralysis, particularly of the motor nerves of the eye. As is the case in other vitamin B deficiencies, the deficiency of thiamine is commonly accompanied by insufficiencies of some of the other vitamins. In relatively rare cases of inborn errors of metabolism, pharmacologic doses of thiamine (5 to 10 mg/day) are required to prevent lactic acidosis due to low activity of the liver pyruvate dehydrogenase complex and ketoaciduria (the presence of keto acids in the urine) due to low activity of the branched-chain α-keto-acid dehydrogenase complex. *See* MALNUTRITION.

Dietary requirements. A dietary source of thiamine is required by all nonruminant animals that have been studied. Thiamine is the most poorly stored of the B vitamins. Individuals eating vitamin-deficient diets are likely to develop beriberi symptoms first. Approximately 5 mg of thiamine can be absorbed per day by normal adults. Excess thiamine given by mouth or parenterally is well tolerated but usually lost through excretion in the urine and feces. Thiamine requirements are related to caloric intake. More thiamine is required in high-carbohydrate than in high-fat diets. Some foods, particularly raw fish, contain enzymes that destroy thiamine. More thiamine is needed in altered physical states, such as hyperthyroidism, pregnancy, and lactation. Thiamine requirements of humans are primarily estimated by means of urinary excretion data and the red cell transketolase assay. The dietary allowances recommended in the United States by the Food and Nutrition Board of the Institute of Medicine are from 1.1 to 1.2 mg of thiamine per day for adults, with an additional increase to 1.4 mg/day for pregnant or breast-feeding women. *See* VITAMIN.

Stanley N. Gershoff; Donald B. McCormick

Manufacture. Industrial synthesis of thiamine is accomplished by linking chloromethylpyrimidine with 4-methyl-5-(β-hydroxyethyl)-thiazole to give thiamine. Another way to produce thiamine on an industrial scale is to convert 4-amino-5-cyanopyrimidine into the thioformyl-aminomethyl derivative via catalytic hydrogenation and reaction with sodium



Metabolic pathways involving thiamine pyrophosphate (TPP). CoA, coenzyme A; P, phosphate.

dithioformate. The resulting compound is then treated with 1-acetoxy-3-chloro-4-pentanone to form the thiazole ring in situ connected to the pyrimidine ring via a methylene bridge (U.S. Patents 2,193,858 and 2,218,350.) Fernand de Montmollin

Bibliography. C. J. Bates, Thiamin, in B. A. Bowman and R. M. Russell (eds.), *Present Knowledge in Nutrition*, 8th ed., pp. 184–190, ILSI Press, Washington, D.C., 2001; Food and Nutrition Board, Institute of Medicine, Thiamin, in *Dietary Reference Intakes: Thiamin, Riboflavin, Niacin, Vitamin B_6, Vitamin B_{12}, Pantothenic acid, Biotin, and Choline*, pp. 58–86, National Academy Press, Washington, D.C., 1998; J. Higdon, Thiamin, in *An Evidence-Based Approach to Vitamins and Minerals: Health Benefits and Intake Recommendations*, pp. 33–38, Thieme, New York, 2003; D. Lonsdale, *A Nutritionist's Guide to the Clinical Use of Vitamin B-1*, 1988; V. Tanphaichitr, Thiamin, in M. E. Shils et al., (eds.), *Modern Nutrition in Health and Disease*, 9th ed., pp. 381–389, Williams and Wilkins, Baltimore, 1999.

Thick-film sensor

A sensor that is based on a thick-film circuit. Thick-film circuits are formed by the deposition of layers of special pastes onto an insulating substrate. The pastes are usually referred to as inks, although there is little resemblance to conventional ink. The printed pattern is fired in a manner akin to the production of pottery, to produce electrical pathways of a controlled resistance. Parts of a thick-film circuit can be made sensitive to strain or temperature. The thick-film pattern can include mounting positions for the insertion of conventional silicon devices, in which case the assembly is known as a thick-film hybrid. The process is relatively cheap, especially if large numbers of devices are produced, and the use of hybrid construction allows the sensor housing to include sophisticated signal conditioning circuits. These factors indicate that thick-film technology is likely to play an increasing role in sensor design.

The three main categories of thick-film inks are

conductors, dielectrics (insulators), and resistors. Conductors are used for interconnections, such as the wiring of bridge circuits. Dielectrics are used for coating conducting surfaces (such as steel) prior to laying down thick-film patterns, for constructing thick-film capacitors, and for insulating crossover points, where one conducting path traverses another. Resistor inks are the most interesting from the point of view of sensor design, since many thick-film materials are markedly piezoresistive.

Fabrication process. The main constituents of a thick-film ink are the binder (a glass frit), the vehicle (an organic solvent), and the active elements (metallic alloys or oxides). After printing, each layer of a thick-film pattern is dried to remove the organic solvents (the vehicle), which give the ink its viscosity. Drying also improves the adhesion properties, bonding the ink to its substrate and rendering the pattern immune to smudging. This stage is usually performed in a conventional oven at 100–150°C (212–302°F).

A final high-temperature firing is required to remove any remaining solvent and to sinter the binder and the active elements. During the firing cycle a thick-film pattern is raised to 500–1000°C (932–1832°F). The glass frit melts, wets the substrate, and forms a continuous matrix, which holds the functional elements. The heating and cooling gradients, the peak temperature, and the dwell time determine the firing profile. This has a critical effect on the production of a thick-film circuit, since it allows the electrical characteristics of the inks to be modified. Resistor materials are especially sensitive to the firing profile, and the resistor layer is usually therefore the last to be fired. However, the need for passivation of a circuit often necessitates covering it with a dielectric layer. To avoid changing the resistor values, a low-melting-point dielectric is often used for the final layer. *See* SINTERING.

The need for high-temperature firing can cause problems if thick-film materials are to be applied to previously heat-treated components. The temperatures used can adversely affect, for example, the properties of toughened or hardened steels.

Thick-film circuits and sensors are created by screen printing. This is essentially a stencil process, in which the printing ink is forced through the open areas of a mesh-reinforced screen onto the surface of a substrate. The screen stencils are formed by photolithography. In this process a photosensitive mesh-filling material is exposed to ultraviolet light through a mask depicting the required pattern. The image is photographically developed, and those parts of the pattern which have not been fixed are subsequently washed away.

The use of thick-film technology was originally introduced as a means of miniaturizing circuits without incurring the expense associated with fabrication in silicon. It was soon noted that thick-film materials had temperature- and stress-dependent properties. Although this was awkward from the point of view of circuit fabrication, it has since been turned to good account in sensor design. The piezoresistive properties of thick-film resistor inks can be used to form strain sensors. This approach is commonly used to manufacture pressure sensors and is exploited to produce accelerometers. *See* ACCELEROMETER; PRESSURE TRANSDUCER; STRAIN GAGE.

Piezoresistive sensors. The basis of most strain-based thick-film transducers is the piezoresistive effect. A piezoresistive material is one in which a change in electrical resistance occurs in response to changes in the applied stress.

Piezoresistive sensors are formed by placing stress-sensitive resistors on highly stressed parts of a suitable mechanical structure. The piezoresistive transducers are usually attached to cantilevers, or other beam configurations, and are connected in a Wheatstone bridge circuit. The beam may carry a seismic mass to form an accelerometer or may deform in response to an externally applied force. The stress variations in the transducer are converted into an electrical output, which is proportional to strain, by the piezoresistive effect. *See* WHEATSTONE BRIDGE.

Piezoresistive devices are relatively easy to construct, provide a low-frequency response extending down to dc (0 frequency), and work well over a relatively large temperature range (−50 to +150°C or −58 to 302°F). Another valuable feature is their ability to include signal-processing and communication functions within the sensor package at little extra cost.

The drawbacks of piezoresistive devices are that the output signal level is moderate (typically 100 mV full scale for a 10-V bridge excitation), the sensitivity can be temperature-dependent, and the usable bandwidth is not as large as that which may be obtained from a piezoelectric transducer.

Analysis of piezoresistance. If a rectilinear resistor has length l, width w, thickness t, and bulk resistivity ρ, its resistance R is given by Eq. (1). The gauge factor or strain sensitivity is defined as k in Eq. (2),

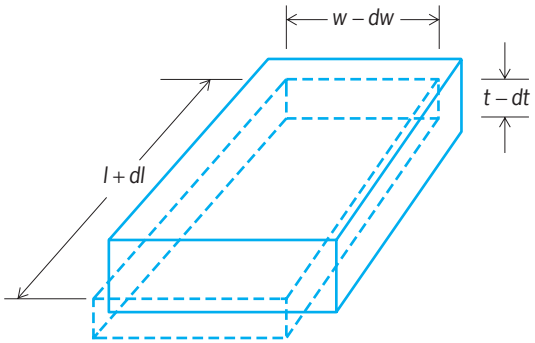
$$R = \frac{\rho l}{wt} \tag{1}$$

$$k = \frac{dR/R}{\varepsilon} \tag{2}$$

where ε is the relative change in length of the resistor (the strain) due to a stress, σ, applied to the substrate parallel to its length. The **illustration** shows the consequences of the applied stress. The length increases by an amount dl, while the width and thickness decrease by amounts dw and dt determined by Poisson's ratio v: $dw = vw\varepsilon$ and $dt = vt\varepsilon$. *See* ELASTICITY.

The original cross section is given by Eq. (3). Owing to the strain, the new cross-sectional area is given

$$A = wt \tag{3}$$



Changes in the dimensions of a rectangular resistor (length *l*, width *w*, and thickness *t*) due to a stress applied parallel to the length.

by Eq. (4). The term $v^2wt\varepsilon$ is very small compared

$$A' = (w - dw)(t - dt) = wt - 2vwt\varepsilon + v^2wt\varepsilon^2 \tag{4}$$

with the other two terms in Eq. (4) and can be neglected. Therefore, the change in cross-sectional area can be written as Eq. (5), given Eq. (6). Differentiating Eq. (1) gives Eq. (7), and hence the gauge factor k is given by Eq. (8).

$$A' - A = dA = -2v\varepsilon A \tag{5}$$

$$\frac{dA}{A} = -2v\varepsilon \tag{6}$$

$$\frac{dR}{R} = \frac{d\rho}{\rho} + \frac{dl}{l} - \frac{dA}{A} \tag{7}$$

$$k = \frac{d\rho/\rho}{\varepsilon} + (l + 2v) \tag{8}$$

Typically v will be between 0.2 and 0.3. Equation (8) therefore shows that the longitudinal gauge factor is a function of changes in both longitudinal resistivity and geometry. In conventional foil or wire strain gauges, the piezoresistive effects are negligible, and the variations in resistance are mainly a function of dimensional changes. For a foil gauge, k is approximately 2. For piezoresistive strain gauges, the first term in Eq. (8) is significant, and higher gauge factors (typically around 10) can be achieved, giving enhanced sensitivity. However, the resistivity of most piezoresistive materials is strongly temperature-dependent, and as a result piezoresistive strain gauges generally have a higher thermal sensitivity than other types.

Temperature sensors. The linear temperature coefficient of resistance possessed by certain platinum-containing conductive inks has allowed thermistors to be printed onto suitable substrates using thick-film fabrication techniques. Thick-film thermistors are very inexpensive and physically small, and have the further advantage of being more intimately bonded to the substrate than a discrete component. It has been shown that thick-film thermistors can have as good, if not better, performance than a comparable discrete component. *See* THERMISTOR.

Chemical sensors. Thick-film materials have been used for a number of chemical sensing applications, including the measurement of gas and liquid composition, acidity (pH), and humidity. It is difficult to define a comprehensive scheme for the classification of thick-film chemical sensors, since there are so many variants. However, a classification based on two categories seems to cover most devices: impedance-based transducers, in which the measurand causes a variation of resistance, capacitance, and so forth; and electrochemical systems, in which the sensed quantity causes a change in electrochemical potential or current. *See* ELECTRICAL IMPEDANCE; ELECTROCHEMISTRY.

Impedance-based chemical sensors. Thick-film gas sensors can be fabricated from a printable paste of semiconducting metal-oxide powder, inorganic additives, and organic binders. The paste is printed over an alumina substrate containing metal film electrodes and a back-heating resistor; the paste is then fired in an infrared or thermal belt furnace. The most frequently used material for this type of sensor is tin oxide (stannic oxide, SnO_2). *See* TIN.

Humidity sensors. Several thick-film humidity sensors have been reported in which capacitive effects are used. A common approach is to fabricate the sensor in the form of a pair of closely spaced interdigitated electrodes, which are screen-printed and fired onto an alumina substrate. The dielectric layer to be tested is then applied on top of this structure, and capacitance changes of up to 3 nanofarads are observed for porous coatings. *See* CAPACITANCE; TRANSDUCER.

John D. Turner

Bibliography. J. D. Turner and M. Hill, *Instrumentation for Engineers and Scientists*, Oxford University Press, 1999; N. M. White and J. D. Turner, Thick-film sensors: Past, present and future, *Meas. Sci. Technol.*, 8(1):1–20, January 1997.

Thickening

The production of a concentrated slurry from a dilute suspension of solid particles in a liquid. In practice, a thickener also usually generates a clear liquid; therefore clarification is generally a concurrent process. Thickening and clarification are outcomes of sedimentation, and both are representative of a group of industrial processes termed mechanical separations. *See* CLARIFICATION; SEDIMENTATION (INDUSTRY).

Operation. Although thickening may be carried out either batchwise or continuously, the latter method is more common. Special means are used to move the concentrated slurry to the discharge outlet in the mechanically agitated continuous thickener (**Fig. 1**). This equipment consists of a settling tank fitted with slow-moving rakes driven by a vertical control shaft. The tank may have a flat bottom or a shallow-cone bottom. Continuous thickeners may be large; diameters range from 20 to 300 ft (6 to 90 m) and the depth from 8 to 12 ft (2.5 to 4 m). Small tanks may be made of wood or steel, intermediate ones

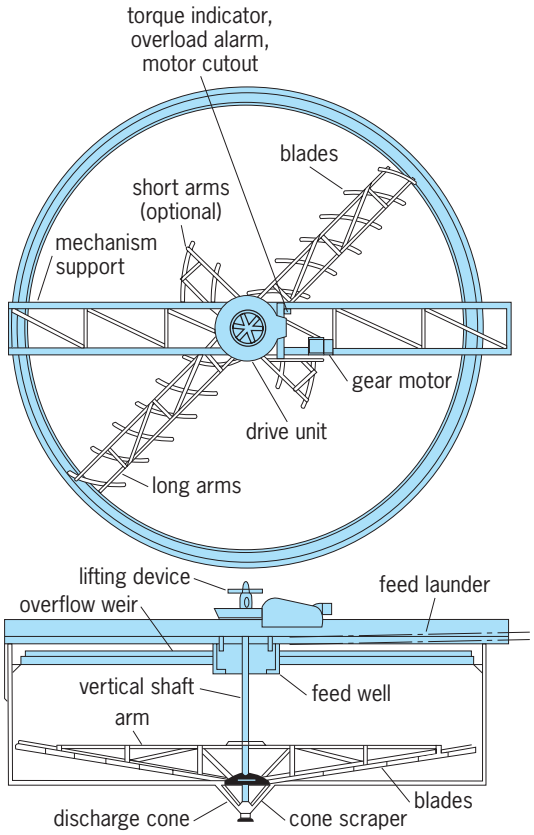


Fig. 1. Top and side views of mechanically agitated, continuous thickener. (*After W. L. McCabe, J. C. Smith, and P. Harriott, Unit Operations of Chemical Engineering, 4th ed., McGraw-Hill, 1985*)

of steel, and large units of concrete. In large thickeners, the rakes may rotate only once every 30 min.

In operation, dilute feed pulp is admitted continuously through a launder to a central well immersed to a depth of 2 to 3 ft (0.5 to 1 m) below the surface of the liquid in the tank. The liquid from the feed moves radially to the wall of the tank and overflows across a weir that bounds the periphery of the tank. As the liquid flows radially from the feed well, the solids it carries settle, so that the liquid is clear when it reaches the overflow weir. The solids settle to the bottom of the tank and form a concentrated slurry. The rakes, without repulping the solids into the liquid, gently agitate the solids, break up the flocs to aid the process of consolidation, and move the thickened solids to the discharge in the center of the tank bottom. From the discharge, the thickened slurry flows to the suction side of a sludge pump.

For a continuous thickener, the concentration of solids in the liquid decreases from the top to the bottom of the tank. There are two main zones, the free-settling and compression, which are separated by an interface (**Figs. 2** and **3**). The upper zone, in which clarification is accomplished, is free from solid in its top layers and supplies the clarified liquid overflow. Within the clarification zone, the solid concentration varies from zero to a low value at the interface between the zones. In the clarification zone, the solid particles are sufficiently far apart that free settling takes place. The bottom zone, in which thickening is accomplished, contains most of the inventory of solid in the tank. Here the concentration of solid changes rapidly from that in the clarification zone to that of the thickened slurry leaving the thickener. The process occurring in this zone is essentially compression.

To attain satisfactory capacities, the feed to a thickener is often flocculated. The performance of a given plant operating on a specified feed slurry depends largely on the major dimensions of the tank. To obtain a clear overflow, the upward velocity of the liquid in the clarification zone must be less than the minimum terminal settling velocity of the smallest particles. Then the capacity of the unit to handle clarified liquid is proportional to the horizontal cross-sectional area and, therefore, to the square of the diameter of the tank. The solid concentration in the underflow, and hence the degree of thickening achieved, depends on the time allowed for action in the compression zone. Once the feed rate of dilute slurry is fixed, the time for compression is proportional to the height of the compression zone. Thus

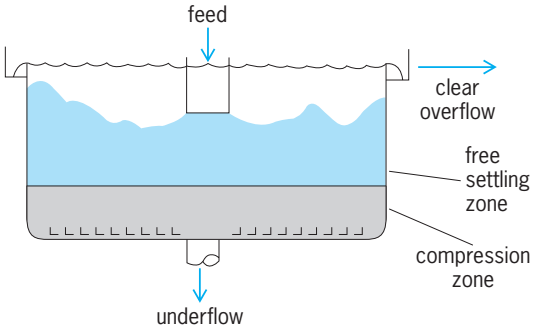


Fig. 2. Zones in continuous thickener. (*After W. L. McCabe and J. C. Smith, Unit Operations of Chemical Engineering, 2d ed., McGraw-Hill, 1967*)

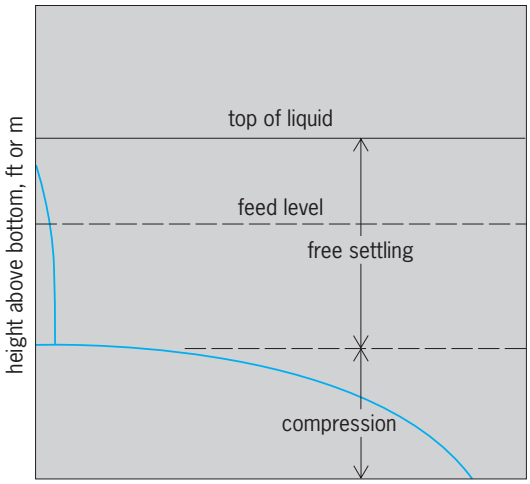


Fig. 3. Variation of solids concentration with the height in the tank of a continuous thickener. (*After W. L. McCabe, J. C. Smith, and P. Harriott, Unit Operations of Chemical Engineering, 4th ed., McGraw-Hill, 1985*)

the performance of the unit as a thickener is a function of tank depth.

Types. Thickeners are especially useful when large volumes of dilute slurries must be treated, as in manufacture of cement, production of magnesium from seawater, treatment of sewage, purification of water, treatment of coal, and dressing of metallurgical ores.

Several variants of the circular, continuous thickener are widely used. One type is fitted with multitray thickeners where flows are in series or parallel. This design reduces base area and also provides either more depth for consolidation of the thickened discharge or more overflow area for clarification. Another type, the National Coal Board deep-cone thickener, is used for processing coal and metallurgical ores. In the Swedish Lamella thickener, the liquid flows upward between inclined plates stacked closely together; the feed enters the stack from a side feed box. In theory the effective settling area is the sum of the horizontal projected areas of all the plates, but it is only about 50% efficient. The sludge can be further consolidated by vibration or raking. *See* MECHANICAL SEPARATION TECHNIQUES.

Vincent W. Uhl

Bibliography. M. C. Bustos et al., *Sedimentation and Thickening: Phenomenological Foundation and Mathematical Theory*, 1999; D. S. Jones, *Elements of Chemical Process Engineering*, 1996; W. McCabe, J. Smith, and P. Harriott, *Unit Operations of Chemical Engineering*, 7th ed., 2004; R. H. Perry and D. W. Green (eds.), *Perry's Chemical Engineers' Handbook*, 7th ed., 1997; J. F. Richardson, J. H. Harker, and J. Backhurst, *Chemical Engineering*, vol. 2, 5th ed., 2002; R. K. Sinnott, *Chemical Engineering Design*, vol. 6, 4th ed, 2005.

Thinner

A material used in paints and varnishes to adjust the consistency for application. Thinners are usually solvents for the vehicle used in the coating and are expected to evaporate after application. Because their only function is to make the application simple, it is important that their cost be low. Water is used as a thinner in emulsion paints and in certain water-soluble paints such as watercolors and calcimines.

Petroleum fractions are most commonly used for oil and resin coatings. The fraction boiling between 300 and 400°F (149 and 204°C), called mineral spirits, is most widely used. A lower-boiling and faster-evaporating solvent is called VM&P (varnish makers' and painters') naphtha. Still faster-evaporating materials are called petroleum ether, lacquer diluent, or rubber solvent. Stronger solvents contain substantial amounts of aromatic hydrocarbons and may be derived from petroleum or coal tar. These may be essentially pure materials, such as toluene or xylene, or mixtures designed to have the solvency and evaporation characteristics desired.

Since numerous coating resins are not sufficiently soluble in hydrocarbons, other materials or mixtures must be used. These include alcohols such as denatured ethyl or isopropyl alcohols for shellac, esters such as amyl acetate for nitrocellulose, and ketones and other compounds for acrylic and vinyl resins. Chlorinated hydrocarbons are used for some materials which are otherwise hard to dissolve, but toxicity limits their usefulness.

The selection of a thinner for a coating formulation depends upon the resins used, the application and curing conditions, and the effects desired. For example, fast-drying solvents will reduce the temperature of the surface, and under humid conditions they may cause moisture to condense on the surface, producing the phenomenon known as blushing. Vinyl, epoxy, and chlorinated rubber, among other types of coatings, commonly employ a mixture of thinners known as a solvent blend. These contain materials having different evaporation rates designed to hold the film open to avoid solvent entrapment and subsequent blistering, or to hasten or retard the drying rate as affected by atmospheric temperature, humidity, and wind conditions.

Historically, the thinner used for conventional paints was turpentine, but because of newer and cheaper solvents it has largely disappeared from paint manufacturing, although it is still used to some extent for thinning paints on the job. *See* PAINT AND COATINGS; SOLVENT; SURFACE COATING; VARNISH.

C. R. Martinson; C. W. Sisler

Thiocyanate

A compound containing the —SCN group, typically a salt or ester of thiocyanic acid (HSCN). Thiocyanates are bonded through the sulfur(s) and have the structure R—S—C≡N. They are isomeric with the isothiocyanates, R—N=C=S, which are the sulfur analogs of isocyanates (—NCO). The thiocyanates may be viewed as structural analogs of the cyanates (—OCN), where the oxygen (O) atom is replaced by a sulfur atom.

Thiocyanic acid is an unstable gas at room temperature and a yellow solid below 5°C (41°F). It is produced by the reaction of barium thiocyanates and sulfuric acid. The thiocyanates are stable to air and oxidation, and form a large number of metal complexes and organic compounds. For example, a test for the presence of ferric iron [Fe(III)] or thiocynate is the formation of the blood red complex $[Fe(SCN)_6]^{3-}$. The nucleophilic displacement of alkyl halides to produce alkyl thiocyanates is an example of the organic chemistry of the thiocyanate anion [reaction (1)]. Furthermore, like the cyanates, thio-

$$RCH_2Cl + SCN^- \longrightarrow RSCH_2SCN + Cl^- \qquad (1)$$

cyanates can be converted to thiocarbamates [reaction (2)].

$$RSCN + H_2O \longrightarrow RSCONH_2 \qquad (2)$$

The principal commercial derivatives of thiocyanic acid are ammonium and sodium thiocyanates.

Thiocyanates and isothiocyanates have been used as insecticides and herbicides. Specifically, ammonium thiocyanate is used as an intermediate in the manufacture of herbicides and as a stabilizing agent in photography. Sodium and potassium thiocyanates are used in the manufacture of textiles and the preparation of organic thiocyanates.

In living systems, thiocyanates are the product of the detoxification of cyanide ion (CN^-) by the action of 3-mercaptopyruvate sulfur transferase. In addition, thiocyanates can interfere with thyroxine synthesis in the thyroid gland and are part of a class known as goitrogenic compounds. *See* CYANIDE; SULFUR; THYROXINE.

Thomas J. Meade

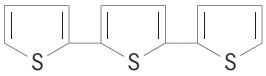
Thiophene

An organic heterocyclic compound containing a diunsaturated ring of four carbon atoms and one sulfur atom. *See* HETEROCYCLIC COMPOUNDS.

Thiophene (**1**), methylthiophenes, and other

(**1**)

alkylthiophenes are found in relatively small amounts in coal tar and petroleum. Thiophene accompanies benzene in the fractional distillation of coal tar. Purification of coal-tar benzene is effected by treatment with concentrated sulfuric acid, which selectively forms water-soluble thiophene-sulfonic acid. Alternatively, treatment with aluminum chloride selectively polymerizes the thiophene in the benzene to nonvolatile materials. 2.5-Dithienylthiophene (**2**)

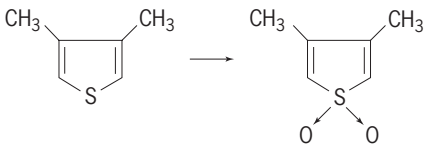


(**2**)

has been found in the marigold plant. Biotin, a water-soluble vitamin, is a tetrahydrothiophene derivative.

Properties. The parent compound (**1**) is nearly insoluble in water (forming 0.02–0.04% solutions at 20°C or 68°F), mp −38.2°C (−36.8°F), bp 84.2°C (183. 6°F), n_D^{20} 1.5287, and specific gravity (20/4) 1.0644. Thiophene has a resonance energy of 29–31 kcal/mole (121–130 kilojoules/mole), is stable to heat, and undergoes electrophilic substitutions (nitration, sulfonation, acetylation, halogenation, chloromethylation, and mercuration). Thus, thiophene is an aromatic compound. Generally, electrophilic substitutions occur with greater ease than with benzene, but less readily than with furan or pyrrole. The entering group favors the α position. Thiophenes are stable to alkali and other nucleophilic agents, and are relatively resistant to disruption by acid. *See* AROMATIC HYDROCARBON.

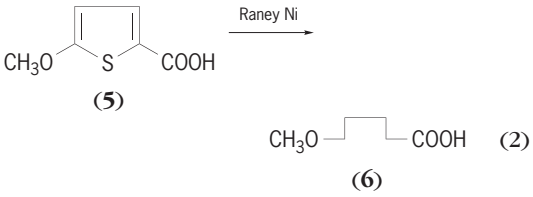
Most oxidative processes (nitric acid, ozone, hydrogen peroxide) involving the nucleus have not proved useful in opening the thiophene ring. Peracetic or perbenzoic acid oxidizes thiophenes such as 3,4-dimethylthiophene (**3**) to the corresponding sulfones (**4**), which behave more as butadiene derivatives than as thiophenes [reaction (1)].



(1)

(**3**) (**4**)

Sodium in liquid ammonia and methanol converts thiophene to a mixture of dihydro and acyclic products. Raney nickel strips sulfur from thiophenes in a ring-opening reaction (2), converting (**5**) to (**6**).



(**5**)

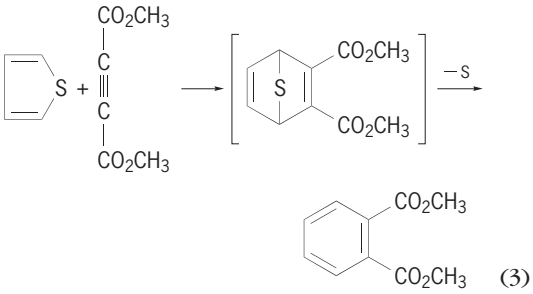
(2)

(**6**)

Catalytic hydrogenation over molybdenum or cobalt sulfide catalysts at high temperature and pressure, as well as over platinum or palladium catalysts in massive amounts, saturates the ring.

Bromine and chlorine react readily with thiophenes, which undergo both substitution and addition reactions. Control of conditions as well as the possibility of dehydrohalogenation by alkali of the products first formed furnishes halogenated thiophenes in practical preparations. Iodination of thiophene in the presence of mercuric oxide or iodination of mercurated thiophenes gives iodinated derivatives.

Thiophene undergoes the Diels-Alder reaction with the more active dienophiles, such as acetylenedicarboxylic ester, to form benzene derivatives by extrusion of sulfur [reaction (3)].



(3)

Preparation. The thiophene ring system is formed by cyclization of 1,4-dicarbonyl compounds in the presence of phosphorus sulfides (for example, 2,5-hexadione gives 2,5-dimethylthiophene; 4-oxo-3-ethylpentanoic acid gives 2-methyl-3-ethylthiophene), or by cyclization of hydrocarbons with sulfur or sulfur compounds at elevated temperatures [for example, the reaction of 2-methylbutadiene with sulfur at 320–420°C (610–790°F) gives 3-methylthiophene; the reaction of ethylbenzene with sulfur in a bimolecular process gives 2,4-diphenylthiophene].

The commercial production of thiophene (**1**) from readily available butane or butadiene awaits only a large-scale demand. A laboratory synthesis converts sodium succinate to thiophene by heating with phosphorus sulfide.

Alkylthiophenes are prepared by ring synthesis, by alkylation of thienylmagnesium halides with sulfate or sulfonate esters, or by reduction of thiophene ketones. 2-Vinylthiophene, potentially of interest as a polymerizable monomer, can be prepared by reducing 2-acetylthiophene to methyl-2-thienylcarbinol, and dehydrating.

Thiophene aldehydes are prepared by treatment of the thiophene with hexamethylenetetramine (Sommelet process), or with the *N*-methylformanilide-phosphorus oxychloride reagent pair. Friedel-Crafts acylation, often with mild catalysts, gives thiophene ketones in good yields. Thiophene carboxylic acids result from the silver oxide oxidation of thiophene aldehydes, the haloform oxidation of acetylthiophene, and the carbonation of thiophenemetal derivatives. Thiophene aldehydes, ketones, and acids show normal chemical behavior, similar to the corresponding benzene derivatives. *See* ORGANOSULFUR COMPOUND.

Walter J. Gensler; Martin Stiles

Bibliography. D. Barton and W. D. Ollis (eds.), *Comprehensive Organic Chemistry*, vol. 4, 1979; S. Gronowitz (ed.), *Chemistry of Heterocyclic Compounds: Thiophene and Its Derivatives*, vol. 44, 4 pts., 2d ed., 1992; A. R. Katritzkey (ed.), *Advances in Heterocyclic Chemistry*, vol. 1, 1963; E. Lukevics and A. E. Skorova, *Thiophene Derivatives of Group IV B Elements*, 1982.

Thiosulfate

A salt containing the negative ion $S_2O_3^{2-}$. This species is an important reducing agent and may be viewed as a structural analog of the sulfate ion (SO_4^{2-}) where one of the oxygen (O) atoms has been replaced by a sulfur (S) atom. The sulfur atoms of the thiosulfate ion are not equivalent. Thiosulfate is tetrahedral, and the central sulfur is in the formal oxidation state 6+ and the terminal sulfur is in the formal oxidation state 2−.

Commercial production of thiosulfates involves the boiling of elemental sulfur in an alkaline solution containing sulfite ions (SO_3^{2-}) [reaction (1)].

$$S + SO_3^{2-} \longrightarrow S_2O_3^{2-} \qquad (1)$$

Alternatively, sulfides may react with sulfur dioxide (SO_2), sulfite, or bisulfite (HSO_3^-) to produce thiosulfates. Thiosulfuric acid ($H_2S_2O_3$) is unstable and readily decomposes to elemental sulfur and sulfurous acid (H_2SO_3), while the sodium salt ($Na_2S_2O_3$) is stable.

Principal uses of thiosulfates include agricultural, photographic, and analytical applications. Ammonium thiosulfate [$(NH_4)_2S_2O_3$] is exploited for both the nitrogen and sulfur content, and it is combined with other nitrogen fertilizers such as urea. Thiosulfate ion is an excellent complexing agent for silver ions (bound through sulfur). The sodium salt and the ammonium salt are well known as the fixing agent "hypo" used in photography. The aqueous (aq) thiosulfate ion functions as a scavenger for unreacted solid (s) silver bromide on exposed film and therefore prevents further reaction with light [reaction (2)]. A

$$AgBr(s) + 2S_2O_3^{2-}(aq) \longrightarrow [Ag(S_2O_3)_2]^{3-}(aq) + Br^-(aq) \qquad (2)$$

well-known and important volumetric analysis is the rapid oxidation of thiosulfate by iodine to produce tetrathionate [reaction (3)].

$$2S_2O_3^{2-} + I_2 \longrightarrow S_4O_6^{2-} + 2I^- \qquad (3)$$

In nature, thiosulfate is converted into hydrogen sulfide (H_2S) via enzymatic reduction. Hydrogen sulfide, in turn, is converted into the thiol group of cysteine by the reaction with *O*-acetylserine. *See* COORDINATION COMPLEXES; OXIDATION-REDUCTION; PHOTOGRAPHIC MATERIALS; SULFUR.

Thomas J. Meade

Thirst and sodium appetite

The sensations caused by dehydration, the continuing loss of fluid through the skin and lungs and in the urine and feces while there is no water intake into the body. Thirst becomes more and more insistent as dehydration worsens. Water and electrolytes are needed to replace losses, and an adequate intake of sodium as well as water is important for maintaining blood volume. Herbivores and human vegetarians, whose diets lack sodium, have a natural appetite for sodium; however, severe sodium deficiency in carnivorous animals and humans can result in the development of a well-marked sodium appetite as well. Water intake varies considerably between individuals and depends on climate, custom, and diet. Reproduction affects drinking behavior; fluid intake increases during pregnancy and especially during lactation. Normally, the amounts of water drunk and taken in food are more than enough to maintain hydration of the body, and the usual mixed diet provides all the electrolytes required.

Drinking behavior. The passage of water through the mouth satisfies a basic urge to drink whether or not the body needs water. An indication of the importance of this urge is the fact that rats infused with water at rates far in excess of their requirements, by routes that bypass the mouth and pharynx, continue to drink. Responding to this urge when the body is in a healthy state is entirely beneficial because the water provides for body fluid homeostasis by the kidney and any surplus is excreted (although in kidney disease, normal drinking behavior may result in an inappropriately large and possibly harmful intake of fluid). Even more demanding than the basic urge to drink are the sensations of thirst and sodium appetite aroused by fluid deficits in the body, which lead to primary or regulatory drinking. The urgency of these

sensations, especially of thirst, indicates how vital drinking is for survival.

Classification. The three categories of drinking behavior are as follows:

Primary, regulatory, or deficit-induced drinking. A deficit of fluid in one or both of the major fluid compartments of the body is the signal to increase drinking.

1. Cellular dehydration, detected by osmoreceptors, causes thirst and vasopressin release.
2. Hypovolemia, detected by volume receptors in the heart and large veins and the arterial baroreceptors, causes immediate thirst, a delayed increase in sodium appetite, activation of the renin-angiotensin system, and increased mineralocorticoid and vasopressin secretion.

Secondary or nonregulatory drinking. Drinking occurs in the absence of a fluid deficit. Drinking associated with feeding, for instance, is important in the day-to-day control of body water. It ensures a regular supply of water so that renal regulation of body fluids can take place. Secondary drinking is distinguished by the following factors:

1. Mealtime or food-associated.
2. Fail-safe, normal.
3. Oropharyngeal cues, dry-mouth.
4. Schedule-induced.
5. Palatability-induced NaCl intake.

Clinical disturbances in drinking behavior. Increases or decreases in amounts drunk in disease may result from normal or abnormal functioning of mechanisms of thirst or sodium appetite:

1. Symptomatic: mechanisms of thirst and sodium appetite respond normally to excessive fluid loss caused by disease.
2. Pathologic: thirst mechanisms are activated excessively (hyperdipsia, polydipsia) or inadequately (hypodipsia) because of abnormalities of the mechanisms.

Physiology and ontogeny. Recent research on the physiology of drinking behavior has yielded extensive information on the mechanisms, neural substrates, and neuropharmacology of the different types of drinking behavior. Many mechanisms, some involving hormones, come into play to cause drinking, depending on the particular circumstances. The role of angiotensin II as a hormone and a central neurotransmitter or neuromodulator of thirst and sodium appetite is becoming clearer. The importance of the hormones of sodium deficiency, of pregnancy and lactation, and of the stress hormones of the hypothalamo-pituitary-adrenocortical axis in sodium appetite has been established. Information on the ontogeny of drinking behavior is now available. In much of this work, the newer specific receptor blockers—and more recently still, antisense oligonucleotides and gene-knockout animal models—have added new insights to theories based on classical behavioral and physiologic methods.

Oropharyngeal versus systemic factors. Theories of thirst in which dryness of the mouth and throat plays a central role are quite old. However, during the nineteenth century, Dupuytren, Latta, Magendie, and Bernard, among others, showed that thirst could be relieved by intravenous fluids and that the passage of water through the oropharynx was neither necessary nor sufficient for relief. Thirst was considered to be a sensation of general origin caused by lack of water in the body. Nevertheless, the dry-mouth theory received strong support from Cannon in the early twentieth century, and dryness of the mouth and throat (such as might be caused by lecturing, singing, mouth breathing, a spicy meal, or apprehension) certainly can cause increased drinking in the absence of any systemic need for water. However, since the early 1950s when B. Andersson and colleagues established that there is a thirst center in the hypothalamus, and as the significance of cellular dehydration and hypovolemia (low blood volume) as causes of thirst and sodium appetite began to be better understood, the dry-mouth theory has become less influential. It is now recognized that there are many other causes of increased drinking behavior, including cases in which there is an obvious systemic need for water as well as cases in which there does not appear to be an immediate need.

Cellular dehydration. H. Wettendorff first proposed the cellular dehydration theory of thirst at the beginning of the twentieth century. In 1937, A. Gilman showed that dogs drank more water after intravenous administration of hyperosmotic saline (a substance which dehydrates cells) than after the same osmotic load of hyperosmotic urea (a substance which penetrates cells), providing crucial supporting evidence for Wettendorff's theory. Observations of other species using a variety of osmotic challenges have extended these findings and have established that hyperosmotic solutions of solutes that are excluded from cells cause more drinking than equiosmolar amounts of solutes that penetrate cells. Thus, the osmotic shift of water out of the cells caused by the excluded solutes provides the critical stimulus to drinking. Continuing water loss in the absence of intake is perhaps a more significant cause of cellular dehydration than administration of an osmotic load, but the same mechanisms apply. *See* OSMOREGULATORY MECHANISMS; OSMOSIS.

Sharing in the overall cellular dehydration, whether caused by osmotic loads or water deprivation, are osmoreceptors which initiate the responses of thirst and renal conservation of water. There has also been considerable discussion about whether or not a sodium-sensitive system in the walls of the cerebral ventricles might account for some of the drinking induced by cellular dehydration, especially in herbivores. Osmoreceptors are mainly located in the hypothalamus. Injection of minute quantities of hyperosmotic solutions of dehydrating substances in the lateral preoptic area and adjacent zona incerta causes water-replete animals to drink, whereas injection of water, or bilateral destruction of the region, prevents the animal with generalized cellular dehydration from drinking. The nervous tissue in the hypothalamus surrounding the anterior third cerebral ventricle and, in particular, the vascular organ of the lamina terminalis also respond to osmotic

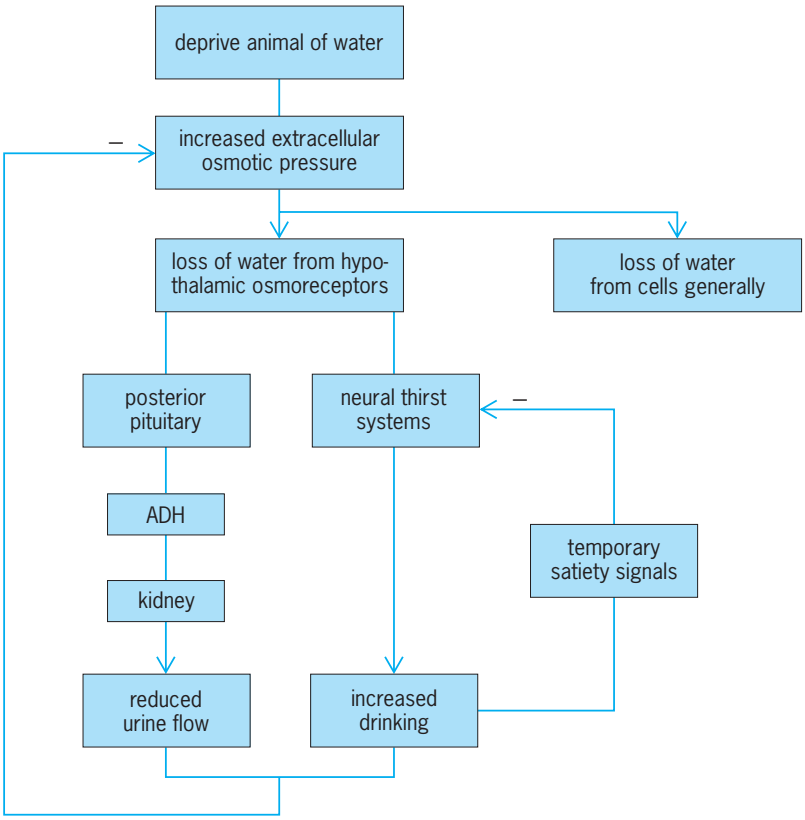


Fig. 1. Regulation of cellular water. ADH = antidiuretic hormone.

stimuli. Osmoreceptors initiating thirst work in conjunction with osmoreceptors initiating antidiuretic hormone (ADH) release to restore the cellular water to its prehydration level (**Fig. 1**). In addition to reducing urine loss, ADH may lower the threshold to the onset of drinking in response to cellular dehydration and other thirst stimuli. The cellular dehydration system is very sensitive, responding to changes in effective osmolality of 1–2%.

Hypovolemia. The cells of the body are bathed by sodium-rich extracellular fluid that corresponds to the aquatic environment of the unicellular organism. Thirst has been known since very early times to accompany hemorrhage, severe diarrhea (such as in cholera), sodium deficiency, and other diseases in which the brunt of fluid loss is borne by the extracellular fluid. Extracellular fluid consists of two main components of similar composition, blood plasma and interstitial fluid, which is the fluid outside the vasculature and in immediate contact with the cells. Loss of sodium is inevitably accompanied by loss of water, resulting in hypovolemia with thirst followed by a delayed increase in sodium appetite. If not corrected, continuing severe sodium loss eventually leads to circulatory collapse.

Among experimental procedures that have been used to cause hypovolemic drinking are (1) removing blood; (2) causing sodium deficiency by dieting, sweating, peritoneal dialysis, and other means; (3) causing extracellular fluid to accumulate outside the vasculature in the peritoneal cavity or subcutaneously by the technique of hyperoncotic dialysis; and (4) mimicking the effects of severe dehydration on the circulation by interfering with the flow of blood back to the heart by obstructing the abdominal inferior vena cava. All of these experimental methods cause thirst and a delayed increase in sodium appetite and either overhydrate the cellular compartment or do not alter its water content. In 1936, R. A. McCance described how three human subjects made sodium-deficient by dieting and sweating experienced a complex sensation of thirst and craving for salt. Initially, the developing sodium deficit and accompanying loss of water led to progressive hypovolemia with little change in osmotic pressure. Later, volume was preserved at the expense of osmotic pressure, which fell. Changes in drinking behavior were caused by hypovolemia, not by cellular dehydration, because the cells are usually overhydrated in sodium deficiency.

Stretch receptors in the walls of blood vessels entering and leaving the heart and in the heart itself are thought to initiate hypovolemic drinking. Volume receptors in the venoatrial junctions and receptors that register atrial and ventricular pressure respond to the underfilling of the circulation with a reduction in inhibitory nerve impulses to the thirst centers, which results in increased drinking. Angiotensin II and other hormones (such as aldosterone and ADH) are also involved in this response. Arterial baroreceptors function in much the same way as the volume receptors on the low-pressure side of the circulation, exerting continuous inhibitory tone on thirst neurons. A fall in blood pressure causes increased drinking, whereas an acute rise in blood pressure inhibits drinking. (It should be noted that the sustained rise in arterial blood pressure of established hypertension does not have any permanent inhibitory effect on drinking, presumably because the baroreceptors are reset to the higher arterial pressure.) The anterior third cerebral ventricle region, which is implicated in angiotensin-induced drinking, plays a crucial role in hypovolemic drinking, body fluid homeostasis, and blood pressure control. Some of the pathways involved are illustrated in **Fig. 2**.

Renin-angiotensin systems and drinking. It is believed that drinking caused by hypovolemic stimuli partly depends on the kidneys because research shows that (1) the amounts of water drunk after such stimuli are reduced by prior nephrectomy but not by ligation of both ureters and (2) extracts of kidney injected into water-replete animals cause drinking. The renal thirst factor is the proteolytic enzyme renin, which is secreted into the circulation by the juxtaglomerular cells of the kidney in response to hypovolemia. Renin cleaves an inactive decapeptide, angiotensin I, from angiotensinogen, an α_2-globulin that is synthesized in the liver and released into the circulation. Angiotensin I is converted to the physiologically active but short-lived octapeptide angiotensin II during the passage of blood through the lungs. Angiotensin II is an exceptionally powerful stimulus of drinking behavior in many

mammals, birds, reptiles, and bony fish when administered systemically or into the brain. Increased activation of the renin-angiotensin system may sometimes account for pathologically increased thirst in humans. Injection of angiotensin II into sensitive limbic structures in the anterior hypothalamus and the anterior third cerebral ventricle region of the brain of a rat causes an almost immediate increase in water intake, often followed by a slower increase in sodium intake. Angiotensin II also produces (1) a rise in arterial blood pressure by causing vasoconstriction and hypertrophy of vascular smooth muscle, release of norepinephrine from sympathetic nerve endings, and secretion of adrenomedullary hormones; and (2) water and sodium retention by causing release of ADH from the posterior pituitary and stimulation of renal tubulular transport of sodium through direct action on the kidney and indirectly through increased aldosterone secretion from the adrenal cortex. (Other less well-defined stimulating actions of angiotensin are on cell growth, membrane function, protein synthesis, prostaglandin release, learning, and memory.) *See* ALDOSTERONE; KIDNEY.

Renin, synthetic renin substrate, and angiotensin I are also effective stimuli to increase drinking, their action being mediated through local generation of angiotensin II. Angiotensin peptides that do not rely on renal renin for their production are produced in many tissues, including nervous tissue. The anterior third cerebral ventricle region, the vascular organ of the lamina terminalis, the median preoptic nucleus, and subfornical organ—which are particularly important in body fluid homeostasis and blood pressure control—are also well provided with angiotensinergic nerve terminals and receptors. The median preoptic nucleus and some other angiotensin-sensitive tissues lie inside the blood-brain barrier and presumably are not accessible to circulating angiotensin but are accessible to peptide released from nerve endings. There are also regions in the brain where production of angiotensin peptides takes place at a distance from receptors. This suggests that angiotensin produced in the brain may act both as a paracrine agent producing volume effects some distance from the point of release (because it diffuses into the extracellular fluid) and as a conventional point-to-point neurotransmitter. The role of cerebral renin and its relation to the more abundant and better-known renal renin remain uncertain.

Genetic aspects of renin-angiotensin systems are beginning to be investigated, and the genes encoding the proteins and receptors are being identified. Use of antagonists of the various stages of the renin-angiotensin cascade and exploitation of the more recently introduced selective antagonist subtypes and antisense oligonucleotides (a deoxyribonucleic acid or ribonucleic acid sequence with two or more covalently linked nucleotides) to block angiotensin synthesis or receptors have been invaluable in helping to understand the physiology of angiotensininduced drinking. At least three angiotensin receptor subtypes have been identified—AT_1, AT_2, and AT_4—

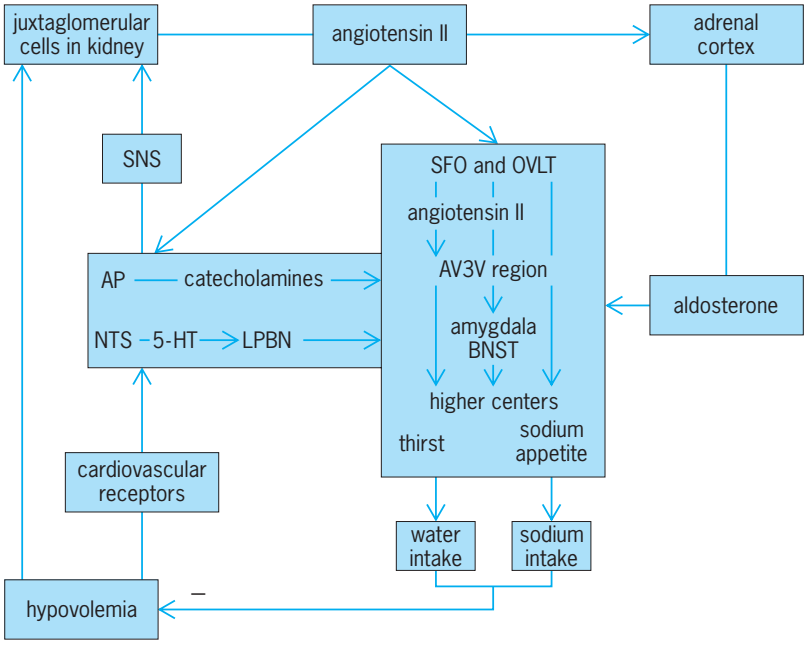


Fig. 2. Hypovolemic thirst and sodium appetite are caused by altered sensory information from an underfilled circulation, reinforced by increases in circulating angiotensin II and mineralocorticoids. Angiotensin peptides generated from components of a cerebral renin-angiotensin system are also involved. (The central nervous system pathways are based on A. K. Johnson and R. L. Thunhorst, 1997.) AP, area postrema; AV3V, anteroventral third ventricle; BNST, bed nucleus of the stria terminalis; 5-HT, 5-hydroxytryptamine or serotonin; LPBN, lateral parabrachial nucleus; OVLT, vascular organ of the lamina terminalis; SFO, subfornical organ; SNS, sympathetic nervous system.

but most known functions of angiotensin, including angiotensin-induced drinking, are associated with AT_1 receptors. However, the AT_2 receptors are the predominant receptors in the fetal brain. Although their function is uncertain, AT_1 and AT_2 receptors may have opposing effects on apoptosis (cell death), vasoconstriction, myoendothelial proliferation, and possibly on drinking behavior. Mutant mice lacking the gene encoding the AT_2 receptor show impaired drinking when water-deprived and an increased pressor response to intracarotid infusion of angiotensin II, indicating that the receptor could be involved in these functions. The potential of approaches such as the use of antisense oligonucleotides and gene-knockout animal models is considerable, though few results are yet available. *See* GENE; OLIGONUCLEOTIDE.

An integrated response to hypovolemia involves both neural and hormonal mechanisms (Fig. 2). The reduced sensory discharge from the cardiovascular stretch receptors resulting from underfilling of the circulation directly activates hypothalamic and limbic drinking systems. The same sensory signals lead to reflex release of renin through sympathetic nerves to the kidney. The resulting increases in circulating angiotensin II act on the subfornical organ, vascular organ of the lamina terminalis, and area postrema, which lie outside the blood-brain barrier, and contribute to the increases in thirst and sodium appetite and hemodynamic responses to hypovolemia by sensitizing hypothalamic and limbic drinking systems to the altered sensory information

from cardiovascular receptors. Angiotensin peptides generated inside the blood-brain barrier may also be involved, but their functional relation to the effects of increases in circulating angiotensin II is unknown. Increases in aldosterone secretion also contribute to the increases in sodium appetite. The renal renin-angiotensin system may have a more important emergency role in the arousal of thirst and sodium appetite in circulatory collapse (for example, in severe hemorrhage or adrenal insufficiency) than in the more modest day-to-day variations in extracellular fluid volume. *See* BRAIN; SYMPATHETIC NERVOUS SYSTEM.

Neuropharmacology of drinking. Many substances released by neurons, and in some cases by neuroglial cells, affect drinking behavior when injected into the brain and may interact with the brain and modify angiotensin-induced drinking. They can act as neurotransmitters and produce effects localized to the postsynaptic membrane or presynaptic endings close to where they are released, they may diffuse into the extracellular fluid and exert paracrine or volume effects on nervous structures some distance from their source, or they may be released into the bloodstream and function as hormones. The time course of action varies, and there may be long-term trophic actions. Substances may stimulate or inhibit drinking, or both, depending on the species and the conditions of the experiment. Acetylcholine is a particularly powerful stimulus to drink in rats, and no inhibitory effects on drinking have been described. Histamine also seems to be mainly stimulatory. However, a lengthening list of neuroactive substances, including norepinephrine, serotonin, nitric oxide, opioids, bombesin-like peptides, tachykinins, and neuropeptide Y, may either stimulate or inhibit drinking with varying degrees of effectiveness, depending on the species or the site of injection in the brain. Natriuretic peptides, prostaglandins, and gamma-amino butyric acid seem to be exclusively inhibitory. *See* ACETYLCHOLINE; NEUROBIOLOGY; SYNAPTIC TRANSMISSION.

Many hormones also affect water or sodium intake. Relaxin stimulates water intake, and ADH (or vasopressin) lowers the threshold to thirst in some species. Vasopressin injected into the third cerebral ventricle may stimulate water intake, suggesting a possible role for vasopressinergic neurons. Increased sodium appetite in pregnancy and lactation depends partly on the conjoint action of progesterone, estrogen, adrenocorticotrophic hormone (ACTH), cortisol, corticosterone, prolactin, and oxytocin. Aldosterone and other mineralocorticoids, the stress hormones of the hypothalamo-pituitary-adrenocortical axis, corticotrophin, ACTH, and the glucocorticoids also stimulate sodium intake. *See* ENDOCRINE MECHANISMS; NEUROHYPOPHYSIS HORMONE.

The effect of many of these substances on drinking behavior shows both species and anatomical diversity. Serotonin stimulates drinking in pigeons, but inhibits drinking in rats when injected into the lateral parabrachial nucleus on the pathway in the hindbrain that responds to overfilling of the circulation (Fig. 2). Tachykinins stimulate drinking in birds but inhibit thirst and sodium appetite in rats. Oxytocin released from the posterior pituitary into the bloodstream contributes to increased sodium appetite caused by the conjoint action of reproductive hormones, but as a neurotransmitter or paracrine agent released from oxytocinergic neurons in the paraventricular nucleus, it may inhibit sodium appetite in circumstances in which water is more urgently and immediately need. The multiplicity of effects of many of these substances makes it impossible to generalize on their role in natural thirst, but none of these substances seems to be as consistent and as universal a stimulus of increased thirst and sodium appetite as angiotensin.

Overview. Mechanisms that ensure a continuing intake of water and sodium are vital. There is a powerful urge to drink independent of need so that in temperate climates and under stable conditions of activity and diet, the body's need for water is fully met. In times of good health, such intake is beneficial, but it may result in overhydration during illness. The role of sodium is less clear. In most western societies, sodium intake in the diet exceeds need, which may be a factor in the increased incidence of hypertension. On the other hand, an adequate intake of sodium is essential, and palatability and increased sodium appetite in sodium deficiency ensure this. When the body lacks water or sodium, it is imperative that physiologic responses restore the normal contents. Renal conservation can slow the rate of fluid loss, but the emergency mechanisms of thirst and sodium appetite in response to deficits are far more important for survival.

James T. Fitzsimons

Bibliography. D. A. Denton, *The Hunger for Salt*, Springer-Verlag, Berlin, 1982; J. T. Fitzsimons, Physiology and pathophysiology of thirst and sodium appetite, in D. W. Seldin and G. Giebisch (eds.), *The Kidney*: *Physiology and Pathophysiology*, 3d ed., Lippincott, Williams and Wilkins, Philadelphia, 2000; A. K. Johnson and R. L. Thunhorst, The neuroendocrinology of thirst and salt appetite: Visceral sensory signals and mechanisms of central integration, *Frontiers Neuroendocrinol.*, 18:292–353, 1997; E. M. Stricker, *Handbook of Behavioral Neurobiology*, vol. 10: *Neurobiology of Food and Fluid Intake*, Plenum Press, New York, 1990.

Thomson effect

A phenomenon discovered in 1854 by William Thomson (Lord Kelvin). He found that there occurs a reversible transverse heat flow into or out of a conductor of a particular metal, the direction depending upon whether a longitudinal electric current flows from colder to warmer metal or from warmer to colder. Any temperature gradient previously existing in the conductor is thus modified if a current is

turned on. The Thomson effect does not occur in a current-carrying conductor which is initially at uniform temperature.

From these observations it may be shown that for copper there is a heat output where positive charge flows down a temperature gradient and a heat input where positive charge flows up a temperature gradient; whereas for iron the reverse is true. All metals may be divided into two classes with respect to the direction of the Thomson effect. These flows of heat require that a distributed seat of electromotive force act at all points in the conductor. The total Thomson emf along the length of a conductor is given by

$$\int_{T_1}^{T2} \sigma \, dT$$

where σ is the Thomson coefficient for the metal in question, and T_1 and T_2 are the temperatures at the two ends of the conductor. With the discovery of the Thomson effect a complete thermodynamical theory of thermoelectricity became possible. *See* THERMOELECTRICITY. John W. Stewart

Thoracica

The major order of the crustacean subclass Cirripedia. The adult animals are permanently attached. The mantle is usually reinforced by calcareous plates (**Fig. 1**). Six pairs of biramous cirri are present, and the abdomen is absent or represented by caudal appendages. Antennules are present in the adult, and cement glands are strongly developed. Most species are hermaphroditic.

Thoracica are subdivided into three suborders: Lepadomorpha, stalked or goose barnacles; Balanomorpha, the common acorn barnacles; and Verrucomorpha, a rare group of asymmetric barnacles. The stalked barnacles are attached by a peduncle, and the body is enclosed in a bivalved fold or mantle, the capitulum, which is typically strengthened by calcareous plates. In acorn barnacles the body is enclosed in a strong shell of four, six, or eight plates rigidly united or fused together, and the mantle opening is protected by a pair of movable opercular valves, each formed of two plates, the tergum and the scutum (Fig. 1*c*, *d*).

Barnacles feed by sweeping the fan of cirri (**Fig. 2**) through the water and straining out minute organisms. The mouth is furnished with a large upper lip or labrum and three pairs of mouth appendages: mandibles, maxillulae, and maxillae. The alimentary canal is divided into a fore-, mid-, and hindgut, with the midgut having the associated digestive diverticula. The anus is terminal. Excretory organs open on the maxillae. The nervous system is typically crustacean, though it is shortened.

The paired ovaries lie in the stalk or in the mantle in the stalkless acorn barnacles, with the oviduct opening at the base of cirrus I. The testes lie in the thorax, and the paired seminal vesicles run backward to the base of the penis. Cross-fertilization is usual, with the sperm being deposited within the mantle by the elongated penis of an adjacent barnacle. Self-fertilization can occur. The eggs are laid as two flat coherent masses into the mantle space, where they hatch into nauplius larvae. The nauplii are planktonic and after passing through five further similar floating naupliar stages, they change into the very different cypris, which ceases to be pelagic and seeks a suitable substrate for attachment and metamorphosis into the adult. Growth is rapid, especially in warm waters, and the young barnacle may be sexually mature in 3 weeks, though occasionally fertilization may not occur until 1 or even 2 years after settlement. Barnacles may live

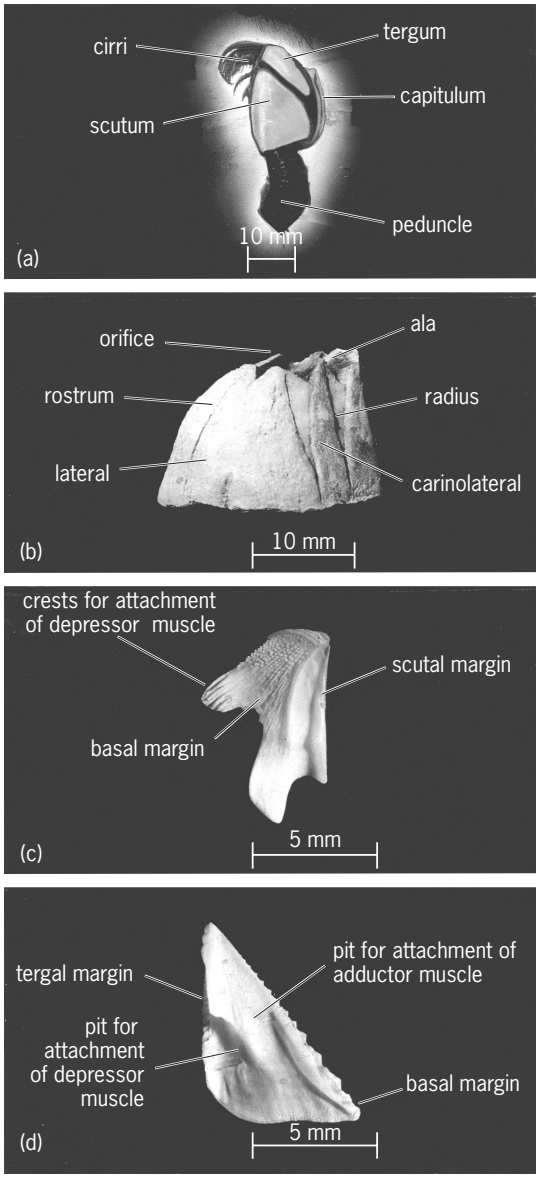


Fig. 1. Morphology of representative Thoracica. (*a*) *Lepas anatifera* (*from D. P. Henry, The Cirripedia of Puget Sound with a key to the species, Univ. Wash. Publ. Oceanogr., 4(1):1–48, 1940*). (*b*) *Balanus eburneus*, lateral view of shell, (*c*) inner view of tergum, and (*d*) inner view of scutum (*from D. P. Henry, American Waters, Friday Harbor Symposium in Marine Biology, University of Washington Press, 1959*).

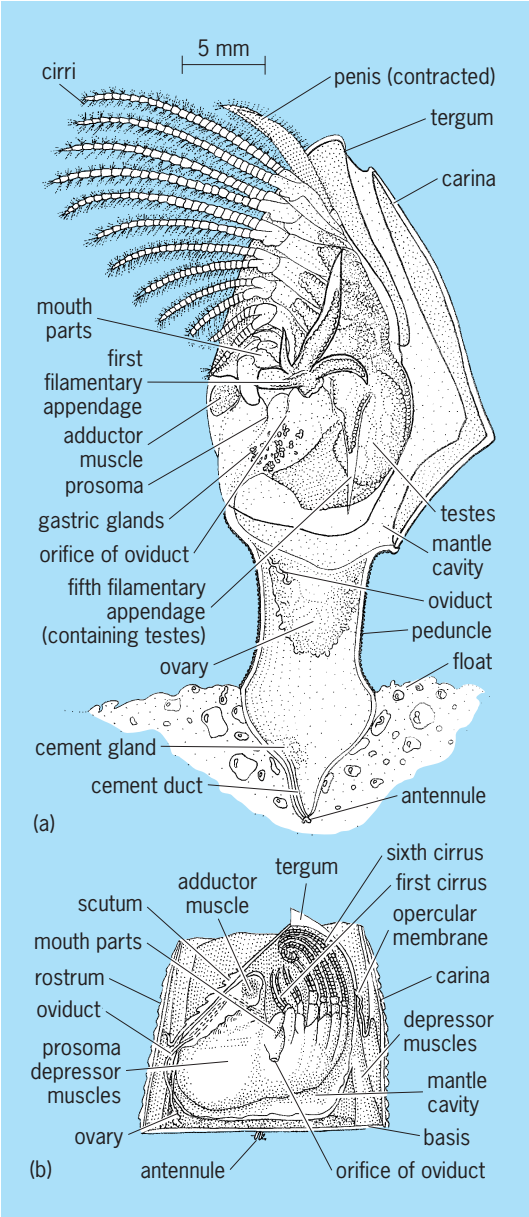


Fig. 2. Thoracica, internal anatomy. (*a*) *Lepas fascicularis*, with right side of capitulum, peduncle, and float removed. (*b*) *Balanus*, with right side of wall removed. (*After R. W. Hegner, Invertebrate Zoology, Macmillan, 1933*)

only months or occasionally as long as 10 years in some slow-growing and deep-water species. *See* CIRRIPEDIA; LEPADOMORPHA.

H. G. Stubbings

Thorite

A mineral, thorium silicate, in which the element thorium was discovered in 1828. Thorite is tetragonal in crystallization and has a crystal structure identical with that of the nesosilicate zircon, $ZrSiO_4$. The idealized chemical formula of thorite is $ThSiO_4$. All natural material departs widely from this composition owing to the partial substitution of uranium, rare earths, calcium, and iron for thorium. Structurally, thorite usually has completely lost its crystallinity because of radiation damage from the contained uranium and thorium (metamict state). The specific gravity ranges between about 4.3 and 5.4. The hardness on Mohs scale is about $4^1/_2$. The color commonly is brownish yellow to brownish black and black. Thorogummite is a chemical variant of thorite with the same crystal structure and very similar properties. It is deficient in silica and contains small amounts of OH in substitution for oxygen.

Thorite occurs chiefly in pegmatites. It also occurs as an accessory mineral in black sands and other detrital deposits derived from granitic or gneissic terrains. Vein deposits containing thorite and thorogummite associated with barite and fluorite occur in the Wet Mountains, Custer and Fremont counties, Colorado. Similar deposits occur in the Lemhi Pass district in Idaho and Montana. A vein deposit of monazite containing thorium is mined at Steenkampskraal near Van Rhynsdorp, Cape Province, South Africa. *See* METAMICT STATE; RADIOACTIVE MINERALS; SILICATE MINERALS; THORIUM.

Clifford Frondel

Bibliography. C. Frondel, *Systematic Mineralogy of Uranium and Thorium*, USGS Bull. 1064, 1958; J. W. Frondel, M. Fleischer, and R. S. Jones, *Glossary of Uranium and Thorium-Bearing Minerals*, USGS Bull. 1250, 1967; E. W. Heinrich, *Mineralogy and Geology of Radioactive Raw Materials*, 1958.

Thorium

A chemical element, Th, atomic number 90. Thorium is a member of the actinide series of elements. It is radioactive with a half-life of about 1.4×10^{10} years. *See* PERIODIC TABLE.

1																	18
1 H	2											13	14	15	16	17	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Thorium oxide compounds are used in the production of incandescent gas mantles. Thorium oxide has also been incorporated in tungsten metal, which is used for electric light filaments. It is employed in catalysts for the promotion of certain organic chemical reactions and has special uses as a high-temperature ceramic material. The metal or its oxide is employed in some electronic tubes, photocells,

and special welding electrodes. Thorium has important applications as an alloying agent in some structural metals. Perhaps the major use for thorium metal, outside the nuclear field, is in magnesium technology. Thorium can be converted in a nuclear reactor to uranium-233, an atomic fuel. The energy available from the world's supply of thorium has been estimated as greater than the energy available from all of the world's uranium, coal, and oil combined.

Monazite, the most common and commercially most important thorium-bearing mineral, is widely distributed in nature. Monazite is chiefly obtained as a sand, which is separated from other sands by physical or mechanical means. *See* MONAZITE.

Thorium has an atomic weight of 232. The temperature at which pure thorium melts is not known with certainty; it is thought to be about 1750°C (3182°F). Good-quality thorium metal is relatively soft and ductile. It can be shaped readily by any of the ordinary metal-forming operations. The massive metal is silvery in color, but it tarnishes on long exposure to the atmosphere; finely divided thorium has a tendency to be pyrophoric in air.

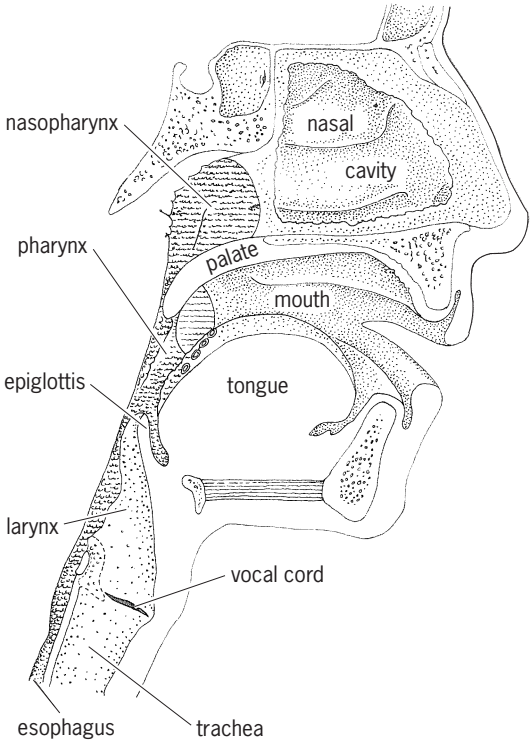
All of the nonmetallic elements, except the rare gases, form binary compounds with thorium. With minor exceptions, thorium exhibits a valence of 4+ in all of its salts. Chemically, it has some resemblance to zirconium and hafnium. The most common soluble compound of thorium is the nitrate which, as generally prepared, appears to have the formula $Th(NO_3)_4 \cdot 4H_2O$. The common oxide of thorium is ThO_2, thoria. Thorium combines with halogens to form a variety of salts. Thorium sulfate can be obtained in the anhydrous form or as a number of hydrates. Thorium carbonates, phosphates, iodates, chlorates, chromates, molybdates, and other inorganic salts of thorium are well known. Thorium also forms salts with many organic acids, of which the water-insoluble oxalate, $Th(C_2O_4)_2 \cdot 6H_2O$, is important in preparing pure compounds of thorium. *See* ACTINIDE ELEMENTS; RADIOACTIVITY.

Harley A. Wilhelm

Bibliography. P. W. Atkins et al., *Inorganic Chemistry*, 4th ed., 2006; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; S. Cotton, *Lanthanide and Actinide Chemistry*, 2d ed., 2006.

Throat

The region that includes the pharynx, the larynx, and related structures. Both the nasal passages and the oral cavity open into the pharynx, which also contains the openings of the Eustachian tubes from the ears (see **illus.**). The lower (inferior) portion of the pharynx leads into the esophagus and the trachea or windpipe. The rather funnel-shaped pharynx is suspended from the base of the skull and the jaws; it is surrounded by three constrictor muscles that function primarily in swallowing. *See* EAR (VERTEBRATE); PHARYNX.



Sagittal section of a human head showing the structure of the pharynx and associated structures. (*After J. C. B. Grant, A Method of Anatomy, Williams and Wilkins, 1952*)

The larynx, or voice box, is marked externally by the shield-shaped thyroid cartilage which forms the Adam's apple. The larynx contains the vocal cords that act as sphincters for air regulation and permit phonation. The lower end of the larynx is continuous with the trachea, a tube composed of cartilaginous rings and supporting tissues. *See* LARYNX.

The term throat is also used in a general sense to denote the front (ventral side) of the neck.

Thomas S. Parsons

Thrombosis

The process of forming a thrombus, which is a solid mass or plug in the living heart or vessels composed of the constituents of the blood. Thrombosis usually occurs in a diseased blood vessel, as a result of arteriosclerosis. The consequences of thrombosis include local obstruction causing both tissue death and hemorrhage. Thrombosis is a significant factor in the death of an individual affected by arteriosclerotic cardiovascular disease, malignancy, and infection. *See* HEMORRHAGE; INFARCTION.

Thrombus formation. Thrombosis is usually initiated by vascular damage and consequent platelet adhesion and clumping. The vascular endothelium plays an important role in protecting against thrombus formation. Normal endothelium releases nitric oxide, a vasodilator. Prostacyclin, a platelet aggregation inhibitor, releases heparinlike anticoagulants and protein inhibitors such as thrombomodulin.

Fig. 1. Organized thrombus in an artery with recanalization.

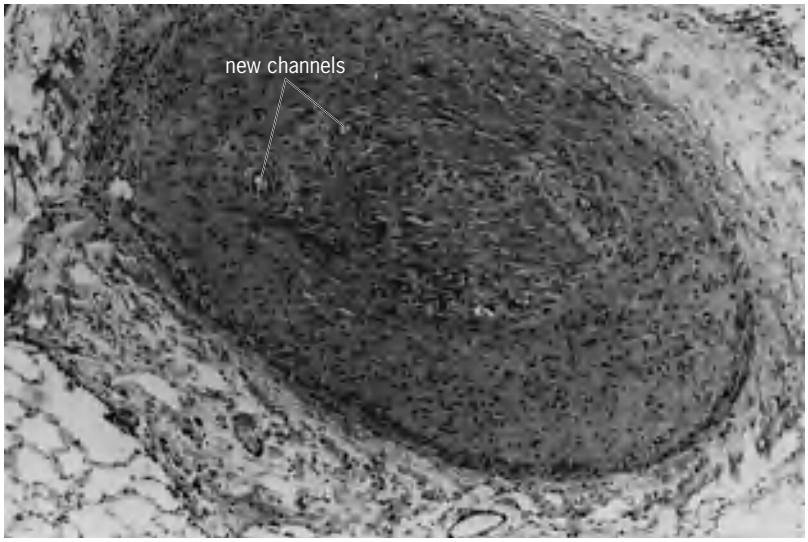


Fig. 2. Organized thrombus in an artery with scar formation; the artery is completely occluded by the fibrotic mass.

Injury to tissues, adequate concentrations of the coagulation elements, and stasis of the blood may all play a role in furthering the process of thrombus formation. The process is initiated when platelets specifically adhere to the subendothelial collagen at the points of damage to the endothelium. At the same time that the platelets begin to aggregate and release products that will further promote thrombus formation, the protein factors of the blood, which help to form the insoluble meshwork of the thrombus, become activated. This latter process is known as blood coagulation. The proteins of the coagulation system, through a series of cascading reactions (the intrinsic and extrinsic pathways), eventually reach a final common pathway to form fibrin, the insoluble protein that forms the scaffolding of the thrombus. As blood flows by the thrombus, more platelets and fibrin are deposited. Red blood cells and white blood cells become entrapped in the thrombus and are integrated into its structure. The thrombus thus consists of alternating zones of platelets and columns of fibrin with irregular layers of red cells and leukocytes.

Location of thrombi. Thrombi may form in arteries, in veins, on heart valves, or on the surfaces of the chambers of the heart. In arteries, which are vessels with relatively rapid flow of blood, the thrombus is predominantly composed of fibrin and platelets, a so-called white thrombus. In veins, vessels with less rapid flow, the thrombus is made up of greater portions of red cells, a so-called red thrombus. In capillaries, a thrombus usually consists of only platelets and fibrin.

Fate of thrombi. Once a thrombus forms, it can have one of four fates. (1) It may be digested, destroyed, and removed by proteolytic enzymes of the plasminogen–plasmin system. (2) If factors favor continued thrombus formation, the thrombus may continue to propagate itself and eventually occlude the vessel. This propagation usually takes place downstream to the site of original thrombus. (3) It may give rise to an embolus. Fresh portions of the thrombus at its outer surface tend to be friable and may break off, giving rise to thromboemboli. These emboli may cause tissue damage at sites distant from the origin of the thrombus. For example, lung infarction may develop from venous emboli, usually arising in the legs, and stroke may occur from thrombi that form in the heart and travel to the brain. (4) Thrombi may undergo a process known as organization. In this process, fibroblasts and capillaries proliferate and grow into the thrombus at its attachment to the vessel wall. The mass may be converted to a mass of vascularized connective tissue; the new channels may reestablish blood flow through the vessel (**Fig. 1**) to some degree. Organization helps stabilize the thrombus, and it may result in incorporation of a contracted fibrous mass into the vessel wall (**Fig. 2**). *See* EMBOLISM.

Prevention and treatment. Maintaining good blood flow (especially in the veins) helps prevent thrombosis. Exercise, support hose, and proper leg elevation when sitting will decrease the tendency of venous stasis in the legs. Treating hypertension and hypercholesterolemia retards atherosclerosis, which is a major cause of arterial thrombosis. Agents that interfere with platelet function, such as aspirin and fish oils, may help avoid thrombotic episodes. Anticoagulants prevent the formation of fibrin and may also be used to prevent thrombosis, especially in the veins or on prosthetic heart valves. If treatment can be given in the early stages of thrombosis, fibrinolytic therapy, utilizing agents that will help form plasmin, can minimize the tissue damage caused by thrombosis. *See* ARTERIOSCLEROSIS; CIRCULATION DISORDERS; PHLEBITIS.

Irwin Nash; Romeo A. Vidone

Bibliography. R. Colman et al., *Hemostasis and Thrombosis*, 4th ed., 2000; J. M. Kissane, *Anderson's Pathology*, 9th ed., 1989; J. L. Robbins et al., *Basic Pathology*, 6th ed., 2000; W. Williams et al., *Hematology*, 5th ed., 1996.

Thrust

The force that propels an aerospace vehicle or marine craft. Thrust is a vector quantity. Its magnitude is usually given in newtons (N) in International System (SI) units or pounds-force (lbf) in U.S. Customary Units. A newton is defined as 1 kilogram mass times an acceleration of 1 meter per second squared. One newton equals approximately 0.2248 lbf. *See* FORCE; UNITS OF MEASUREMENT.

The thrust power of a vehicle is the thrust times the velocity of the vehicle. It is expressed in joules (J) per second or watts (W) in SI units. In U.S. Customary Units thrust power is expressed in foot-pounds per second, which can be converted to horsepower by dividing by 550. *See* TURBINE ENGINE SUBSYSTEMS; POWER; RECIPROCATING AIRCRAFT ENGINE; ROCKET; TURBOJET.

J. Preston Layton

Thulium

A chemical element, Tm, atomic number 69, atomic weight 168.934. It is a rare metallic element belonging to the rare-earth group. The stable isotope ^{169}Tm makes up 100% of the naturally occurring element. *See* PERIODIC TABLE.

1												13	14	15	16	17	18
1 H	2																2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

The salts of thulium possess a pale green color and the solutions have a slight greenish tint. The metal has a high vapor pressure at the melting point. When ^{169}Tm is irradiated in a nuclear reactor, $_{170}$Tm is formed. The isotope then emits strongly an 84-keV x-ray, and this material is useful in making small portable x-ray units for medical use. *See* RARE-EARTH ELEMENTS.

Frank H. Spedding

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; K. A. Gschneidner Jr., J.-C. Bünzli, and V. K. Pecharsky (eds.), *Handbook on the Physics and Chemistry of Rare Earths*, 2005.

Thunder

The acoustic radiation produced by thermal lightning channel processes. The lightning return stroke is a high surge of electric current (~20,000 A) that occurs when the lightning flash makes contact with the Earth. The current surge has a very short duration, depositing approximately 95% of its electrical energy during the first 20 microseconds with its peak power occurring at 2 μs. Spectroscopic studies have shown that the lightning channel is heated to temperatures in the 20,000–30,000 K (36,000–54,000°F) range by this process. The lightning channel at this time has a diameter of approximately 1 cm (0.4 in.), and the pressure of the hot channel exceeds 10 atm (10^6 pascals). The hot, high-pressure channel expands supersonically and reaches a radius of 5 cm (2 in.) within the 20-μs period during which it is being heated. The channel continues expanding and forms a shock wave as it pushes against the surrounding air. Because of the momentum gained in expanding, the shock wave overshoots, causing the pressure in the core of the channel to go below atmospheric pressure temporarily. The outward-propagating wave separates from the core of the channel, forming an N-shaped wave that eventually decays into an acoustic wavelet. *See* SHOCK WAVE; STORM ELECTRICITY.

If the lightning channel were a perfectly straight line, the thunder produced by a single return stroke would consist of a single acoustic wavelet; the sound would be similar to that of a passing supersonic aircraft. However, real lightning channels are very crooked or tortuous. The formation of a single acoustic wavelet as described above applies, therefore, not to the channel as a whole but to the many, approximately straight segments of the lightning channel that locally behave as line sources of shock waves. The sound that is eventually heard or detected, thunder, is the sum of many individual acoustic pulses, each a remnant of a shock wave, that have propagated to the point of observation from the generating channel segments. The first sounds arrive from the nearest part of the lightning channel and the last sounds from the most distant parts.

To explain the amplitude variability in thunder, it is necessary to examine the radiation pattern from a channel segment. Laboratory experiments with sparks have shown that 80% of the acoustic energy from a line source is confined to a zone within ±30° of a plane perpendicular to and bisecting the line source. This means that the collection of thunder pulses from a long section of the channel perpendicular to the observer's line of sight (line of sound, actually) will all have higher amplitudes than the collection of pulses arriving from a long channel section more nearly parallel to the line of sound. Hence, the claps of high-amplitude thunder and the rumbles of low-amplitude thunder are each associated with the orientation of lightning channel segments with respect to the observer's location. It follows from this reasoning that the sound of thunder is unique to the shape of each lightning channel and to the location of the observer in relation to the lightning channel.

Another complicating aspect of thunder is the temporal structure of lightning. The description of thunder given above represents acoustics of a

single stroke of lightning. A lightning flash may produce several return strokes. In addition to the return strokes, there are lightning leader processes that also produce acoustic signals; however, they are thought to be much less energetic. For a complete description of thunder, the acoustic signals of all of these sources must finally be added together.

A further complicating aspect is the propagation path of the thunder pulses in the atmosphere. An acoustic ray is the path followed by a narrow beam of acoustic signals. In the atmosphere, which has temperature gradients and wind shears, the acoustic rays are bent into curved paths. Because of the decrease of air temperature with height, rays are most frequently bent upward. Therefore, observers on the surface may not hear all of the thunder from a visible lightning flash. Specifically, the higher the source of the sound, the farther it can be heard. Frequently, the thunder that is heard originates in the cloud and not in the visible channel. On some occasions, the observer may hear no thunder at all; this is more frequent at night when lightning can be seen over long distances and thunder can be heard only over a limited range (~10 km or 6 mi). *See* LIGHTNING; THUNDERSTORM.

Arthur A. Few

Bibliography. A. A. Few, Thunder, *Sci. Amer.*, 233(1):80–90, 1975; National Research Council Staff (eds.), *The Earth's Electrical Environment*, 1986; H. S. Ribner and D. Roy, Acoustics of thunder: A quasilinear model for tortuous lightning, *J. Acous. Soc. Amer.*, 72:1911–1925, 1982; M. A. Uman, *The Lightning Discharge*, 1987, reprint 2001.

Thunderstorm

A convective storm accompanied by lightning and thunder and a variety of weather such as locally heavy rainshowers, hail, high winds, sudden temperature changes, and occasionally tornadoes. The characteristic cloud is the cumulonimbus or thunderhead, a towering cloud, generally with an anvil-shaped top. A host of accessory clouds, some attached and some detached from the main cloud, are often observed in conjunction with cumulonimbus. The height of a cumulonimbus base above the ground ranges from 1000 to over 10,000 ft (300 to 3000 m), depending on the relative humidity of air near the Earth's surface. Tops usually reach 30,000–60,000 ft (9000–18,000 m), with the taller storms occurring in the tropics or during summer in midlatitudes. Thunderstorms travel at speeds from near zero to 70 mi/h (30 m/s). In many tropical and temperate regions, thunderstorms furnish much of the annual rainfall. *See* LIGHTNING; THUNDER.

Development. Thunderstorms are manifestations of convective overturning of deep layers in the atmosphere and occur in environments in which the decrease of temperature with height (lapse rate) is sufficiently large to be conditionally unstable and the air at low levels is moist. In such an atmosphere, a rising air parcel, given sufficient lift, becomes saturated and cools less rapidly than it would if it remained unsaturated because the released latent heat of condensation partly counteracts the expansional cooling. The rising parcel reaches levels where it is warmer (by perhaps as much as 18°F or 10°C over continents) and less dense than its surroundings, and buoyancy forces accelerate the parcel upward. The convection may be initiated by a combination of surface heating, cooling of the upper layers of the troposphere, forced ascent of moist low-level air over rising terrain or at fronts and other boundaries (for example, drylines, cold-air outflow boundaries from other thunderstorms) or in gravity waves, and forced lifting of air by upper-air weather disturbances. The rising parcel is decelerated and its vertical ascent arrested at altitudes where the lapse rate is stable, and the parcel becomes denser than its environment. The forecasting of thunderstorms thus hinges on the identification of regions where the lapse rate is unstable, low-level air parcels contain adequate moisture, and surface heating or uplift of the air is expected to be sufficient to initiate convection. *See* FRONT.

Occurrence. Thunderstorms are most frequent in the tropics, and rare poleward of 60° latitude. In the United States, the Florida peninsula has the maximum activity with 60 thunderstorm days (days on which thunder is heard at a given observation station) per year. Thunderstorms occur at all hours of day and night, but are most common during late afternoon because of the diurnal influence of surface heating. The weak nighttime maximum of thunderstorms in the Mississippi Valley of the central United States is still a topic of debate.

Structure. Radar is used to detect thunderstorms at ranges up to 250 mi (400 km) from the observing site. Much of present-day knowledge of thunderstorm structure has been deduced from radar studies, supplemented by visual observations from the ground and satellites, and in-place measurements from aircraft, surface observing stations, and weather balloons. *See* METEOROLOGICAL INSTRUMENTATION; RADAR METEOROLOGY; SATELLITE METEOROLOGY.

Thunderstorms occur in isolation, in chaotic patterns over wide areas, in the eye walls and spiral bands of hurricanes, in clusters within large-scale weather systems, and in squall lines perhaps several hundred miles long. An individual thunderstorm typically covers a surface area of 10–400 mi^2 (25–1000 km^2) and consists of one or more distinct cells, each of which is several miles across, lasts about an hour, and undergoes a characteristic life cycle. In the cumulus or growing stage, a cell consists primarily of updrafts (vertical speeds of 20–110 mi/h or 10–50 m/s) with precipitation suspended aloft; in the mature stage, updrafts and downdrafts coexist and heavy rain falls to the ground; in the dissipating stage, a cell contains weakly subsiding air and only light precipitation. During the mature stage, downdrafts may reach 35 mi/h (15 m/s). The downdraft air is denser than its surroundings due to evaporational cooling, which occurs as clear air is entrained into the cloud from outside, and is forced downward by gravitational pull and by the drag of falling

precipitation. The downflowing air spreads outward in all directions as it nears the surface, and forms a cold, gusty wind that is directed away from the precipitation area. This advancing cold air may provide the necessary lift in neighboring warm moist air for the formation of new updraft cells. Intense, narrow downdrafts (sometimes from innocuous-looking cumulonimbi) produce locally damaging, divergent outflow winds as they impact the ground. These small-scale flow features, known as microbursts, contain large wind shears that are extremely hazardous to low-flying aircraft. *See* HURRICANE; PRECIPITATION (METEOROLOGY); SQUALL LINE.

In an environment where the winds increase and veer with height, and midlevel air is dry enough to provide the potential for strong downdrafts, a thunderstorm may become organized so as to maintain a nearly steady state for hours. In such a strong vertical shear of the horizontal wind, the updraft is tilted so that precipitation falls out of the updraft instead of through it, and updraft and downdraft can coexist for several hours in the configuration shown in **Fig. 1**. A long-lived storm in a sheared environment may consist of a single intense cell (supercell) or of many cells with an organized growth of new cells on one side of the storm (generally, the southwest in the Northern Hemisphere) and decay of old cells on the opposite flank. *See* WIND.

Severe storms. Thunderstorms are considered severe when they produce winds greater than 58 mi/h (26 m/s or 50 knots), hail larger than $\frac{3}{4}$ in. (19 mm) in diameter, or tornadoes. While thunderstorms are generally beneficial because of their needed rains (except for occasional flash floods), severe storms have the capacity of inflicting utter devastation over narrow swaths of the countryside. The greatest frequency of severe storms occurs in the Great Plains region of the United States during the spring, but only a small percent of the thunderstorms are severe. Severe storms are most frequently supercells which form in environments with high convective instability and moderate-to-large vertical wind shears. The supercell may be an isolated storm or part of a squall line.

Since severe storms constitute a hazard to aircraft, their internal dynamics has been deduced largely from radar measurements. Doppler radar is specialized to measure the velocity of radar targets parallel to the radar beam, in addition to the intensity of precipitation. Doppler radar studies and analysis of surface pressure falls have shown that large hail, high winds, and tornadoes often develop from a rotating thunderstorm cell known as a mesocyclone. Large hail, high winds, and weak tornadoes may form from nonrotating (on broad scale) multicellular storms, but are less likely. Maximum tangential winds around the typical mesocyclone are roughly 50 mi/h (20 m/s) and are located in a circular band which is 1–3 mi (1.5–4.5 km) in radius. A surface pressure deficit of several millibars exists at the mesocyclone center. In one case, a pressure drop of 34 mbar (3.4 kilopascals) was measured. Identification of a mesocyclone signature on radar has been

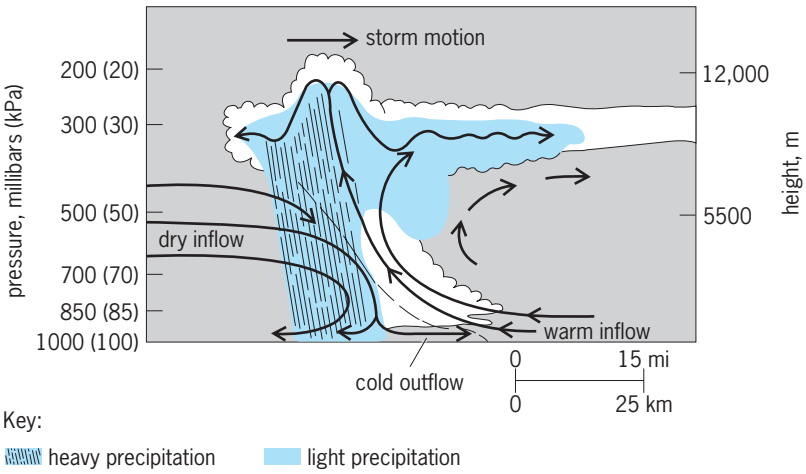


Fig. 1. Cloud boundaries and simplified circulation (arrows denote flow) of a typical mature thunderstorm in winds which blow from left to right and increase with height. Vertical scale has been exaggerated fivefold compared with the horizontal scale. 1 m = 3.3 ft; 1 km = 0.6 mi.

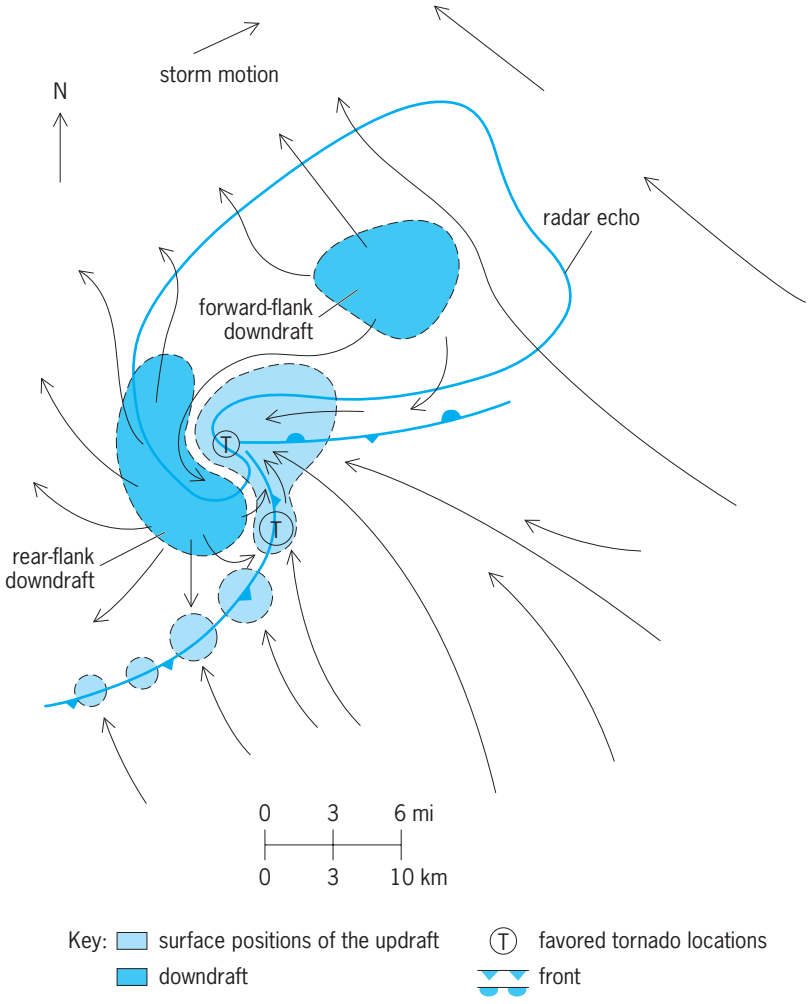


Fig. 2. Schematic surface plan view of a tornadic thunderstorm. The gust front is the boundary between unmodified warm, moist, inflowing air and rain-cooled, outflowing air. Arrows depict flow streamlines relative to the storm. The northern T at the mesocyclone center (wave apex) indicates where a major tornado is most likely. The southern T indicates the favored place for new mesocycline and tornado development. For storms in the Southern Hemisphere, transpose north and south. (*After R. Davies-Jones, Tornado dynamics, in E. Kessler, ed., Thunderstorm Morphology and Dynamics, 2d ed., University of Oklahoma Press, 1986*)

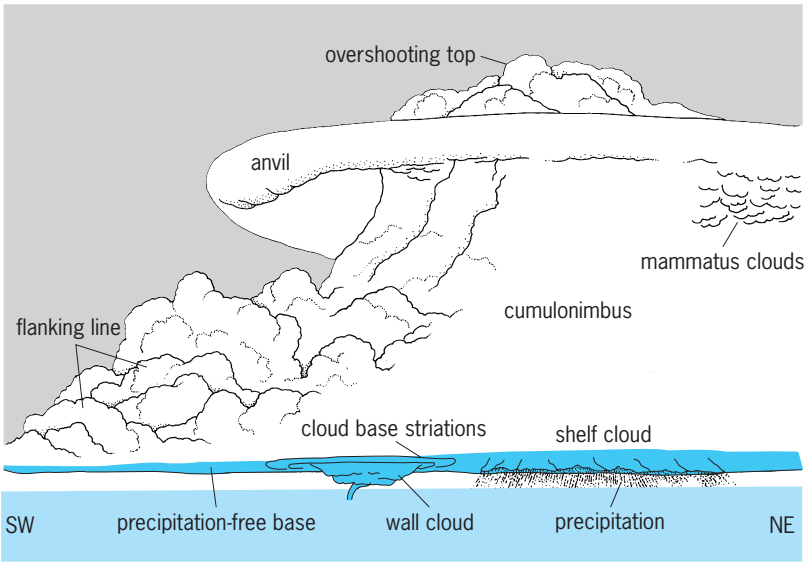


Fig. 3. Composite view of a typical tornado producing cumulonimbus as seen from a southeasterly direction. Horizontal scale is compressed, and all the features shown could not be seen from a single location. (*NOAA picture by C. Doswell and B. Dirham*)

used to issue severe weather warnings. The structure of a supercell storm is shown in **Fig. 2**.

On conventional radar displays, hook-shaped appendages to echoes are also good indications of mesocyclones, but unfortunately a large percentage of tornadic storms never exhibit such a hook. A mesocyclone sometimes is recognizable visually by rotation of a wall cloud, a discrete and distinct lowering of the cumulonimbus base (**Fig. 3**). The wall cloud is often seen visually to be rotating as an entity. The wall cloud is frequently the seat of intense vertical motions at low levels. The initial rotation of the mesocyclone at midlevels of the storm stems from the vertical shear of the winds flowing into the storm at low levels. An example of this effect is what happens to an initially vertical line drawn by a skywriter; changing wind speed (direction) with height causes the line to turn about a horizontal axis perpendicular (parallel) to the wind. The physically relevant winds are the ones relative to to the storm since it is the updraft that converts horizontal vorticity (spin) into vertical vorticity. In the extreme case when the storm-relative winds are constant in speed but veer (turn anticyclonically) with height, air parcels flowing into the storm have streamwise vorticity (that is, they spin about their direction of motion). As the parcels flow into the updraft, their spin axes are tipped upward, like a propeller on an aircraft entering a climb, and so the updraft rotates as a whole.

If the storm-relative winds back instead of veer with height, the vorticity is antistreamwise, resulting in anticyclonically rotating updrafts. In storm environments (on a rotating Earth), ground-relative winds generally veer and increase with height owing to effects of friction and flow of warmer air into the region. Generally the storm motion is such that the storm-relative winds also veer. Thus the majority of updrafts rotate cyclonically.

When the storm-relative winds increase with height without changing direction, air parcels spin like wheels normal to their direction of travel in the storm's reference frame. As the parcels flow into the updraft, their spin axes are tilted toward the vertical, resulting in cyclonic and anticyclonic spin on the right and left sides, respectively. Thus, the two halves of the updraft rotate in different directions in this case. In strong shear, the initial storm splits into two supercells that move to the right and left of the original storm motion. The cyclonic and anticyclonic halves of the initial updraft become the updrafts of the supercells. These updrafts continue to rotate as a whole after the split because net streamwise and antistreamwise vorticity is present in the inflows of the right- and left-moving supercells, respectively, owing to the deviate motions.

Once vertical vorticity has been generated, it can be amplified by the vertical stretching of air parcels in the lower part of the updraft where the flow converges horizontally. This process is analogous to a spinning ice skater. The development of rotation at the ground is a more complicated process, involving thermal generation of vorticity and the storm's downdraft as well as its updraft. *See* TORNADO; VORTEX.

Attempts have been made to modify thunderstorms to increase areal rainfall and suppress hail. The results of such experiments have been inconclusive. *See* HAIL; SQUALL; STORM DETECTION; WEATHER MODIFICATION.

For further discussion of storm electricity *see* ATMOSPHERIC ELECTRICITY; STORM ELECTRICITY.

Robert Davies-Jones

Bibliography. D. Atlas (ed.), *Severe Local Storms*, Amer. Meteorol. Soc., Meteorol. Monogr., vol. 5, no. 27, 1963; E. Kessler (ed.), *Instruments and Techniques for Thunderstorm Observation and Analysis*, 2d ed., 1988; E. Kessler (ed.), *The Thunderstorm in Human Affairs*, 2d ed., 1983; E. Kessler (ed.), *Thunderstorm Morphology and Dynamics*, 2d ed., 1992; F. H. Ludlam, *Clouds and Storms*, 1980.

Thyme

Any of a large and diverse group of plants in the genus *Thymus* utilized for their essential oil and leaves in both cooking and medicine.

Hundreds of different forms, or ecotypes, of thyme are found in the Mediterranean area, where thyme occurs as a wild plant. *Thymus vulgaris*, generally considered to be the true thyme, is the most widely used and cultivated species. Both "French" and "German" thyme are varieties of this species. Other species, such as *T. serpyllum* (mother-of-thyme) and *T. herba-barona* (caraway-scented), are grown or used in much the same way as *T. vulgaris*, but they do not have the same flavor or scent characteristics. Most types of thyme are low-growing perennials that do not exceed 15 in. (38 cm) in height. Typically, small smooth-edged leaves to 0.4 in. (1 cm) long are closely spaced on stems that become woody with age. Depending on soil type and climate, thyme may

live 10 years or longer, though in culture thyme is replanted every 5 years or less.

Cultivation of thyme is similar to that of rosemary. Much of the thyme seed available is harvested from wild plants. Seeds from a large number of forms are often mixed together. Though seed may be suitable or even desirable for small plantings, it is often unsuitable for large-scale cultivation. Rooted cuttings are usually employed for farm-size plantings, thus avoiding the variability of seed. Once established, thyme requires little care, since it has few pests or associated diseases. Verticillium wilt and *Rhizoctonia solani* have however, been reported to infect thyme.

Wild European thyme, the source of much imported material, is usually harvested only once a year, while cultivated plants in the United States are harvested mechanically up to three times a year. As with most herbs, both stems and leaves are harvested and then dehydrated. Dried stems and leaves are separated mechanically. Thyme oil is extracted from fresh material.

Thyme is a widely used herb, both alone and in blends such as "fines herbes." Thyme oil is used for flavoring medicines and has strong bactericidal properties. *See* SPICE AND FLAVORING.

Seth Kirby

Thymosin

A polypeptide hormone synthesized and secreted by the endodermally derived reticular cells of the thymus gland.

Action. Thymosin exerts its actions in several loci: (1) in the thymus gland, either on precursor stem cells derived from fetal liver or from bone marrow, or on immature thymocytes, and (2) in peripheral sites, on either thymic-derived lymphoid cells or on precursor stem cells. The precursor stem cells, which are immunologically incompetent whether in the thymus or in peripheral sites, have been designated as predetermined T cells or T_0 cells, and mature through stages termed T_1 and T_2, each reflecting varying degrees of immunological competence. Thymosin promotes or accelerates the maturation of T_0 cells to T_1 cells as well as to the final stage of a T_2. In addition to this maturation influence, the hormone also increases the number of total lymphoid cells by accelerating the rate of proliferation of both immature and mature lymphocytes. Thus, both the numbers and state of immunological maturity of the lymphocyte, one of the prime classes of cells contributing to host immunity, are under the influence of thymosin. It is evident, therefore, that the thymus gland and its hormone, thymosin, play an essential role in the development and maintenance of host immunological competence. *See* IMMUNITY.

Host immunological competence. This phenomenon can be broadly divided into two types, namely humoral immunity and cell-mediated immunity. In the former, T cells participate cooperatively with lymphoid cells derived from bone marrow (B cells) in the synthesis of humoral circulating antibody as a response to the presence of a foreign soluble or insoluble antigen. In cell-mediated immunity, the immunologically competent T cell functions in phenomena based upon cell-mediated immunological responsivity of the host, for example, ability to recognize and reject foreign tissue or organ transplants or grafts, resistance to infections, particularly of a viral or fungal nature, and responsivity to foreign allergens, that is, normal or adequate responsivity in the delayed hypersensitivity reaction. It is also recognized that additional classes of T cells may exert other specialized functions. Thus, the ability to recognize "self," as contrasted with "nonself" or cells of foreign origin, resides apparently in a class of specialized T cells that have been termed suppressor T cells. *See* CELLULAR IMMUNOLOGY.

The actions of thymosin are the basis of its roles, and that of the thymus, in the regulation of the structure and functioning of host lymphoid tissue, including parameters of immunological competence.

Bioassay. There are several assay methods for thymosin; two are discussed here. One is done with cultures of sheep erythrocytes and mouse spleen cells; the other is done by injecting lymphoid cells into animals.

Spontaneous rosette cell-forming assay. This assay for thymosin activity is based upon the adherence of sheep erythrocytes to mouse spleen cells because of recognition by the spleen cells of a foreign antigen on the sheep erythrocytes. This results in four to six sheep erythrocytes adhering to the perimeter of each individual lymphoid cell, giving a "rosette" appearance to the latter. The numbers of spontaneous rosette-forming cells (SRFCs) formed in cell cultures by spleen cells of normal mice are inhibited by a concentration of azathioprine approximately one-tenth of that required to decrease the number of SRFCs in spleen cells of adult thymectomized mice. However, incubation of the latter cells with thymosin in cultures restores the sensitivity of these cells to the inhibitory action of azathioprine. Thus, it is possible to assay the potency of a particular thymosin preparation by titrating to the minimum quantity of thymosin required to produce a 50% inhibition of SRFCs numbers in the presence of a quantity of azathioprine that is inhibitory for spleen cells from a normal mouse. This assay is highly sensitive and has been utilized to detect picogram quantities of thymosin as well as to assess blood levels of circulating thymosin and the half-life (rate of disappearance from the blood) of intravenously injected thymosin.

GVH assay. One assay for thymosin is based upon the graft-versus-host (GVH) reaction. In this assay, lymphoid cells, usually spleen cells, of an animal (for example, mouse or rat) are injected into an animal of the same species but of differing histocompatibility and immunologically suppressed. This will result in an "attack" by the injected cells upon the spleen of the recipient animal, inducing a marked proliferative response with a significant splenic enlargement (splenomegaly) evident within 5 to 7 days after the injection. The magnitude of the splenomegaly is dependent upon the number of spleen cells

administered and the status of the immunological competence of the injected animal. Thus, injection of allogenic spleen cells into a mouse rendered immunologically incompetent will not cause splenomegaly in the recipient animal. In contrast, administration of thymosin to the immunologically incompetent animal will restore to a degree, dependent upon the amount and period of thymosin pretreatment, such competence and the animal will be able to respond to allogenic spleen cell injection with a resultant splenomegaly, or GVH reaction.

Purified preparations. Utilizing the above assays to assess biological activity, it has been possible, by classical methods of protein fractionation and purification, to obtain thymosin from fresh calf thymus glands in a purified, physically and chemically homogeneous form. A highly purified preparation of thymosin has also been obtained from human thymic tissue, and thymosinlike activity has been demonstrated in thymic tissue of the mouse, rat, guinea pig, and hog, as well as in human, mouse, and pig serum.

Physical and chemical studies of the homogeneous preparation of thymosin from calf thymus glands have established that the hormone belongs to the class of polypeptide hormones. Purified bovine thymosin has a molecular weight of $12{,}500 \pm 200$, consists of 108 amino acid residues, and is free of carbohydrate and lipid. No unusual amino acids are present in the molecule. Both calf and human thymosin are weakly antigenic and, when administered to rabbits under appropriate experimental conditions, stimulate antibody production. The antibody-containing serum has been utilized as an experimental tool for counteracting endogenous circulating thymosin, thereby producing an immunologically depressed or incompetent animal. Antibody-containing serum has also been used for the development of a radioimmune assay for assessing circulating blood levels of thymosin. *See* PROTEIN; RADIOIMMUNOASSAY.

Blood levels. The radioimmune assay and the spontaneous rosette cell-forming assay applied to human serum have revealed that thymosin blood levels are maximal at birth, with the concentration maintained at this level until approximately age 20. At this time, thymosin concentration in the blood begins to decline slowly and is nondetectable beyond 45 years of age. Preliminary data for serum thymosin concentrations in persons suspected of having a lower than normal immunological responsivity indicate that blood thymosin levels may be lower than normal in such individuals when compared with healthy individuals of the same age group.

Verifying hormone activity. One of the basic tenets for designating a substance as a hormone of an endocrine gland, such as the thymus, requires evidence that administration of the preparation to animals deprived of the gland in question will prevent the development of deleterious consequences, that is, the hormone can function in the absence of the gland. To provide such evidence for the thymus, a variety of laboratory animals were subjected to one of two experimental procedures: surgical removal of the thymus in the early neonatal period (within 24 to 48 h following birth); or thymectomy of the adult animal followed by involution or destruction of the major lymphoid organs either by exposure to whole-body x-radiation or by injection of an immunosuppressive drug such as cyclophosphamide or Imuran. *See* HORMONE.

Either of these experimental approaches results primarily in a loss of host immunological competence. This is reflected in failure of normal growth and development, a depressed or total loss of capacity to respond to foreign antigens by synthesizing antibodies, the inability to reject a foreign organ or tissue transplant, and a lack of resistance to infectious agents. Thus, the experimental animal exhibits depressed ability to function in the two major divisions of immunity: humoral immunity and cell-mediated immunity. *See* IMMUNOGLOBULIN; TRANSPLANTATION BIOLOGY.

Thymosin has been shown to function in the absence of the thymus gland. The effects of neonatal thymectomy on immunological capability can be prevented from occurring by treatment of the thymectomized animal with daily injections of thymosin preparations. These replacement therapeutic studies with thymosin, and other experimental approaches, have established that the thymus is an endocrine gland and that thymosin is a hormone, synthesized and secreted by the thymus and exerting its influence both within the thymus and in extrathymic loci, notably in other lymphoid organs (discrete lymph nodes) as well as the spleen. **Tables 1** and **2** list some of the demonstrated biological activities of thymosin in two test systems.

Immunodeficiency disorders. At least 21 well-recognized antibody and cellular immunodeficiencies have been described in humans. Certain of these disorders, characterized as primary immunological deficiency diseases, are directly related to the absence or hypoplasia of the thymus gland at birth. In others,

TABLE 1. Biological activities of thymosin in cell culture models

Activity	System
Conversion of precursor cells to immunologically competent lymphocytes	Rosette assay Cytotoxicity assay: expression of σ and TL antigens Responsivity to mitogens Mixed lymphocyte interaction Primary and secondary antibody response Conversion of cultured bone marrow cells into cells reactive in the graft versus host assay T and B cell cooperation (incubation in cell cultures, cells tested in animals)
Inhibition	Autosensitization of human lymphocytes Autologous rosette formation

TABLE 2. Biological activities of thymosin in animal models

Model	Activity
Normal mice	Lymphocytopoiesis Enhanced rate of allograft rejection Enhanced resistance to progressive growth of Moloney virus-induced sarcoma Enhanced mixed lymphocyte reaction (thymosin injected into animal, cells tested in culture) Enhanced lymphoid cell response to mitogens (thymosin injected into animal, cells tested in culture) Enhanced antibody synthesis
Germ-free mice	Lymphocytopoiesis
Adrenalectomized mice	Enhanced lymphocytopoiesis
Neonatally thymectomized mice	Increased survival and rate of growth Lymphocytopoiesis Restoration of ability to reject skin allograft Enhancement of mixed lymphocyte interaction Stimulation of antibody synthesis
Adult thymectomized mice	Restoration of sensitivity of spleen cells to azathioprine Inhibition of autologous rosette cell formation
Athymic "nude" mice	Lymphocytopoiesis Restoration of response of lymphoid cells to mitogen Reduction of allogeneic and xenogeneic tumor growth rate
Immunosuppressed mice	Enhancement of ability to reject skin allograft
NZB mice*	Delay of appearance of abnormal thymocyte differentiation, with loss of suppressor function

*Animals developing spontaneous autoimmune disease.

immune disorders may be associated with thymic malfunction. Studies indicate that aberrant or perhaps inadequate production of thymosin, or other thymic factors, may be a significant etiological factor in a variety of primary immune disorders, such as autoimmune diseases, in certain malignancies, as well as in resistance to oncogenic viruses and to many fungal and mycobacterial pathogens.

In selected numbers of immune disorders in which the deficiency is one of cell-mediated immunity, as for example, the DiGeorge syndrome, the number of circulating lymphocytes may be normal, or only slightly lower than normal. However, the maturation of the lymphoid cells has been arrested due to a failure of normal thymic development, including probable lack of functioning of cells responsible for the synthesis and secretion of thymosin and, perhaps, other thymic factors. Studies in selected immunodeficient individuals injected with a partially purified preparation of calf thymosin suggest that the hormone may induce the maturation of lymphoid cells of an immature or early T type to T_1 and T_2 cells with resultant enhancement of host immunological competence, as reflected, for example, in responsivity in tests for delayed hypersensitivity. *See* ENDOCRINE SYSTEM (VERTEBRATE); IMMUNOLOGICAL DEFICIENCY; THYMUS GLAND.

Abraham White

Bibliography. J. Kuby, *Immunology*, 3d ed., 1997; D. O. Norris, *Vertebrate Endocrinology*, 3d ed., 1996; E. R. Stiehm and V. A. Fulginiti, *Immunological Disorders in Infants and Children*, 5th ed., 2001; D. Van Bekkum and A. Kruisbeek (eds.), *Biological Activity of Thymic Hormones*, 1975; A. White and A. L. Goldstein, The endocrine role of the thymus, and its hormone, thymosin, in the regulation of the growth and maturation of host immunological competence, *Advan. Metabol. Dis.*, 8:361–376, 1975.

Thymus gland

An important central lymphoid organ in the neck or upper thorax of all vertebrates from elasmobranchs to mammals. The most primitive representatives of vertebrates which have been shown to possess this organ are the cyclostomes *Eptatretus stoutii* (California hagfish) and *Petromyzon marinus* (sea lamprey).

Embryologically, the thymus gland arises as an endodermal outgrowth from the pharyngeal portion of the alimentary canal and is invested by the surrounding mesodermal tissue, which gives rise to its connective tissue elements and blood vessels. The organ is later invaded by additional mesodermal blood-borne stem cells from the blood islands or bone marrow, and these differentiate into the precursor cells for the thymus-derived lymphocytes (thymocytes and T lymphocytes). The thymus gland is most prominent during early life. In many laboratory species of mammals and in humans it reaches its greatest relative weight at the time of birth, but its absolute weight continues to increase until the onset of puberty. Then it undergoes an involution and progressively decreases in size throughout adult life. The degree of involution varies greatly from species to species and appears to be correlated with the hormonal status of the animal.

Lymphocytes. The thymic stem cells generate a large population of small lymphocytes (thymocytes) through a series of mitotic divisions. Simultaneously these dividing lymphocytes show evidence of cellular differentiation within the special thymic environment. During this division and maturation phase the developing thymocytes undergo an intrathymic migration from the peripheral cortical area to the medullary core of the organ. Some thymocytes degenerate within the organ, but many enter the circulating blood and lymph systems at various stages of maturity. A small percentage of the T lymphocyte population (5–10%) within the thymus is antigenically competent and capable of recognizing antigenic determinants on foreign cells or substances. Some of the T lymphocytes have the capacity to lyse the foreign tissue cells, while others are involved in recognizing the "foreignness" of the antigens and assisting a second subpopulation of bone-marrow-derived lymphocytes (B lymphocytes) to respond to the antigen by producing a specific antibody. These two types of immunocompetent T lymphocytes are called killer cells and helper cells, respectively. They are involved in both tissue transplantation and

humoral antibody responses. These cells become a part of the memory cell components of the lymphoid tissue and recirculate between the blood and lymph systems. On the other hand, the vast majority of the thymic lymphocytes are immunologically incompetent (90–95%). Some thymocytes are thought to give rise to the smaller pool of immunocompetent T lymphocytes, but many emigrate into the circulating blood. Their extrathymic fate and function are presently unknown.

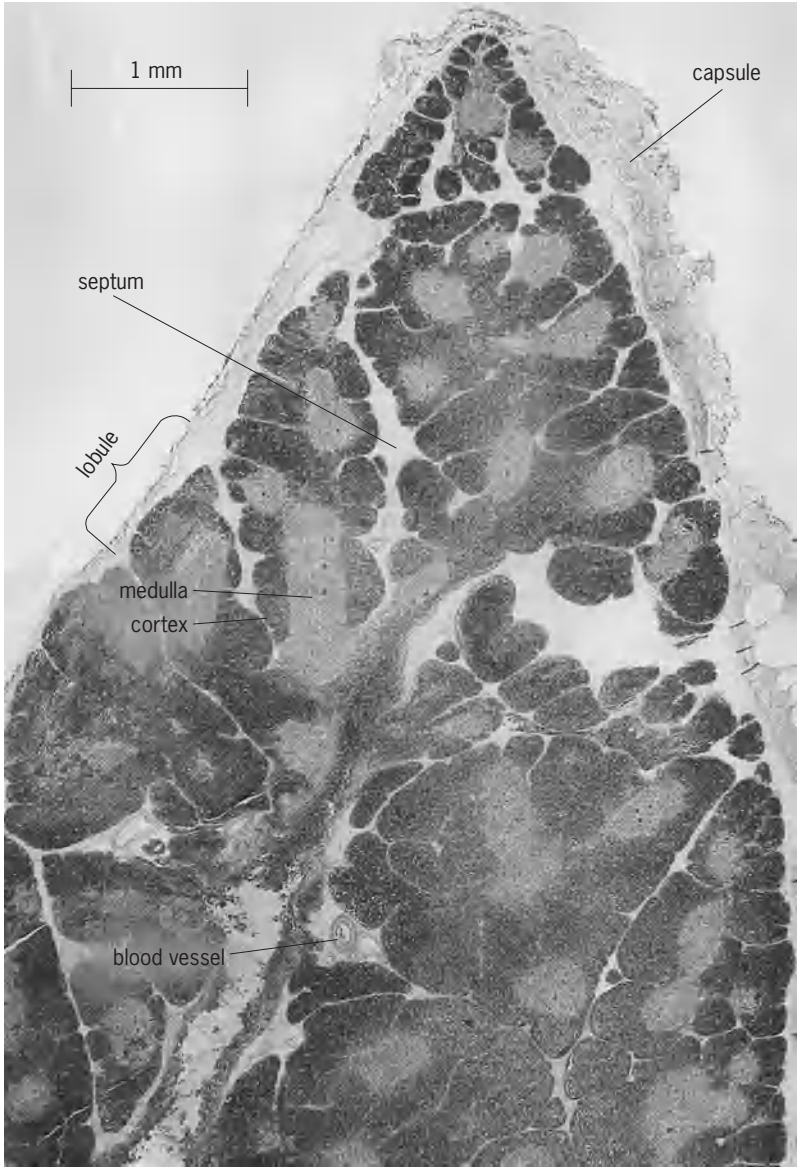
Comparative anatomy and embryology. Throughout life in most fishes and amphibians, and during development in other forms, the thymus is closely associated with the epithelial lining (endoderm) of the gill cavities, or the corresponding pharyngeal pouches of nonaquatic forms. *See* PHARYNX.

Fishes. The lamprey has a rudimentary thymus which is composed of a focus of lymphoid cells associated with the lining of the pharyngeal pouches. Elasmobranch fishes, sharks, and rays have a well-developed thymus which demonstrates a definitive cortex and medullary arrangement. These primitive vertebrates are also the first to show lymphoid aggregations in other areas of the body and immunological specificity toward foreign antigens.

Amphibians. The thymus occupies essentially the same position as in fishes with respect to the pharyngeal pouches. The tissue may be somewhat further removed from the pharyngeal mucous membrane and fuse into a bilateral organ which is composed of one or two masses in the neck.

Reptiles and birds. Thymus glands in these species usually occur as a bilateral chain of organs in the neck dorsolateral to the esophagus and medial to the jugular vein at approximately the level of the thyroid gland, or more rostrally. The individual anlage is derived mainly from the third and fourth pharyngeal pouches and may separate into several masses which correspond roughly to the spaces between the segmental nerves of the neck.

Mammals. The thymus gland is derived from the caudal pharyngeal pouches, III and IV. During development the thymuses become dissociated from the pharynx and appear to migrate both caudally and ventromedially. In many species the right and left glands approximate each other in the midline to form a single bilobed organ in the lower cervical or anterior thoracic region, or both. In humans and rodents the gland is principally situated in the anterior mediastinum between the sternum and the pericardium. Because of the migration pattern of this organ during development, inconstant accessory thymuses may be found in other locations between the pharynx and the thorax.



Thymus of a 3-month-old human infant. Histological section through a portion of one lobe.

Histology of mammalian thymus. Each lobe is covered by a connective tissue capsule which subdivides the lobe into numerous lobules by trabeculalike extensions into the gland. These connective tissue septa provide a pathway for blood vessels to enter the gland. Beneath the connective tissue capsule and septa are the parenchymal components of the thymus. The parenchyma consist of a peripheral cortical region which is adjacent to the capsule and septa and a centrally positioned medulla. In humans each lobule is 0.02–0.08 in. (0.5–2 mm) in diameter and continuous with adjacent lobules by parenchymal extensions (see **illus.**).

The thymic mass is composed principally of two types of cells, reticular cells and lymphocytes. A small number of macrophages are also present within the gland. The reticular cells have a histological appearance similar to those found in other lymphatic organs. They possess a relatively large reticulated, pale-staining nucleus with one or more visible nucleoli. The nucleus is surrounded by a scant acidophilic cytoplasm which contains a small number of subcellular organelles. The thymic reticular cells interconnect with each other in a spongelike framework by extensions of their cytoplasm. They differ from the mesodermally derived reticular cells of the other lymphomyeloid tissues in that they are derived from endoderm, lack the ability to produce reticular fibers, and are nonphagocytic. The proposed

functions of reticular cells are to support the lymphocyte population within the thymus, assist in the formation of the thymic-blood barrier, and possibly produce a thymic hormone.

The thymus lymphocytes include large, medium, and small forms. The larger lymphocytes are concentrated mainly in the subcapsular and septal regions of the cortex and possess a large round nucleus, 9 micrometers in diameter. These cells are mitotically active in producing smaller lymphocytes and usually show abundant amounts of strongly basophilic cytoplasm. The small lymphocytes (thymocytes) of the cortex superficially resemble the typical small lymphocytes of other lymphatic organs and circulating blood. They make up the vast majority of the cells found within the cortical region and possess a densely stained nucleus, 4–5 μm in diameter, with a slight amount of surrounding cytoplasm. Thymocytes differ from the small lymphocytes in the peripheral blood and other lymphatic tissue in that they are easily destroyed by corticosteroids and x-radiation; they are less well provided with intracellular organelles (nucleoli, mitochondria, and endoplasmic reticulum); they are immunologically incompetent; and they possess a unique array of surface antigens which allows them to be distinguished from the immunocompetent T lymphocytes and the bone-marrow-derived lymphocytes.

The medullary region contains fewer small lymphocytes relative to the number of reticular cells. This cellular organization produces a less compact appearance, and more large blood vessels are evident. The reduced cellular density of the medulla contrasts markedly with the densely populated cortex. The interface between these two distinct regions of the gland is termed the corticomedullary junction. The medulla contains the majority of the immunocompetent T lymphocytes, which are believed to arise from a population of maturing thymocytes. The majority of thymus macrophages are also present in the medulla and may be of blood-borne origin. In addition, a real histological peculiarity of the thymus exists within the medulla: Hassall's bodies, which are concentrically layered, flattened cells, probably of reticular origin. The overall size of these bodies ranges from 30 to 100 μm in diameter. The central cells in these structures tend to degenerate and hyalinize in a manner reminiscent of stratified epithelium and epithelial pearls of other organs. They are not present in the thymus at birth but do arise early in life during the development of immunological maturity. Their origin and function are unknown, but they serve as a histological marker for thymic tissue even during stages of severe involution.

The thymic blood supply and vascular organization have been extensively studied and provide some insight into possible mechanisms of thymic function. Arterioles leave the connective tissue septa and enter the parenchyma at the corticomedullary junction. They ramify within the medulla and supply a capillary network in the cortex. Blood in the cortex returns to the postcapillary venules, which are also located at the corticomedullary junction and in the medulla. The capillaries within the cortex are modified to prevent the movement of high-molecular-weight macromolecules into the thymic parenchyma. This specialized capillary structure is the basis for the thymic blood barrier, which functions to prevent antigenic material from penetrating the cortical region of the thymus. The blood vessels within the medulla are, however, permeable to antigens and to other macromolecules as well.

Physiology. Lymphopoiesis within the body may be induced from a stem cell source or an antigenic stimulation of immunocompetent small lymphocytes. The production of lymphocytes from stem cell sources occurs almost entirely within the central lymphoid tissue: bone marrow and thymus. Antigenically induced lymphopoiesis can occur at any site within the body into which the necessary cells can migrate and accumulate in the presence of the inducing antigen. These sites are most commonly found within the spleen and lymph nodes.

Lymphopoiesis. The lymphopoiesis within the thymic cortex is unique in that the lymphocytes are being produced and differentiating in an antigen-free environment. This specialized environment results from the presence of the thymic blood barrier within the cortical region of the thymus. As the cortical thymocytes mature, they undergo an intrathymic migration from the peripheral region of the cortex to the corticomedullary junction. This migration takes 2–4 days. The fate of the mature thymocytes is not entirely clear. Many of these cells undergo a premitotic death within the thymus and form visible structures known as tingible bodies. These are dark-staining nuclear concentrations principally found in the cortex, and each contains a double complement of DNA. Many of the remaining thymocytes emigrate from the thymus into the peripheral blood. It is believed that their route of emigration is via the venules and lymphatics present at the corticomedullary junction. This leaves a small percentage (5–10%) of the thymocytes which undergo an intrathymic maturation to give rise to that immunocompetent T-cell population which is found mainly in the medulla. The immunologically incompetent thymocytes can be distinguished from the immunologically competent T lymphocytes by the presence of specific surface antigens. The thymocytes that emigrate into the circulating blood may also be capable of transforming into mature T lymphocytes in the peripheral lymphoid tissue. *See* HEMATOPOIESIS.

Immunology. The T lymphocytes both within and outside the thymus display a functional heterogeneity and are involved in a wide range of immunological functions. They are responsible for tissue transplantation immunity and the rejection of homografts, for delayed hypersensitivity responses, and for immune responses against many microorganisms and viruses. In the tissue transplantation and delayed hypersensitivity responses, these cells interact with foreign cells and produce cell destruction by lysing the foreign or altered cells. An independent subpopulation of T lymphocytes functions as a group of

antigen-recognizing cells and assists the B lymphocytes in the production of humoral antibodies. The T lymphocytes are also responsible for much of the long-term immunological memory which allows higher animals to respond in a more efficient manner to recurring exposure to antigens. These cells make up the major portion of the population of small lymphocytes which constantly recirculate through the major lymphatic channels. *See* CELLULAR IMMUNOLOGY; TRANSPLANTATION BIOLOGY.

Thymectomy. Much knowledge about the function of the thymus has been derived from experiments on animals in which the thymus has been surgically or chemically removed. Such thymectomized animals show profound defects in their abilities to populate peripheral lymphoid tissues with lymphocytes and induce immune responses. If the thymectomy is performed in the prenatal or early postnatal periods, the thymectomized animals show severe lymphopenia and virtually total incompetence in their immunological reactivity to thymus-dependent antigens. The population of lymphocytes in the thoracic duct system of neonatally thymectomized animals may be only 2–3% of that found in normal, nonthymectomized control animals. The lymph nodes and spleens of these neonatally thymectomized animals are also poorly populated with lymphocytes. If the thymectomy is performed in an adult animal, little immediate change in population or function is noted. Peripheral lymphatic organs and pools of lymphocytes show normal populations of T lymphocytes. Immunological activity is also normal. With time, however, the T lymphocyte population begins to decrease and is no longer replaced in the absence of the thymus. Adult thymectomized animals will eventually show the same effects as those seen in neonatally thymectomized animals, particularly in responses to newly encountered antigenic stimuli.

Involution. The functional activity of the thymus is most intensive from the late fetal stages to puberty in most mammals. In humans the thymus begins to involute after puberty but is still an active and functional organ throughout most of adult life. The process of involution decreases the parenchymal elements of the thymus and gradually replaces these elements with fat. During involution, thymocytes are first to disappear, followed by the T lymphocytes and reticular cells. Histologically, the cortex becomes indistinct as the thymocytes are depleted, and finally the medullary region may be distinguishable only by the presence of keratinized Hassall's bodies. Premature involution is produced in times of stress, mainly caused by increased levels of circulating corticosteroids. Therapeutic dosages of corticosteroids may deplete a normal thymus of its lymphocytic population within 48 h. Repopulation of the thymus is then initiated by blood-borne stem cells from the bone marrow as the circulating steroid levels again approach normal concentrations.

Hormones. In addition to lymphopoiesis, production of a hormone has been proposed as a second major thymic function. There is much speculative evidence for the presence of a thymus hormone, but a great deal of controversy still surrounds its specific nature and action. In fact, more than one hormone has been proposed. It has been suggested that the endodermally derived reticular cells of the thymus are the sources of these hormones and that they act to produce a maturing influence on the differentiating thymocytes. Their influence is not only intrathymic but systemic, as suggested by studies involving thymic extract injections and transplants with thymuses enclosed in millipore chambers. What role these putative hormones may play in thymic function remains to be determined through further investigation. The thymic response to pituitary growth hormone and adrenal corticosteroids, together with its endocrinelike developmental history, suggests that some hormonal activity is plausible. *See* ENDOCRINE SYSTEM (VERTEBRATE).

Clinical disorders. While the preceding discussion emphasizes the role of the thymus in providing the animal with an efficient means of protection against certain diseases, a defect in thymic function may underlie some of the etiology for other diseases. Myasthenia gravis has been shown to correlate closely with thymic disorders. Thymoma and the presence of thymic germinal follicles are commonly found in individuals with myasthenia gravis. Some evidence for a common antibody against thymic epithelial cells and muscle cells suggests that myasthenia gravis may be an autoimmune disease. Thymomas have also been associated with anemias and other suspected autoimmune disorders. These diseases appear to be enhanced when an immunological deficiency is present. It has been postulated that clones of cells that could react with self-antigen escape from a suppressed stage and proliferate. These lymphocytes then react with the various organ systems which contain these antigens. It has also been shown that immune deficiencies can enhance the incidence of carcinogenesis (cancer). Thymic disorders and involution are associated with increased carcinogenesis. On the other hand, some leukemias require an active thymus to sustain the disease. This may be due to the requirement for some hormone factors or to the intrathymic growth of a viral agent that could promote the leukemia. In these cases a thymectomy may be an asset in the treatment of the disease even though it results in an immunological deficiency with time. Thymocytes are also affected in acquired immune deficiency syndrome (AIDS). A human immune deficiency virus (HIV) can infect thymocytes and lead to their destruction and a subsequent immunodeficiency. *See* AUTOIMMUNITY; LEUKEMIA; MYASTHENIA GRAVIS; ONCOLOGY.

Bursa of Fabricius. The development of the bursa of Fabricius as a gut-associated lymphoepithelial organ in birds appears to be a unique step in evolution. The bursa serves as a second central lymphoid organ and, together with the thymus, plays an important role in the development of immunological competency in birds. Mammals lack a definitive bursa, and attempts to define a bursal equivalent have thus far failed. Some scientists feel that the gut-associated lymphoid tissue may represent this avian structure, but it is equally probable that the

bone marrow lymphocyte population contains the bursal equivalent in mammals. These cells would be the source of the B lymphocyte population. *See* LYMPHATIC SYSTEM.

Charles E. Slonecker

Bibliography. P. S. Amenta, *Histology and Human Anatomy*, 6th ed., 1991; N. A. Byron and J. R. Hobbs (eds.), *Thymic Factor Therapy*, 1984; J. A. Goss, Y. Nakafusa, and W. M. Flye, *The Thymus and Cellular Immunity*, 1993; M. W. Hess, *Experimental Thymectomy*, 1968; L. C. Junqueira et al., *Basic Histology*, 9th ed., 1998; M. D. Kendall (ed.), *The Thymus Gland*, 1982; D. Metcalf, *The Thymus*, 1966.

Thyrocalcitonin

A hormone, the only known secretory product of the parenchymal or C cells of the mammalian thyroid and of the ultimobranchial glands of lower forms. The hormone has been isolated and characterized from porcine and human thyroid tissue as well as from fish ultimobranchial tissue. In all cases it is a polypeptide of 3800 molecular weight characterized by a single polypeptide chain. The amino acid sequence of the porcine hormone is

H_2N-Cys-Ser-Asn-Leu-Ser-Thy-Cys-Val-Leu-Ser-Ala-
(disulfide bridge between the two Cys residues)
Tyr-Trp-Arg-Asn-Leu-Asn-Asn-Phe-His-Arg-Phe-Ser-
Gly-Met-Gly-Phe-Gly-Pro-Glu-Thr-Pro-$CONH_2$

The sequences of both the human and fish hormone have also been determined and are different from each other and from the porcine hormone except in the *N*-terminal seven amino acids. *See* ULTIMOBRANCHIAL BODIES.

Physiological activity. In conjunction with the parathyroid hormone, thyrocalcitonin is of prime importance in regulating calcium and phosphate metabolism. Its major function is to protect the organism from the dangerous consequences of elevated blood calcium. Its sole known effect is that of inhibiting the resorption of bone. It thus produces a fall in the concentration of calcium and phosphate in the blood plasma because these two minerals are the major constituents of bone mineral and are released into the bloodstream in ionic form when bone is resorbed. The inhibition of bone resoption also leads to a decreased excretion in the urine of hydroxyproline-containing peptides, which are end products in the breakdown of bone matrix, collagen. *See* BONE; CALCIUM METABOLISM; PARATHYROID GLAND; PARATHYROID HORMONE.

Thyrocalcitonin also causes an increased excretion of phosphate in the urine under certain circumstances, but a question remains as to whether this is a direct effect of the hormone upon the kidney or an indirect consequence of the fall in blood calcium which occurs when the hormone inhibits bone resorption. *See* PHOSPHATE METABOLISM.

Biochemical mechanisms. Practically nothing is known about the biochemical basis for the hormonal inhibition of bone resorption, although it has been shown that thyrocalcitonin inhibits the resorption of bone grown in tissue culture when this resorption is increased by the addition of parathyroid hormone, vitamin D, or vitamin A. *See* VITAMIN A; VITAMIN D.

Repeated administration of the hormone to young animals leads to an increased bone mass, but it is not known whether this is due solely to the fact that the hormone has inhibited bone resorption while bone formation has continued unchanged, or whether it has also enhanced bone formation.

Age factors. One of the most unusual features of this hormone's effect is the considerable decrease in sensitivity with age. With many hormones, a given dose, on a unit per weight or surface area basis, is one-half to one-third less effective in adults as compared to young animals. However, in the case of thyrocalcitonin, adult animals are considerably less sensitive, by a factor of one-fiftieth or more, than young growing animals. There is a good correlation between this decreased sensitivity and the fact that the rate of bone remodeling, both formation and resorption, is very high in young growing animals but decreases greatly in adult life.

Control mechanisms. Another unusual aspect of thyrocalcitonin physiology is the fact that the content of thyrocalcitonin in thyroid tissue increases greatly when there is no stimulus for its release, that is, when the blood calcium is low for any reason. In most other endocrine organs there appears to be a close relationship between the rate of synthesis of the hormone and the rate of its secretion, so that the content of hormone increases with increased secretion rates and decreases with decreased secretion rates.

Thus the only known stimulus to thyrocalcitonin secretion is a rise in blood calcium. The increased secretion, so induced, leads to a fall in blood calcium so that a negative feedback relationship exists between thyrocalcitonin secretion, bone resorption, and blood calcium.

Therapeutic value. There are no known diseases which have been clearly recognized as being due to chronic overproduction or underproduction of this hormone. Nevertheless, it has been found effective in the treatment of two known human diseases: idiopathic hypercalcemia of infancy, and hypercalcemia in adults caused by overproduction of parathyroid hormone. In both conditions, the concentration of calcium in the blood is greater than normal, and this leads to serious consequences, particularly calcification of the kidneys, which may eventually cause sufficient damage to the kidneys to cause death. The administration of thyrocalcitonin to people with these diseases leads to a prompt fall of the plasma calcium to normal levels.

It is hoped that this hormone will also be useful in several other disorders of bone metabolism, but because of the extremely small amount present in thyroid glands, and the difficulty of purifying it from these glands, only limited amounts have been available for medical use. *See* THYROID GLAND.

Howard Rasmussen

Bibliography. D. O. Norris, *Vertebrate Endocrinology*, 3d ed., 1996.

Thyroid gland

An endocrine gland found in all vertebrates that produces, stores, and secretes the thyroid hormones. The primary function of the thyroid, in warm-blooded vertebrates at least, is to regulate the rate of metabolism. In humans, the gland is located in front of, and on either side of, the trachea (**Fig. 1**). Thyrocalcitonin, a hormone of the thyroid gland, assists in regulating serum calcium by reducing its levels. The thyroid gland is capable of accumulating inorganic iodides and uniting them with the amino acid tyrosine to produce iodinated proteins. This activity is regulated by thyrotropic hormone from the anterior lobe of the pituitary gland. Microscopically (**Fig. 2**), the tissue consists of thyroid follicles composed of a single layer of secretory cuboidal epithelium which secretes hormone as a colloid into a blind lumen. The height of the secretory cells and the amount of colloid vary with the functional state of the gland. A main site of action of thyroid may be in the mitochondria, where energy-rich phosphate bonds are formed.

The thyroid receives an exceptionally rich blood supply. Postganglionic fibers from the cervical ganglia and vagus enter with the blood vessels and form extensive plexuses around the smaller arteries. This innervation does not appear to be essential for normal thyroid function except as it controls the rate of blood flow through the gland.

Comparative Anatomy

The thyroid is usually a well-encapsulated single gland, often with two distinct lobes connected by a narrow isthmus. However, there are morphological variations among the various vertebrate classes.

Fishes. In most fishes the thyroid is represented by diffuse masses of follicles scattered along the large arteries which enter the gills ventrally, or tucked between muscles in the pharyngeal floor. In elasmobranchs the follicles are aggregated in a single mass. In the remaining fishes (except lampreys) and in higher vertebrates a pair of thyroid masses, often connected by a median strand (isthmus), usually occurs. Minute accessory thyroid masses are common.

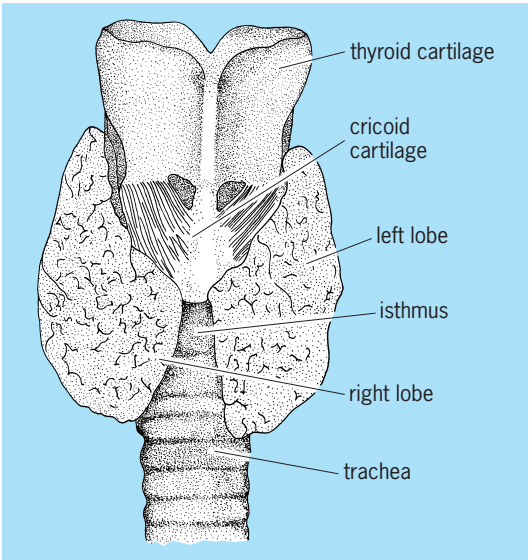


Fig. 1. Ventral view of human thyroid gland shown in relation to trachea and larynx. (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)

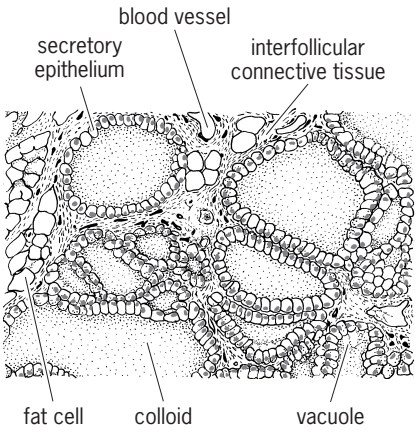


Fig. 2. Microscopic view of histologic features of the normal thyroid gland of the rat. (*After C. D. Turner, General Endocrinology, 4th ed., Saunders, 1966*)

Amphibians. In amphibians the thyroid glands lie under cover of certain muscles of the buccal or pharyngeal floor near the caudal angle of the jaws (anurans), or at the base of the branchial arches (urodeles).

Amniotes. In amniotes the two oval or elongated glands lie against the trachea, immediately below the larynx in mammals (Fig. 1), partway down the trachea in lizards, still farther caudad in other reptiles, or just above the bifurcation of the bronchi in birds. In most mature reptiles the gland is unpaired. In humans the thyroid glands are attached to the thyroid (shield-shaped) cartilage of the larynx.

Phylogeny

A clue to the phylogenetic history of the thyroid is found in lampreys. In marine species an elongated rod of cells capable of selectively absorbing iodine-rich substances is located in the pharyngeal floor between the second and fifth gill pouches. In larval brook lampreys the iodine-capturing cells are part of a complicated subpharyngeal gland (endostyle) which evaginates from the embryonic pharyngeal floor. At metamorphosis the gland loses its connection with the pharynx and remains as isolated thyroid masses underneath the pharynx. Embryonic origin of the thyroid of higher vertebrates as a pharyngeal outpocketing probably represents a recapitulation of the phylogeny of the gland. Although the embryonic pharyngeal connection is usually lost, a duct remains patent in some elasmobranchs. Even in humans remnants of the duct may persist.

George C. Kent

Embryonic Origin

In the lamprey, the thyroid arises from some of the cells that line the larval endostyle, a groove lying in the floor of the pharynx. In all other vertebrates it originates from the same region, usually arising as a groove or pit at the level of the first pair of gill pouches (**Fig. 3**). In some fishes and amphibians

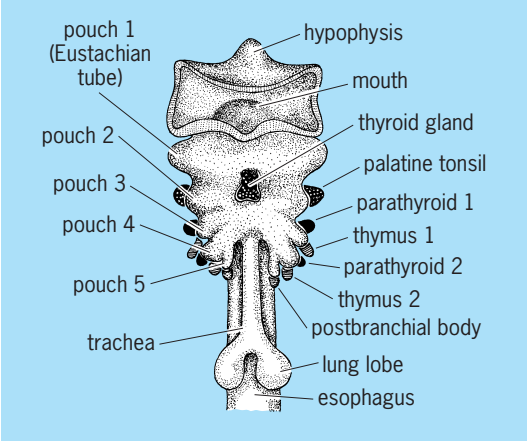


Fig. 3. Ventral view of pharyngeal region of a human embryo showing the pharyngeal pouches and their glandular derivatives; semidiagrammatic. (*After H. V. Neal and H. W. Rand, Chordate Anatomy, Blakiston, 1939*)

the cells appear as a solid bud rather than a hollow structure. In further development the bud or pit separates from the pharynx and migrates backward to lie ultimately in the throat region below the ventral aorta or the trachea. During this migration the cells of the primordium multiply and arrange themselves into elongate cords or flattened plates. The cords finally break up into discontinuous groups of cells, the early follicles, which with the onset of secretion become blind sacs. The definitive follicles consist of a single epithelial cell layer surrounding a cavity, the lumen, filled with a fluid, the colloid. The follicles are bound together by connective tissue (Fig. 2), and the whole gland becomes covered by a well-defined capsule. Frequently, small groups of functional follicles are left behind along the embryonic route of migration. These are called accessory thyroids or thyroid "rests." In some bony fishes, in which the thyroid is not encapsulated, follicles may wander rather far from the pharyngeal area to such unusual regions as the kidney, spleen, eye, or brain.

Various vertebrate groups differ as to the time when the thyroid first shows evidence of secretory activity and when its hormones begin to affect development. In the frog the first organized follicles are seen quite early, when the tadpole has completed only about 10% of larval life; in the chick they appear later, when about 30% of the period of incubation has passed; in the human, at approximately halfway through intrauterine development; and in the rat, after 80–90% of intrauterine life. The onset of thyroid function, determined by the point at which the gland is able to take up radioiodine, precedes the actual appearance of colloid, but there is evidence that organic combination of iodine, and therefore formation of true hormone, is closely correlated with the time of origin of colloid.

W. Gardner Lynn

Development, Differentiation, and Morphogenesis

The thyroid gland as a distinct histologic entity is found only in vertebrates. The origin of vertebrates from the lower forms is obscure, with the closest relatives being the protochordates, tunicates, and amphioxus. Even in these groups, however, evidence of thyroid evolution from prevertebrate ancestry is inconclusive. In forms below vertebrates, there is no significant thyroid hormone formation, although iodotyrosines (iodinated amino acids; of unknown function) are found. Conversion from the sessile form of the jellyfish to the medusa or free-floating form depends partly on iodine concentration and suggests some form of control by iodotyrosine.

The role of thyroid hormones in lower vertebrates, therefore, is of interest. The thyroid hormones thyroxine (T4) and triiodothyronine are synthesized in the thyroid gland, but most of triiodothyronine is formed by the peripheral deiodination of thyroxine. Triiodothyronine is considered to be the active hormone, with thyroxine being considered a prohormone. Triiodothyronine plays an important role in thermal regulation and in the control of postembryonic development. *See* THERMOREGULATION.

Developmental effects are found most strikingly during the metamorphosis of a tadpole into a frog; metamorphosis will not occur in the absence of thyroid hormones. Interestingly, the deiodinase enzyme responsible for converting thyroxine to triiodothyronine is detectable in the skin of premetamorphic tadpoles, and enzyme activity significantly increases as the tadpoles go through spontaneous metamorphic climax. The major tissues responsible for producing triiodothyronine from thyroxine in amphibians are gut and skin. Significant deiodinase activity is present in tail tissue once the fin starts to be resorbed.

Thyroid hormones and retinoic acid. Development depends on morphogenetic signals that determine changes in gene expression in particular cells. The role played by thyroid hormones in development became clearer after it was discovered that retinoic acid is a major factor in morphogenesis and the retinoic acid receptor is a member of the thyroid receptor family. Furthermore, the retinoic acid receptor can activate gene expression through a thyroid hormone response element. The implication is that thyroid hormones and retinoic acid, acting through their respective receptors, control overlapping gene networks involved in the regulation of vertebrate development, differentiation, and morphogenesis.

The relationship between the retinoid receptor and the thyroid hormone receptor is surprising because retinoids and thyroid hormones bear little resemblance to one another. However, this probably reflects a common mode of action by which they elicit their particular regulatory effects. Thus, the interaction of retinoic acid with its receptor would induce a cascade of regulatory events that result from activation of specific sets of genes by the hormone receptor complex. The demonstration that the retinoic acid receptor is part of the steroid receptor superfamily suggests that mechanisms controlling morphogenesis may be more universal than previously suspected.

Metamorphosis. Thyroid hormones are essential for amphibian metamorphosis. Extensive structural changes occur during the transition of the larval tadpole to the adult frog (**Fig. 4**): Resorption of gills and tail occurs with development of lung and limb buds. Profound biochemical changes also occur.

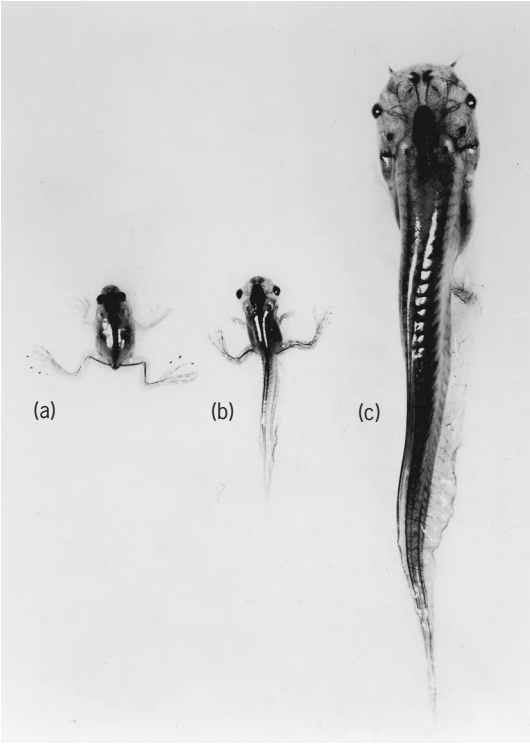


Fig. 4. Effect of thyroid hormone on metamorphosis of the amphibian *Xenopus laevis* (South African clawed toad). (*a*) Newly metamorphosed toad. (*b*) Metamorphosing tadpole. (*c*) Giant tadpole showing no metamorphosing signs after thyroidal hormone synthesis was blocked by immersion in perchlorate, an antithyroid drug. (*From J. A. Dodd and A. J. Matty, Comparative aspects of thyroid function, in R. Pitt-Rivers and W. R. Trotter, eds., The Thyroid Gland, Butterworth, 1964*)

Metamorphosis has been divided into three stages: premetamorphosis, prometamorphosis, and metamorphic climax. Premetamorphosis is characterized by rapid body growth without differentiation and without thyroid hormone. Prometamorphosis is characterized by diminishing growth rate, the beginnings of differentiation, and increasing thyroid hormone concentrations. The metamorphic climax is associated with growth cessation, maximal rates of differentiation, and a thyroid hormone surge that eventually drops to undetectably low levels. Although the tadpole is very sensitive to thyroid hormone effects, the frog seems not very sensitive.

Tremendous diversity exists within the metamorphic process. Some organs are stimulated under the influence of thyroid hormones while others resorb. Despite the similarity between thyroid hormones and steroid receptors, gonadal development in amphibians is not influenced by thyroid hormones. If early metamorphosis is evidenced by exogenous administration of thyroid hormones, precocious sexual maturity does not occur.

The sequencing and spacing of metamorphosis are determined by the secretory activity of the thyroid and by differences in the rate of response of affected tissues. For instance, hindlimb growth begins early and requires weeks or months for completion. In contrast, tail resorption and loss of mouthparts in the tadpole occur later in metamorphosis and more rapidly.

The various tissue and organ responses differ among groups of amphibians. Anurans lose their tails at metamorphosis while urodeles do not. Rapid growth of the hindlimbs occurs in the anurans but not in the urodeles. The characteristic differences in the mode of response are genetically determined. Tissues of very young larvae are not capable of manifesting any metamorphic response even when exposed to high concentrations of thyroid hormones. Sensitivity to thyroid hormones appears at the stage of development when the external gills become covered by the opercular folds. The concept of a time window for the action of thyroid hormones occurs with higher species as well.

Neoteny. Some amphibians, such as the Mexican axolotl (*Ambystoma mexicanum*), fail to metamorphose. Instead, they mature and reproduce but retain a larval form; this is known as neoteny. However, treatment with thyroid hormones will induce these salamanders to undergo metamorphosis. Neoteny is characterized by the cessation of metamorphosis at the larval stage, coupled with an increase in size of the larvae and maturity of their gonad systems.

The hypothalamus is essential for amphibian metamorphosis. The preoptic nucleus of the hypothalamus may be the location of the control center for thyroid-stimulating hormones (TSHs). Also, an important fraction of the brain thyrotropin-releasing hormone (TRH) is located in this area. In birds and mammals, TRH stimulates thyrotrophic cells in the pituitary to produce TSH; TRH is also present in the hypothalamus of amphibians, including the axolotl, and may control thyroid activity in amphibians.

While thyroxine is indispensable for metamorphosis of amphibian larvae, the pituitary hormone prolactin is the growth-promoting hormone. An antagonistic interaction exists between prolactin and thyroxine, but it does not appear to be mediated through the thyroid gland.

Fishes. Extraordinarily high concentrations of triiodothyronine have been found in the ammocoete stage of the lamprey, a cyclostome. Nuclear triiodothyronine receptors in the liver of these animals appear to resemble mammalian receptors. Despite the presence of thyroid hormones and thyroid hormone receptors, the function of thyroid hormones in fishes has remained elusive. Little convincing direct evidence has been found to indicate that thyroid hormones are involved in growth and development of fishes. Thyroid hormones characteristically function in coordination with other hormones and growth factors. If the coordinate signal is missing, thyroid hormones may not produce any effects.

Exogenous thyroid hormones do induce metamorphosis in flounder larvae. The antithyroid drug, thiourea, blocks the effect of thyroid hormones. The role of thyroid hormones in metamorphosing larvae, at least in flounder, is comparable to its role in amphibians.

Development in homeotherms. Thyroid hormone also plays an essential role in the ontogenesis of higher vertebrates. As in amphibians, it does this at a

well-defined stage of development. The precise stage at which thyroid hormones influence development in a given species is related to duration of gestation and degree of maturity at birth.

In the rat the diminished growth rate characteristic of a deficiency of thyroid hormones (hypothyroidism) does not occur until the neonate is 8 days old. Clinical recognition of hypothyroidism in humans at birth is extremely difficult. Apparently, the major effects of thyroid hormones in both the rat and the human become manifest after a surge of thyroid hormones in a manner analogous to the amphibian metamorphic climax.

The manifestation of neonatal hypothyroidism in humans is most noticeably expressed in marked skeletal retardation with profound disturbances in the central nervous system. Thyroid hormone deficiency involves the maturation of individual cells as well as the retardation of general body growth. Although deficiency of thyroid hormones in the adult can be easily corrected, abnormalities in the neonate can be prevented only if the hormones are replaced at specific times. Thyroid hormones appear to function by an increase in general growth as well as by acceleration of differentiation of specific cell types. Thyroid hormones are essential for the pituitary gland to produce growth hormone and perhaps other growth factors as well. *See* ADENOHYPOPHYSIS HORMONE.

Gerard N. Burrow

Physiology

The thyroid gland synthesizes, stores, and secretes the iodine-containing hormones, thyroxine and triiodothyronine. Thyroid hormones regulate metabolic rate in warm-blooded animals and are essential for normal growth and development.

Several technical advances have contributed greatly to the understanding of thyroid physiology. One is the use of radioactive isotopes of iodine, which has enabled investigators to study the fate of iodine in the body. Another is the development of the technique of radioimmunoassay, which has afforded a means of quantifying the minute quantities of hormones and other compounds in the blood and tissues. *See* AUTORADIOGRAPHY; IODINE; RADIOIMMUNOASSAY.

Metabolism of hormones. Upon entering the blood, both thyroxine and triiodothyronine become bound to specific plasma proteins. The interaction of the thyroid hormones with these so-called transport proteins conforms to a reversible binding equilibrium in which almost all the hormone is bound and only a very small proportion is unbound or free. Only the unbound hormone is available to the tissues for metabolism and induction of hormone action. The binding of triiodothyronine to the transport proteins is much weaker than that of thyroxine. A major route of metabolism of the thyroid hormones is by way of deiodination, in which the iodine atoms—four in the case of thyroxine and three in the case of triiodothyronine—are removed progressively from the compound. In the case of thyroxine, the initial monodeiodination can occur either in the lower ring, resulting in the formation of triiodothyronine, or in the upper ring, yielding reverse triiodothyronine (**Fig. 5**). This process accounts for about 80% of the triiodothyronine produced, the remainder being secreted by the thyroid gland. It also accounts for all the reverse triiodothyronine generated. Since triiodothyronine is about three times more active than thyroxine, whereas reverse triiodothyronine is inactive, the relationship between the initial upper- or lower-ring monodeiodination of thyroxine may represent a level of regulatory control of hormone action in the tissues. Of the iodide liberated by deiodination, part is reaccumulated by the thyroid gland and utilized in hormone synthesis, and the remainder is excreted by the kidneys. Other routes of metabolism of the thyroid hormones include conjugation with glucuronate and sulfate in the liver followed by their excretion into the bile and modification of the side chain to form other derivatives, such as tetra- and triiodothyroacetic acids. *See* THYROXINE.

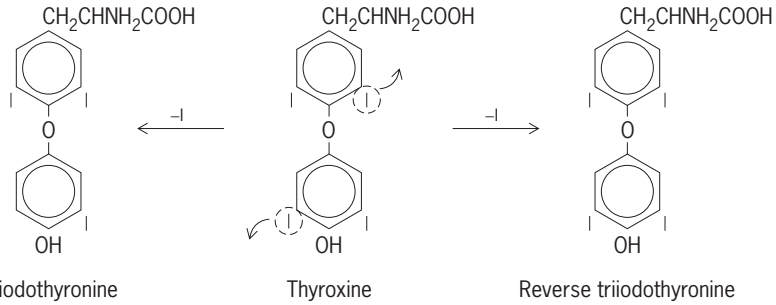


Fig. 5. Monodeiodination of thyroxine yielding either triiodothyronine or reverse triiodothyronine.

Regulation of thyroid function. The function of the thyroid gland is regulated by the thyroid-stimulating hormone of the anterior lobe of the pituitary gland. The secretion of TSH is regulated by a negative feedback mechanism in the pituitary, under hypothalamic control, which is sensitive to the concentration of unbound thyroxine or triiodothyronine in the perfusing blood. A decrease in the concentration of hormone in the blood stimulates the secretion of TSH, which tends to increase the function and size of the thyroid gland, thereby restoring to normal the concentration of thyroid hormones. Conversely, an increase in the concentration of hormone in the blood decreases the secretion of TSH and thereby tends to decrease the function and size of the gland. The threshold of feedback control of the secretion of TSH is set by the hypothalamic peptide, TSH-releasing hormone. In addition to regulation by the hypothalamus-pituitary complex, the thyroid gland possesses an intrinsic regulatory mechanism whereby the glandular content of organic iodine, present as iodinated compounds in thyroglobulin, determines in an inverse manner the activity of the iodide-trapping mechanism and the sensitivity of the gland to TSH. *See* ENDOCRINE MECHANISMS.

When the intrinsically normal thyroid gland is deprived of TSH stimulation as a result of hypothalamic or pituitary disease, atrophy of the thyroid gland and hypothyroidism ensue.

Physiological actions. Two cardinal actions of the thyroid hormones are their stimulation of the basal metabolic rate (BMR) in warm-blooded animals and their influence on the growth and development of tissues. The metabolism of carbohydrates, proteins, and fats is influenced by the thyroid gland; too much hormone in the circulation intensifies symptoms of diabetes. *See* DIABETES.

A deficiency of thyroid hormones (hypothyroidism), resulting from either a qualitative or a quantitative deficiency of thyroid tissue, produces certain distinctive changes in humans or experimental animals. The decrease in basal metabolic rate is reflected in a decrease in oxygen consumption in the whole animal and in isolated tissue preparations. Body temperature may be subnormal. Cardiac output, heart rate, and respiratory rate are decreased. There is a slowing of mental activity and depression of neuromuscular excitability. In humans, the skin is cool, dry, and coarse, and accumulation in the dermis of a mucinous material gives it a puffy appearance (myxedema). When hypothyroidism begins in early life and treatment is delayed, severe retardation of mental development and growth results (cretinism). If begun in time, treatment with thyroid hormones reverses virtually all the manifestations of the hypothyroid state.

An excess of thyroid hormones (hyperthyroidism) results from toxic goiter in humans, or its effects can be produced experimentally by the administration of large quantities of thyroid hormones. In general, the effects are the converse of those that occur in hypothyroidism. The basal metabolic rate is increased. Cardiac output, heart rate, and respiratory rate are increased, and neuromuscular excitability is enhanced. There is increased sweating, weight loss despite an increased appetite, and muscle weakness. Protrusion of the eyeballs (exophthalmos) occurs in some patients with toxic goiter; the reason for this is not known, but it is not due to the excess of thyroid hormones per se, since it cannot be produced by their administration.

Thyroid hormones do not appear to stimulate the metabolic rate in cold-blooded animals.

Kenneth A. Woeber

Biochemistry

The size of the normal thyroid is subject to more variation than other organs in the body, fluctuating with age, reproductive state, diet, and external environment. The average weight in the adult human is from 0.08 to 1.4 oz (25 to 40 g).

The thyroid has the greatest ability to trap iodine, binding one-third to one-quarter of the total amount of this element in the body. The gland takes up iodine and fixes it extremely rapidly. As the first step in the synthesis of the thyroid hormone, iodine is extracted from the circulation, and in the gland it combines with the amino acid tyrosine. The colloid that fills the follicles of the gland is a protein globulin, known as thyroglobulin. Thyroglobulin is a glycoprotein of molecular weight 669,000. It is stable in salt solutions and ranges between pH 5 and slightly alkaline pH, and dissociates into subunits under a variety of conditions. After iodination, thyroglobulin contains approximately twelve molecules of monoiodotyrosine plus diiodotyrosine, two of thyroxine, one-third of triiodothyronine, and traces of other tyrosine derivatives (Fig. 5). It cannot pass through the cell membrane into the circulation in this form, however, because of the large size of the molecule, so the thyroglobulin is broken down by an enzyme system into its constituents, one of which is thyroxine. Thyroxine is the form in which the hormone enters the circulation. The thyroxine may be degraded or changed by the tissues to still another compound, L-triiodothyronine, which is biologically more active than thyroxine itself. *See* METABOLISM; THYROID GLAND DISORDERS.

Choh Hao Li

Bibliography. R. Ekholm, *Control of the Thyroid Gland: Regulation of Its Normal Function and Growth*, 1989; M. Greer, *The Thyroid Gland*, 1990; S. H. Ingbar and L. E. Braverman (eds.), *Werner's The Thyroid: A Fundamental and Clinical Text*, 5th ed., 1986; J. Tepperman, *Metabolic and Endocrine Physiology*, 4th ed., 1980.

Thyroid gland disorders

Disorders of the thyroid gland may be classified according to anatomical and functional characteristics. Those thyroid disorders that are primarily anatomical include goiter and neoplasia; those that are primarily functional result in either hyperthyroidism or hypothyroidism.

Thyroid gland enlargement. Thyroid gland enlargement, or goiter, is the most common disorder. Its classification is based upon both the anatomy and function of the gland (see **table**). An enlarged but normally functioning thyroid gland is termed a nontoxic goiter. This condition affects hundreds of millions of people throughout the world in areas where the diet is deficient in iodine. In other areas it may be caused by subtle disorders in the biosynthesis of thyroid hormone. In both cases, there is compensatory enlargement of the gland that can be diffuse and symmetrical or can produce a multinodular goiter. When goiter is present in more than 10% of the population it is termed endemic and can represent a major public health issue. If the condition is untreated, the thyroid gland can increase 10–20 times its normal size. Dietary iodine supplementation is effective in preventing endemic goiter. Surgical removal of the enlarged gland is rarely required.

The multinodular goiter can grow independently from pituitary gland control and produce excess thyroid hormone, causing hyperthyroidism. However, hyperthyroidism is most often the result of Graves' disease. Thyroid enlargement can also be caused by Hashimoto's thyroiditis, in which the individual's immune system develops abnormal antibodies that react with proteins in the thyroid gland. This autoimmunity can make the gland enlarge or become underactive. *See* AUTOIMMUNITY.

Neoplasia. Tumors of the thyroid account for a small fraction of human neoplasms and an even smaller fraction of deaths due to cancer. The vast

Diseases of the thyroid gland and their effects on thyroid size and hormone production

Disease	Thyroid size	Hormone production
Multinodular goiter	Increased	Normal or increased
Thyroid adenoma	Normal or increased	Normal or increased
Thyroid carcinoma	Increased	Normal
Hashimoto's thyroiditis	Increased	Normal or decreased
Graves' disease	Increased	Increased
Hypothyroidism	Normal, increased, or decreased	Decreased

majority of thyroid neoplasms are follicular adenomas, which are benign; however, some thyroid neoplasms are malignant. Malignant tumors generally arise from thyroid follicular epithelium (follicular, papillary, and anaplastic carcinoma), but may also derive from parafollicular cells (medullary thyroid carcinoma) or from lymphocytes within the gland (lymphoma). Rarely, tumors arising elsewhere in the body can metastasize to the thyroid gland. Risk factors for the development of thyroid neoplasia include exposure of the thyroid to ionizing radiation (tumors of follicular epithelium), a family history of thyroid cancer (medullary carcinoma of the thyroid), and lymphocytic or Hashimoto's thyroiditis (lymphoma of the thyroid). Most thyroid tumors are detected as a solitary nodule when the thyroid is palpated. Relative to normal thyroid tissue, these nodules are usually hypofunctional and thus appear inactive on isotope scans of the thyroid gland. Aspiration (removal of fluid by suction) of these nodules for diagnostic cytologic evaluation has become standard practice. If malignant cells are detected in this manner, the tumor is surgically removed and radioative iodine is usually administered to ablate any residual thyroid tissue. The survival of patients 5 years after diagnosis is excellent (approximately 95%). *See* ONCOLOGY; RADIOACTIVE TRACER.

Hyperthyroidism. Hyperthyroidism is the clinical condition that results from excessive levels of the circulating thyroid hormones thyroxine and triiodothyronine, which are secreted by the thyroid gland. Signs and symptoms include weight loss, tachycardia (increased heart rate), heat intolerance, sweating, and tremor. Graves' disease, the most common form of hyperthyroidism, is mediated by an abnormal antibody directed to the thyroid-stimulating hormone (TSH) receptor on the surface of the thyroid cell, which stimulates secretion of thyroid hormone. Unique to Graves' disease is the associated protrusion of the eyes (exophthalmos). As noted above, single or multiple thyroid nodules can also cause hyperthyroidism when they produce excessive levels of thyroid hormone. Acute or subacute inflammation of the thyroid (thyroiditis) causes hyperthyroidism when thyroid hormone is released by the inflamed gland into the bloodstream.

In all cases, the goal of therapy is to promptly reduce thyroid hormone levels to normal. This is most rapidly accomplished by the administration of drugs that impair thyroid hormone biosynthesis and the release of preformed thyroid hormone from the gland. Other possible treatments include administration of radioactive iodine or surgical excision of the thyroid gland.

Hypothyroidism. Hypothyroidism is the clinical state that results from subnormal levels of circulating thyroid hormones. Manifestations in infancy and childhood include growth retardation and reduced intelligence; in adults, cold intolerance, dry skin, weight gain, constipation, and fatigue predominate. Individuals with hypothyroidism often have a slow pulse (bradycardia), puffy dry skin, thin hair, and delayed reflexes. In its most extreme form, hypothyroidism can lead to coma and death if untreated.

The vast majority of cases are due to thyroid gland failure as a result of chronic inflammation (Hashimoto's thyroiditis). When the thyroid begins to fail, the pituitary gland senses a reduction of thyroid hormone in the circulation and responds by secreting thyrotropin. The resultant rise in serum thyrotropin is the biochemical hallmark of primary hypothyroidism. Hypothyroidism is treated by the oral administration of thyroxine. *See* ADENOHYPOPHYSIS HORMONE; ENDOCRINE MECHANISMS; PITUITARY GLAND; THYROID GLAND; THYROID HORMONES.

Leslie J. DeGroot; David A. Ehrmann

Bibliography. R. H. Cobin and D. K. Sirota (eds.), *Malignant Tumors of the Thyroid*: *Clinical Concepts and Controversies*, 1992; M. Greer, *The Thyroid Gland*, 1990; M. H. Wheeler and J. H. Lazarus, *Thyroid Disease*, 1993.

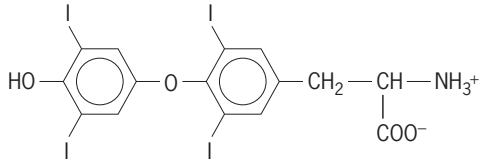
Thyroid hormones

Any of the chemical messengers produced by the thyroid gland. For many years it was generally believed that the thyroid gland produced only two closely related hormones, thyroxine and triiodothyronine. However, in 1961 a new hormonal principle, calcitonin, was discovered by Harold Copp. He originally thought that this substance was secreted by the parathyroid glands, but in 1963 it was established that the source of this hormone in mammals was the thyroid gland, and it was renamed thyrocalcitonin. In 1967 it was isolated in pure form and in 1968, synthesized. It is a polypeptide containing a single chain of 33 amino acids, in contrast to thyroxine and triiodothyronine which are iodinated thyronines. *See* HORMONE.

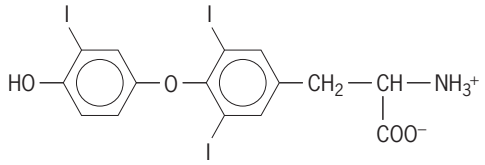
Thyroid gland. The thyroid gland is in reality two separate endocrine organs composed of two different cell types: acinar cells, manufacturing both thyroxine and triiodothyronine; and interstitial, parenchymal, or C cells, producing thyrocalcitonin. These two cell types differ in location, embryologic origin, histochemical characteristics, and control of the secretion of their respective hormones. The acinar cells are derived embryologically from the fourth branchial pouch; the parenchymal cells, from the fifth branchial pouch. In higher animals the parenchymal cells are enveloped by the lateral lobes

of the thyroid during subsequent development; in birds, reptiles, fish, and amphibia, they persist as separate organs, the ultimobranchial bodies which, rather than the thyroid gland, are the source of thyrocalcitonin. *See* THYROID GLAND.

Biochemistry and biosynthesis. Thyroxine (abbreviated T4) and triiodothyronine (T3) have very similar chemical structures (shown below) and are iodine-



Thyroxine



Triiodothyronine

containing derivatives of the amino acid tyrosine. [Thyrocalcitonin, with its entirely different function from T3 and T4, is discussed separately. *See* THYROCALCITONIN.] Because these two thyroid hormones are the only important biological compounds in higher animals that contain iodine and because radioisotopes of iodine are readily available to researchers, the biosynthesis, secretion, and distribution throughout the body of thyroid hormones has been extensively investigated and is well understood. *See* IODINE.

Ordinarily, cells manufacture chemical compounds of this small size using biochemical pathways. In a biochemical pathway, common metabolic intermediates (such as amino acids) are chemically modified in a series of enzyme-catalyzed steps until the desired compound is produced. Thyroxine and triiodothyronine, by contrast, are synthesized in a completely different and unusual manner. The acinar cells of the thyroid gland produce a large protein called thryoglobin, which contains approximately 600 amino acid residues. Iodine is chemically attached to approximately 30 of this protein's tyrosine residues, forming 2,5-diiodotyrosine residues. These iodinated tyrosine residues are then chemically linked to one another to form thyroxine-like and triiodothyronine-like residues within the thryoglobin protein. Lastly, the modified thyroglobin protein is broken down into many tiny fragments, most of which are degraded. Among the resulting fragments are the desired products, thyroxine and triiodothyronine. Five or six active hormone molecules are derived from a single thyroglobin protein.

The biosynthesis of these hormones is tightly regulated. The thyroid-stimulating hormone (TSH), released by the pituitary gland, controls how much thyroxine and triiodothyronine are produced and released into the bloodstream. Furthermore, the amount of iodine in one's diet influences the rate at which these hormones can be produced and released. *See* ADENOHYPOPHYSIS HORMONE; PITUITARY GLAND.

Mode of action. Thyroxine and triiodothryonine are nonpolar molecules that are not readily soluble in the aqueous environments of the bloodstream and cellular interiors. Consequently, upon release into the bloodstream from the thyroid gland, these hormones bind to a protein found in the blood called thyroxine-binding protein. The thyroxine-binding protein transports these hormones through the bloodstream to the hormone's target cells in the body. Another protein in the blood, serum albumin, can also transport these hormone molecules. *See* ALBUMIN.

Thyroid hormones are taken up by nearly all cells of the body, but the rate of uptake differs from cell type to cell type. Liver cells take up these hormones very rapidly, but many cells of the brain take up these hormones at nearly undetectable levels. Because of their nonpolar character, these hormones readily pass through the membranes of their target cells, where they bind to another thyroid hormone–binding protein in the cell's cytoplasm. This protein–hormone complex does not itself enter the cell's nucleus; instead it is believed to function in maintaining a reservoir of thryroid hormone within the cell.

In order to act upon the target cell, the thyroid hormone molecules leave the cytoplasmic binding protein and pass directly through the nuclear membrane. Once inside the nucleus, these hormone molecules bind to the thyroid hormone receptor protein, which is structurally related to the steroid hormone receptor proteins. The thyroid hormone receptor protein is bound to the chromosomal DNA at defined positions. When this protein binds to the hormone, it activates transcription of nearby genes that begins the production of a variety of metabolic enzymes.

It appears that the majority of the effects of thyroid hormones on cells comes from their ability to enter the cell's nucleus and stimulate the transcription of certain genes. However, there is evidence that these hormones also enter the mitochondria, the energy-producing organelles of the cell, and directly stimulate oxidative phosphorylation—a process through which adenosine triphosphate (ATP) is formed.

Physiological activities and abnormalities. The maintenance of a normal level of thyroxine is critically important for normal growth and development as well as for proper bodily function in the adult. The hormone is also critically important in amphibian metamorphosis. Its absence leads to delayed or arrested development. It is one of the few hormones with general effects upon all tissues. Its lack leads to a decrease in the general metabolism of all cells, characteristically measured as a decrease in nucleic acid and protein synthesis, and a slowing down of all major metabolic processes. These widespread and profound changes have been the subjects of intense investigative interest. *See* THYROXINE.

Two abnormalities have been described that involve either too little or too much thyroid hormone in the bloodstream. Hypothyroidism is an abnormality in which the levels of thyroid hormone

are too low. Individuals with hypothyroidism have a slowed metabolism and are generally lethargic and obese. Hyperthyroidism is an abnormality in which the levels of thyroid hormone are too high. Individuals with hyperthyroidism show the opposite effects, namely increased oxygen consumption, a speeding up of the metabolic processes, and hyperactivity. *See* THYROID GLAND DISORDERS.

M. Todd Washington; Howard Rasmussen

Bibliography. W. Green (ed.), *The Thyroid*, 1986, reprint 2001; J. E. Griffin and S. R. Ojeda (eds.), *Textbook of Endocrine Physiology*, 4th ed., 2000; D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 4th ed., 2004; D. Voet and J. G. Viet, *Biochemistry*, 3d ed., 2004.

Thysanoptera

An order of small, slender insects, commonly called thrips, having exopterygote development, sucking mouthparts, and highly modified wings. The order is a relatively small one, but individuals are often very numerous in favorable environments.

The mouthparts are conical and used for scraping, piercing, and sucking; flowers, leaves, and even twigs are attacked, resulting in severe injury to the epidermal cells of the plant. Vegetables and fruit trees are especially subject to damage by these insects. A very few species of thrips are predacious, feeding on mites and aphids.

The wings of these insects are exceptionally narrow, with few or no veins, and are bordered by long hairs. The tarsi terminate in an inflatable membranous bladder, which has remarkable adhesive properties.

The eggs of thrips are laid on the surface of twigs (suborder Tubulifera) or in small cuts made by the ovipositor (suborder Terebrantia). There are usually four nymphal stages, the last of these being quiescent and pupalike. There are from one to several generations produced in a single year. *See* INSECTA.

Frank M. Carpenter

Bibliography. N. F. Johnson and C. A. Triplehorn (eds.), *Borror and DeLong's Introduction to the Study of Insects*, 7th ed., 2004.

Thysanura (Zygentoma)

An order of wingless insects with soft, fusiform bodies 0.12–0.8 in. (3–20 mm) long, often covered with flat scales forming diverse patterns. The mouthparts are free with dicondylous mandibles used for scraping and chewing. Antennae are long and threadlike (filiform) with muscles present only in the scape and pedicel. Visual organs may be a cluster of simple ommatidia (cylinder-shaped units of a compound eye) or lacking altogether. The abdomen terminates in three "tails": a pair of lateral cerci and a median caudal filament (telson). Females have well-developed ovipositors; males have a penis and often one or rarely two pairs of parameres.

Classification. The order Thysanura (also called Zygentoma) includes about 400 described species which most taxonomists group into three families. The Lepidotrichidae, known as forest silverfish, are well represented in Oligocene fossils but were thought to be extinct until 1959, when living specimens were discovered in northwestern California. The single extant species lives in decaying bark and rotten wood. Members of the family Nicoletiidae are typically found underground in caves, mammal burrows, or the nests of ants or termites. Some species mimic ants and steal their food. The family Lepismatidae is worldwide in distribution. It encompasses more than 300 species of fast-running insects that feed as scavengers on algae, lichens, starchy vegetable matter, fungal hyphae (mycelium), and woodland debris. This family includes all of the common species found in human dwellings, notably the silverfish, *Lepisma saccharina*, and the firebrat, *Thermobia domestica*. Silverfish (sometimes known as fishmoths) are silvery gray, active at night, and often regarded as nuisance pests. Firebrats are usually tan or brown. These heat-tolerant insects prefer warm domestic habitats like kitchens, attics, and boiler rooms. They have been reported as serious pests in bakeries. Both species can cause extensive damage to household goods by feeding on stored food, cardboard packaging, wallpaper paste, book bindings, and the starch sizing of textiles.

Reproduction and development. Mating behavior of males and females involves external (indirect) fertilization. Males package sperm in desiccation-resistant spermatophores which they attach to the substrate. Insemination occurs when a female finds a spermatophore and draws it into her genital opening. Silverfish have an elaborate courtship ritual to ensure efficient exchange of sperm. The male spins a silken thread between the substrate and a vertical object. He deposits a spermatophore beneath this thread and then coaxes a female to walk under it. When her cerci contact the silk, she picks up the spermatophore with her genital opening. Sperm are released into her reproductive system, and then she ejects the empty spermatophore and eats it.

Thysanurans may be long-lived. A 3-year life-span is probably typical, and firebrats have been known to live up to 6 years. Development is ametabolous: other than an increase in size, there is no change in physical appearance from immatures (called young) to adults. These insects molt frequently, even as adults, and may complete more than 40 instars (stages between molts). This unusual adaptation could help reduce the risk of infection by parasitic fungi. *See* APTERYGOTA; INSECTA.

John Meyer

Bibliography. L. M. Ferguson, Insecta: Microcoryphia and Thysanura, pp. 935–949 in D. Dindal (ed.), *Soil Biology Guide*, Wiley, New York, 1990; N. F. Johnson and C. A. Triplehorn, (eds.), *Borror and DeLong's Introduction to the Study of Insects*, 7th ed., 2004; J. A. Wallwork, *Ecology of Soil Animals*, McGraw-Hill, New York, 1970; P. Wygodzinsky, A review of the silverfish (Lepismatidae, Thysanura) of the United States and Caribbean area, *Amer. Mus. Novit.*, 2481:1–26, 1972.

Tick paralysis

A loss of muscle function or sensation in humans or certain animals following the prolonged feeding of female ticks. Paralysis, of Landry's type, usually begins in the legs and spreads upward to involve the arms and other parts of the body. Evidence suggests that paralysis is due to a neurotoxin formed by the feeding ticks rather than the result of infection with microorganisms. *See* IXODIDES; TOXIN.

Present information indicates that (1) in endemic areas, resistance to tick paralysis is found in older animals as well as in certain animal species, (2) some animals which have recovered are not immune, (3) only occasional female, but not male, ticks may induce the disease under favorable host or environmental conditions, the specific requirements being as yet unknown, (4) paralysis of adult persons and domestic animals as large as a 1000-lb (450-kg) bull has resulted from only one partially engorged tick usually but not necessarily attached about the head or upper spine, (5) death may ensue if respiratory centers are reached by the ascending paralysis before the offending tick completes feeding or is removed, and (6) in the United States recovery is prompt, a matter of hours, when the ticks are removed. It is highly important, therefore, that the disease not be confused with infantile paralysis, that it be properly and promptly diagnosed, and that search for the tick be instituted immediately.

The disease has been reported in North America, Australia, South Africa, and occasionally in Europe, and is caused by appropriate species of indigenous ticks. In Australia, *Ixodes holocyclus* causes frequent cases in dogs, and occasionally in humans, and paralysis has been known to progress even after removal of ticks; serum from recovered dogs has been shown to have some curative properties and was at one time produced for treatment. *Ixodes cubicundus* is associated with the disease in South Africa.

Since 1903 over a hundred human cases and many outbreaks in cattle, sheep, and even domesticated bison have been recorded in the northwestern United States and southern British Columbia, due to attacks by *Dermacentor andersoni*. Several human fatalities, one in Idaho in June 1958, and some losses of stock have occurred when deticking was delayed. April to June are the months of most prevalence. Incidence is highest in children of 1–5 years of age, with more than twice as many girls affected as boys, presumably because their longer hair conceals feeding ticks. However, the sex ratio is reversed among the fewer cases in adults because of difference in exposure. Young to yearling stock are most prone to the disease in sporadic years for reasons still unknown.

The related American dog tick (*D. variabilis*) has paralyzed persons and dogs in the southeastern United States, and a few cases have been associated with the lone star tick (*Amblyomma americanum*).

The female tick requires 4–5 days of feeding before initial symptoms appear, and the disease progresses rapidly during the next 2–4 days. Experimental, fatal paralysis has been produced by *D. andersoni* females in woodchucks, ground squirrels, wood rats, hamsters, guinea pigs, dogs, and lambs. Signs of the disease occur within a few hours of transfer of partially fed females, but not males, to fresh animals. Isolation of the toxic principle has been claimed only by G. H. Kaire from the Australian tick (*I. holocyclus*) by chromatographic methods. Cornelius B. Philip

Bibliography. B. Fifaz, T. Petney, and I. Horak (eds.), *Tick Vector Biology: Medical and Veternarian Aspects*, 1993.

Tidal bore

A part of a tidal rise in a river which is so rapid that water advances as a wall often several feet high. The phenomenon is favored by a substantial tidal range and a channel which shoals and narrows rapidly upstream, but the conditions are so critical that it is not common. A shoaling channel steepens the tidal curve. If the curve becomes vertical or nearly so, a bore results (**Fig. 1**). A narrowing channel increases the tidal range. Since the tidal range is greatest at spring tides, some rivers exhibit bores only then. Although the bore is a very striking feature, Fig. 1 shows that the tide continues to rise after the passage of the bore and that this subsequent rise may be greater. Bores may be eliminated by changing channel depth or shape. *See* RIVER TIDES; TIDE.

In North America three bores have been observed: at the head of the Bay of Fundy (**Fig. 2**), at the head of the Gulf of California, and at the head of Cook Inlet, Alaska. The largest known bore occurs in the Tsientang Kiang, China. At spring tides this bore is a

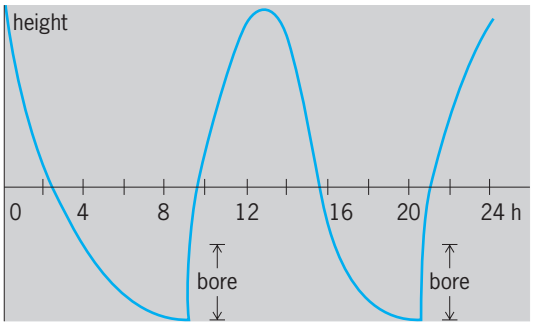


Fig. 1. Tidal curve of a river with a tidal bore.

Fig. 2. Tidal bore of the Petitcodiac River, Bay of Fundy, New Brunswick, Canada. Rise of water is about 4 ft (1. 2 m). (*New Brunswick Travel Bureau*)

wall of water 15 ft (4.5 m) high moving upstream at 25 ft/s (7.5 m/s). *See* OPEN CHANNEL. Blair Kinsman

Tidal datum

A reference elevation of the sea surface from which vertical measurements are made, such as depths of the ocean and heights of the land. The intersection of the elevation of a tidal datum with the sloping shore forms a line used as a horizontal boundary. In turn, this line is also a reference from which horizontal measurements are made for the construction of additional coastal and marine boundaries.

Since the sea surface moves up and down from infinitely small amounts to hundreds of feet over periods of less than a second to millions of years, it is necessary to stop the vertical motion in order to have a practical reference. This is accomplished by hydraulic filtering, numerical averaging, and segment definition of the record obtained from a tide gage affixed to the adjacent shore. Waves of periods up through wind waves are effectively damped by a restricting hole in the measurement well. Recorded hourly heights are averaged to determine the mean of the higher (or only) high tide of each tidal day (24.84 h), all the high tides, all the hourly heights, all the low tides, and the lower (or only) low tide. The length of the averaging segment is a specific 19 year, which averages all the tidal cycles through the regression of the Moon's nodes and the metonic cycle. [The metonic cycle is a time period of 235 lunar months (19 years); after this period the phases of the Moon occur on the same days of the same months.] But most of all, the 19-year segment is meaningful in terms of measurement capability, averaging meteorological events, and for engineering and legal interests. However, the 19-year segment must be specified and updated because of sea-level changes occurring over decades. The present sea-level epoch is 1960 through 1978. It has been updated about every 20 years.

The legally recognized tidal datums of the United States are Mean Higher High Water (MHHW), Mean High Water (MHW), and Mean Lower Low Water (MLLW). Depths of the ocean, especially in harbors and coastal regions, are measured from the tidal datum of Mean Lower Low Water, called Chart Datum, and printed on nautical charts. This lowest of the ordinary tidal datums was chosen as a safety factor for early mariners who might not have been in possession of tide predictions. Later, as predictions became easily available, the predictions at any time could be added to the printed depths for harbor access and safety.

Mean Sea Level (MSL), obtained from averaging all the hourly heights, is used to monitor and study apparent secular trends in sea level.

Over the decades, land elevations have been based on a geodetic network supported by the tidal datums of several groupings of tide stations. They are now computed within the network of the North American (Canada, United States, Mexico) Vertical Datum of 1988. This network is attached to the tidal datum at only one tide station, Father Point, Quebec, Canada.

Tidal datums are legal entities. Because of variations in gravity, semistationary meteorological conditions, semipermanent ocean currents, changes in tidal characteristics, ocean density differences, and so forth, the sea surface (at any datum elevation) does not conform to a mathematically defined spheroid. Therefore, a series of tide stations along the marine and Great Lakes coastlines of the United States is necessary in order to provide vertical continuity over horizontal distances. *See* GEODESY; TIDE.

Steacy D. Hicks

Bibliography. G. W. Groves, Numerical filters for discrimination against tidal periodicities, *Trans. Amer. Geophys. Union*, 36(6):1073–1084, 1955; S. D. Hicks, *Tide and Current Glossary*, NOAA, 1989; E. Lisitzin and J. Pattullo, The principal factors influencing the seasonal oscillations of sea level, *J. Geophys. Res.*, 66:845–853, 1961; H. A. Marmer, *Tidal Datum Planes*, 5th ed., USCGS Spec. Publ. 135, 1951; J. Pattullo et al., The seasonal oscillations in sea level, *J. Mar. Res.*, 14:88–156, 1955; J. R. Rossiter, Note on methods of determining monthly and annual values of mean water level, *Int. Hydrogr. Rev.*, May 1958.

Tidal power

Tidal-electric power is obtained by using the recurring rise and fall of coastal waters in response to the gravitational forces of the Sun and the Moon. Marginal marine basins are enclosed with dams, making it possible to create differences in the water level between the ocean and the basins. The oscillatory flow of water filling or emptying the basins is used to drive hydraulic turbines that propel electric generators.

Electric power could be developed in the world's coastal regions having tides of sufficient range, although even if fully developed this would amount to only a small percentage of the world's potential water (hydroelectric) power. Nevertheless, tidal-electric power could become locally important, particularly because it produces no air or thermal pollution, consumes no exhaustible resource, and produces relatively minor impacts on the environment.

The use of ocean tides for power purposes dates back to the tidal mills in Europe during the Middle Ages and to those in America during colonial times. The first tidal development producing electric power in operation, the Rance development in northwestern France, was completed in 1967. It has an installed capacity of 240,000 kW in 24 units and is capable of producing about 500×10^6 kWh annually.

Tidal range is measured as the difference in level between the successive high and low waters. Although there are variations at certain locations in the intervals between successive high tides, at most places the tides reach the highest levels at intervals of about 12 h 25 min. The tidal ranges vary from

day to day. The highest tides, known as spring tides, occur twice monthly near the time of the new moon and the full moon when the Sun and Moon are in line with the Earth. The lowest tides, known as neap tides, occur midway between the spring tides when the Sun and Moon are at right angles with the Earth. The highest spring tides occur near the time of the equinoxes in the spring and fall of the year. Except for variations caused by meteorological changes, the tides are predictable and follow similar patterns from year to year.

Large tidal ranges occur when the oscillation of the ocean tides is amplified by relatively shallow bays, inlets, or estuaries. There are limited locations where the tidal ranges are sufficiently large to be considered favorable for power development. The largest tidal ranges in the world, reaching a maximum of over 50 ft (15 m), are said to occur in the Bay of Fundy in Canada. Other locations with large maximum tidal ranges are the Severn Estuary in Britain, 45 ft (14 m); the Rance Estuary in France, 40 ft (12 m); Cook Inlet in Alaska, 33 ft (10 m); and the Gulf of California in Mexico, 30 ft (9 m). Large tidal ranges also occur at locations in Argentina, India, Korea, Australia, and on the northern coast of Russia. *See* ELECTRIC POWER GENERATION; TIDE.

George G. Adkins

Bibliography. American Society of Civil Engineers, *Civil Engineering Guidelines for Planning and Designing Hydroelectric Developments*, vol. 5: *Pumped Storage and Tidal Power*, 1989; American Society of Civil Engineers, *Developments in Tidal Energy*, 1990; American Society of Civil Engineers, *Tidal Power: Symposium Proceedings*, 1987; G. Boyle, *Renewable Energy*, 2d ed., 2004.

Tidalites

Sediments deposited by tidal processes. Until recently, "tidalites" referred to sediments deposited by tidal processes in both the intertidal zone (between normal low- and high-tide levels) and shallow,

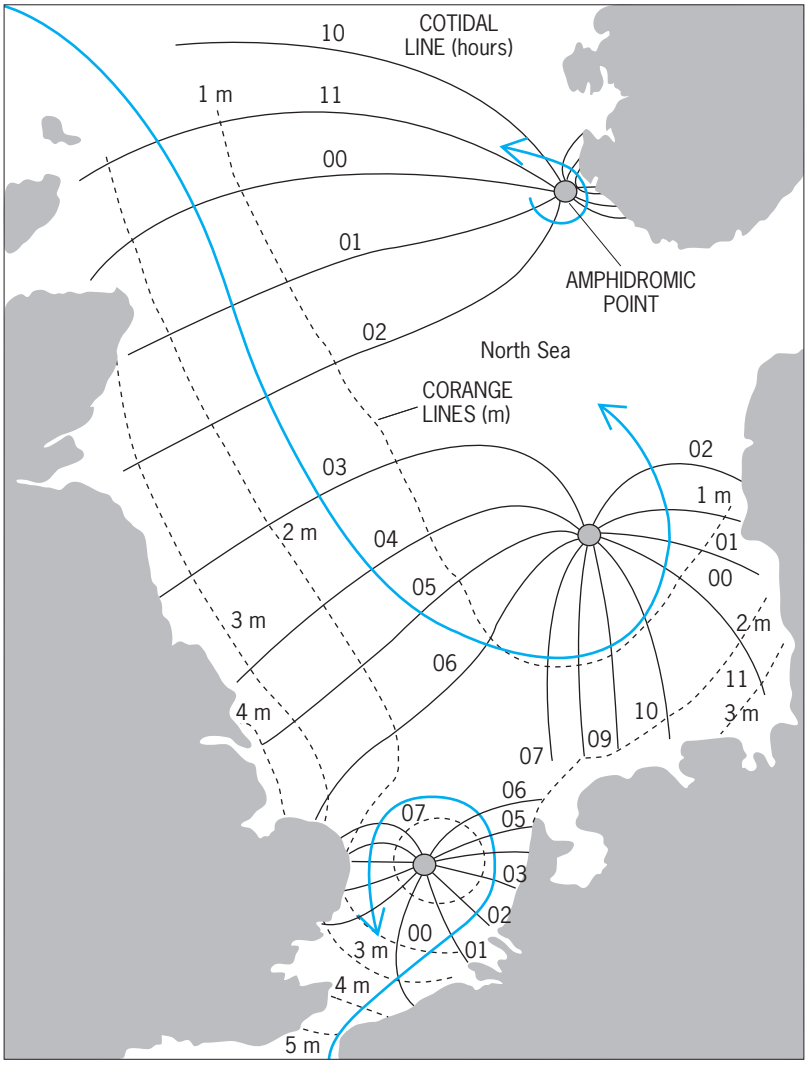


Fig. 1. North Sea amphidromic tidal system. Corange lines indicate equal tidal range. Cotidal lines show times of high water. Arrows show rotation directions of the tidal waves. (*Modified from R. W. Dalrymple, Tidal Depositional Systems, in R. G. Walker and N. P. James, eds., Facies Models Response to Sea Level Changes, pp. 195–218, Geological Association of Canada, 1992*)

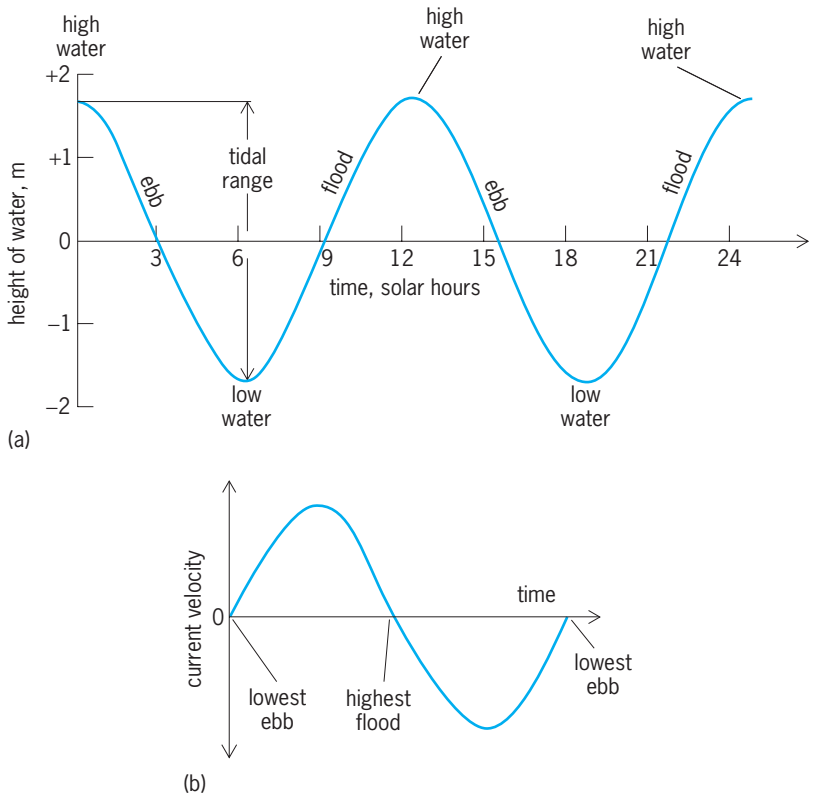


Fig. 2. Flood–ebb cycle. Idealized (*a*) semidiurnal tidal cycle and (*b*) time–current velocity curve. (*Modified from S. D. Nio and C. S. Yang, Recognition of tidally influenced facies and environments, Short Course Note Ser. 1, International Geoservices BV, Leiderdorp, Netherlands, 1989*)

subtidal (permanently submerged), tide-dominated environments less than 200 m (660 ft) deep. Tidalites are now known also to occur within supratidal environments (above normal high tide and flooded only during storms or very high spring tides) and submarine canyons at depths much greater than 200 m. Common usage has drifted toward describing tidalites as ripple- and dune-scale features rather than more composite deposits such as large linear sand ridges of tidal origin present on continental shelves or point bars associated with migrating tidal channels. Both of these larger-scale features, however, would be composed of tidalites.

Recognition criteria. By identifying tidalites in either the modern or the ancient geological record, geologists are implying that they know that the sediments were deposited by tidal processes rather than by storms or waves. Tidalites are not always easy to identify with certainty, especially in the rock record. In order to do so, it is necessary to understand the basic tidal cycles that can influence sedimentation.

Tidal theory. Tides are generated by the combined gravitational forces of the Moon and Sun on the Earth's oceans. Some sources are misleading in suggesting that the tidal forces from the Moon and Sun, in combination with centrifugal forces associated with the spin of the Earth, produce oceanic bulges on opposite sides of the Earth. While it is true that the combined gravitational forces of the Moon and, to a lesser extent, the Sun produce tides on the Earth, the Earth does not spin through two oceanic bulges that form on opposite sides of the Earth. This conceptual model has little bearing on real-world tides. Rather, water within each of the Earth's ocean basins is forced to rotate as discrete waves about a series of fixed (amphidromic) points (**Fig. 1**). For a fixed point along an ocean coastline, a tidal system is referred to as diurnal if it experiences the passing of the resultant tidal wave once every 24 h 50 min. The tidal system is semidiurnal if the resultant tidal wave passes the fixed point twice during the same time.

In the open ocean, the motion of a tidal wave is largely expressed as a vertical movement of water masses. In shallow basins along the coast, water movements are more horizontal, with tides moving in and out of estuaries and embayments, resulting in a change in water level as the tidal wave passes. The daily or semidaily rise in tides is called the flood tide, and the fall is referred to as the ebb tide (**Fig. 2***a*). Tidal currents are maximized between flood and ebb tides and minimized at highest flood or lowest ebb tides (Fig. 2b). The difference between the high tide and the low tide is called the tidal range.

The intensity or height of the daily or twice-daily tides can vary in a number of ways. Cyclic semimonthly changes in daily tidal heights associated with neap–spring tidal cycles are the most pronounced of these. Spring tides occur twice a month when the tidal range is greatest, and neap tides occur twice a month when tidal range is least. Neap–spring

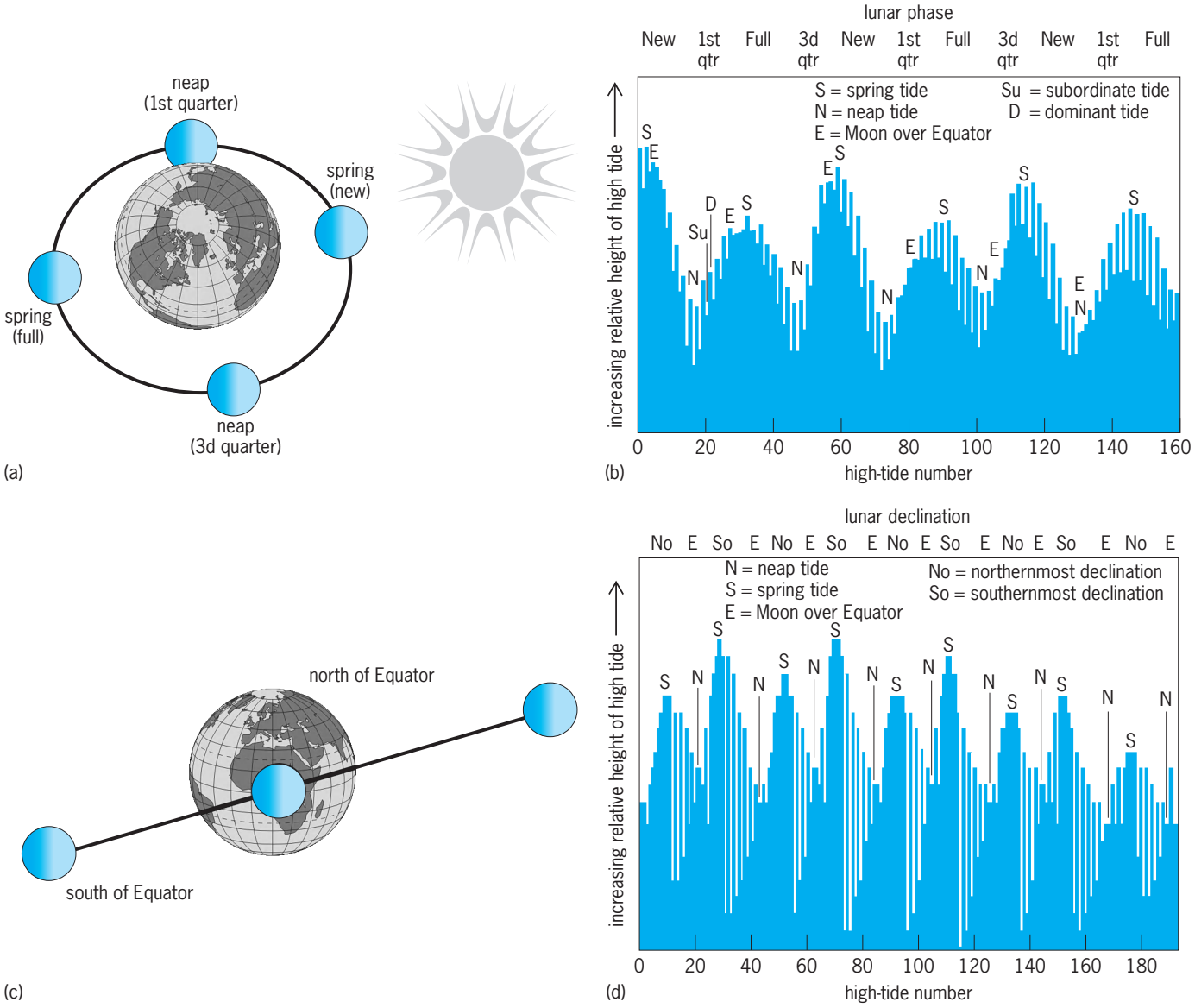


Fig. 3. Idealized models of origin of neap–spring tidal cycles: (*a*) Synodic month. (*b*) A segment of the 1991 predicted high tides from Kwajalein Atoll, Pacific. (*c*) Tropical month. (*d*) A segment of the 1994 predicted high tides from Barito River estuary, Borneo. (*Modified from E. P. Kvale, K. H. Sowder, and B. T. Hill, Modern and ancient tides: Poster and explanatory notes, Society for Sedimentary Geology, Tulsa, OK, and Indiana Geological Survey, Bloomington, IN, 1998*)

cycles can be generated in two ways. The familiar neap–spring cycle is related to the phase changes of the Moon. Spring tides occur every 14.76 days when the Earth, Moon, and Sun are nearly aligned at new or full moon (**Fig. 3***a*). Neap tides occur when the Sun and Moon are aligned at right angles from the Earth at first- and third-quarter phases of the Moon. The result is that spring tides are higher than neap tides (Fig. 3*b*). The time from new moon to new moon is called the synodic month, which has a modern period of 29.53 days. This type of neap–spring cycle is referred to as synodically driven, and it dominates the coastlines of western Europe and the eastern coastline of North America.

A second type of neap–spring cycle is less well known but no less common, and is related to the orbit of the Moon around the Earth. The Moon's orbital plane is inclined relative to the Earth's equatorial plane. The period of the variation in lunar declination relative to the Earth's Equator is called the tropical month, and is the time the Moon takes to complete one orbit, moving from its maximum northerly declination to its maximum southerly declination and return (Fig. 3*c*). In this type of neap–spring cycle, the tidal force depends on the position of the Moon relative to the Earth's Equator. The tide-raising force at a given location is greater when the Moon is at its maximum declination every 13.66 days. These periods correspond to the generation of spring tides (Fig. 3*d*). The neap tides occur when the Moon is over the Earth's Equator. The current length of the tropical month is 27.32 days, and neap–spring cycles in phase with the tropical month are referred to as tropically driven. These types of neap-spring cycles dominate coastlines in the Gulf of Mexico and many areas in the Pacific.

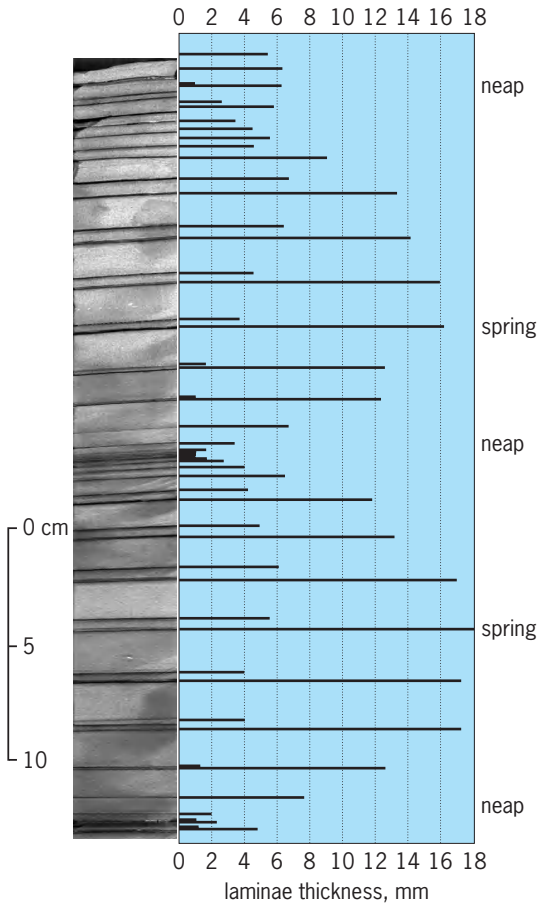


Fig. 4. The core shows small-scale tidalites from the Hindostan Whetstone beds, Mansfield Formation, Indiana. The chart shows thicknesses of layers as measured between dark clay-rich bands. The interval shows approximately one synodic month of deposition.

Besides generating neap–spring cycles in many parts of the world, the changing position of the Moon relative to the Earth's Equator through the tropical month causes the diurnal inequality of the tides in semidiurnal tidal systems. In tidal systems that experience two high tides and two low tides per day, the tropical monthly cycle results in the morning high tide being greater or lesser than the evening high tide. The diurnal inequality is reduced to zero when the Moon is over the Equator, resulting in the morning tide and the evening tide being of equal magnitude (Fig. 3*b* and *d*).

Other tidal cycles besides those mentioned above can influence sedimentation and have been documented in the geologic record. These include monthly, semiannual, and multiyear tidal cycles. *See* EARTH ROTATION AND ORBITAL MOTION; MOON; TIDE.

Examples of tidalites. How the various tidal cycles manifest themselves in the geologic record and how geologists can identify their influence on sedimentation has been studied for nearly 75 years. To recognize tidalites in the geologic record, geologists look for evidence of one or more of the following:

1. Sediment deposited by reversing currents (that is, flood–ebb cycles).

2. A stacked sequence of sediments that show a recurring change from sediments transported (and deposited) by currents at maximum current velocity to sediments deposited from suspension at minimum current velocity (Fig. 2*b*).

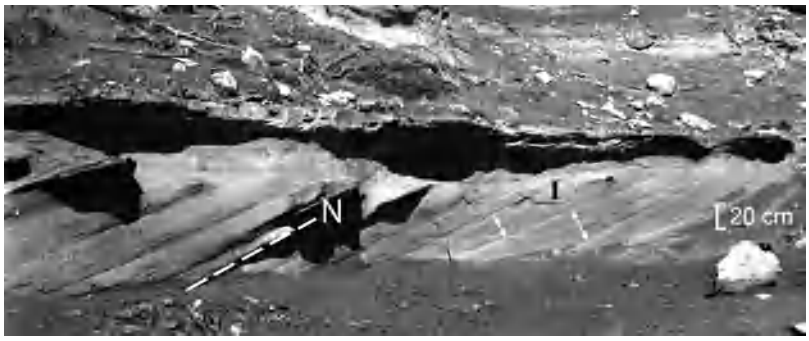
3. Stacked packages of sediments in which each package shows evidence of subaerial exposure superimposed on sediments deposited in subaqueous settings (sediments transported and deposited during flood tides and exposed during low ebb tide).

4. A sequence of sediment packages in which the thickness or accretion of successive packages of sediments varies in a systematic way, suggesting diurnal, semidiurnal, and/or neap–spring tidal cycles.

An example of a small-scale tidalite can be found in the Mansfield Formation (Pennsylvanian Period) in Orange County, Indiana (**Fig. 4**). The sample shown is from a rock core taken through this interval. The lighter-colored layers are siltstone, and the thin dark layers are finer-grained claystone. The regular and repeating change in deposition from siltstone to claystone indicates systematic current velocity fluctuations related to the tidal cycle over a 12-h period (see item 2 above). The thick–thin pairing of the lighter bands of siltstone suggests the influence of the semidiurnal inequality over a 24-h period (see item 4). In addition, the regular and systematic overall thickening and thinning of the siltstone layers, as shown in the bar chart next to the core in Fig. 4, suggests that neap–spring tidal cycles controlled the thicknesses of the silt layers. The higher spring tides resulted in thicker accumulations of silt than the lower neap tides. *See* PENNSYLVANIAN; SEDIMENTARY ROCKS.

An example of a large-scale tidalite can be found in the Jurassic Sundance Formation of northern Wyoming (**Fig. 5**). This tidalite is the remnant of a migrating subtidal dune or sandwave. The preserved inclined beds of the avalanche face (I) indicate the migration direction of the dune from right to left (Fig. 5*a*). The evidence for tidal influence, however, lies within the inclined, less resistant, and more recessed lighter-colored bands (examples marked by arrows in Fig. 5*a*). In this interval (Fig. 5*b*), one sees evidence of (1) cessation of dune migration and a reversal of current direction from flood tide to ebb tide with small ripples migrating up the avalanche face (II); (2) a mud drape (III) resulting from fine-grained materials settling out of suspension when the current velocities reached zero as the tide reversed (Fig. 2); and (3) a reactivation of the migrating dune above the mud drape (IV) as current velocity increased during the next flood tide. The right-to-left migration of the dune was also controlled by the neap–spring cycle, with greater migration (interval between lighter-colored bands) occurring during spring tides and lesser migration during neap tides (Fig. 5*a*). In the example shown, the neap tide deposits are centered on line N. *See* JURASSIC.

Geologic record. Deposits of tidalites are known from every geologic period from the modern back into the Precambrian and from depositional environments with water chemistries ranging from fresh to hypersaline. Studies of tidalites are important



(a)

(b)

Fig. 5. Photographs from Sundance Formation of northern Wyoming. (*a*) Example of large-scale tidalites. (*b*) Closeup of inclined light-colored band showing evidence of current reversals.

because geologists have used these features not only to interpret the original depositional settings of the deposits but also to calculate ancient Earth–Moon distances, interpret paleoclimates existent during deposition, and calculate sedimentation rates. *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; GEOLOGIC TIME SCALE; MARINE SEDIMENTS; SEDIMENTOLOGY.

Erik P. Kvale

Bibliography. C. Alexander, R. Davis, and V. Henry (eds.), *Tidalites: Processes and Products*, Geological Society Publishing, 1998; D. E. Cartwright, *Tides: A Scientific History*, Cambridge University Press, 1998; G. deV. Klein, A sedimentary model for determining paleotidal range, *Geol. Soc. Amer. Bull.*, 82:2585–2592, 1971; G. deV. Klein, Determination of paleotidal range in clastic sedimentary rocks, *XXIV International Geological Congress*, 6:397–405, 1972; D. T. Pugh, *Tides, Surges and Mean Sea Level*, Wiley, 1987; H. G. Reading (ed.), *Sedimentary Environments: Processes, Facies and Stratigraphy*, 3d ed., Blackwell Science, Cambridge, MA, 1996.

Tide

Stresses exerted in a body by the gravitational action of another, and related phenomena resulting from these stresses. Every body in the universe raises tides, to some extent, on every other. This article deals only with tides on the Earth, since these are fundamentally the same as tides on all bodies. Sometimes variations of sea level, whatever their origin, are referred to as tides.

Introduction. The tide-generating forces arise from the gravitational action of Sun and Moon, the effect of the Moon being about twice as effective as that of the Sun in producing tides. The tidal effects of all other bodies on the Earth are negligible. The tidal forces act to generate stresses in all parts of the Earth and give rise to relative movements of the matter of the solid Earth, ocean, and atmosphere. The Earth's rotation gives these movements an alternating character having principal periodicities of 12.42 and 12.00 h, corresponding to half the mean lunar and solar day, respectively.

In the ocean, the tidal forces act to generate alternating tidal currents and displacements of the sea surface. These phenomena are important to shipping and have been studied extensively. The main object of tidal studies has been to predict the tidal elevation or current at a given seaport or other place in the ocean at any given time.

The prediction problem may be attacked in two ways. Since the relative motions of Earth, Moon, and Sun are known precisely, it is possible to specify the tidal forces over the Earth at any past or future time with great precision. It should be possible to relate tidal elevations and currents at any point in the oceans to these forces, making use of classical mechanics and hydrodynamics. Such a theoretical approach to tidal prediction has not yet yielded any great success, owing in great part to the complicated shape of the ocean basins. However, use of numerical-hydrodynamical models (such as the work of K. T. Bogdanov, N. Grijalva, W. Hansen, M. C. Henderschott, and C. L. Pekeris) has yielded some satisfactory results and undoubtedly will have practical importance.

The other approach, which consists of making use of past observations of the tide at a certain place to predict the tide for the same place, has yielded practical results. The method cannot be used for a location where there have been no previous observations. In the harmonic method the frequencies of the many tidal constituents are derived from knowledge of the movements of Earth, Moon, and Sun. The amplitude and epoch of each constituent are determined from the tidal observations. The actual tide can then be synthesized by summing up an adequate number of harmonic constituents. The method might loosely be thought of as extrapolation.

A "convolution" method of tidal analysis and prediction has been proposed by W. H. Munk and D. E. Cartwright. In this method past observations at a place are used to determine a numerical operator which, when applied to the known tide-producing forces, will calculate the resulting tide.

In the following discussion only the lunar effect is considered, and it is understood that analogous statements apply to the solar effect.

Tide-generating force. If the Moon attracted every point within the Earth with equal force, there would

be no tide. It is the small difference in direction and magnitude of the lunar attractive force, from one point of the Earth's mass to another, which gives rise to the tidal stresses.

According to Newton's laws, the Moon attracts every particle of the Earth with a force directed toward the center of the Moon, with magnitude proportional to the inverse square of the distance between the Moon's center and the particle. At point A in **Fig. 1**, the Moon is in the zenith and at point B the Moon is at nadir. It is evident that the upward force of the Moon's attraction at A is greater than the downward force at B because of its closer proximity to the Moon. Such differential forces are responsible for stresses in all parts of the Earth. The Moon's gravitational pull on the Earth can be expressed as the vector sum of a constant force, equal to the Moon's attraction on the Earth's center, and a small deviation which varies from point to point in the Earth (**Fig. 2**). This small deviation is referred to as the tide-generating force. The larger constant force is balanced completely by acceleration (centrifugal force) of the Earth in its orbital motion around the center of mass of the Earth-Moon system, and plays no part in tidal phenomena. *See* GRAVITATION.

The tide-generating force is proportional to the mass of the disturbing body (Moon) and to the inverse cube of its distance. This inverse cube law accounts for the fact that the Moon is 2.17 times as important, insofar as tides are concerned, as the Sun, although the latter's direct gravitational pull on the Earth, which is governed by an inverse-square law, is about 180 times the Moon's pull.

The tide-generating force, as illustrated in Fig. 2, can be expressed as the gradient of the tide-generating potential, Eq. (1), where λ is the zenith

$$\psi = \frac{3}{2}\frac{\gamma M r^2}{c^3}\left(\frac{1}{3} - \cos^2\lambda\right) \qquad (1)$$

distance of the Moon, r is distance from the Earth's center, c is distance between the centers of Earth and Moon, γ is the gravitational constant, and M is the mass of the Moon. In this expression, terms containing higher powers of the smaller number r/c have been neglected. As ψ depends only on the space variables r and λ, it is symmetrical about the Earth-Moon axis.

It helps one visualize the form of the tide-generating potential to consider how a hypothetical "inertialess" ocean covering the whole Earth would respond to the tidal forces. In order to be in equilibrium with the tidal forces, the surface must assume the shape of an equipotential surface as determined by both the Earth's own gravity and the tide-generating force. The elevation of the surface is given approximately by Eq. (2), where ψ is evaluated

$$\bar{\zeta} = -\frac{\psi}{g} + \text{const} \qquad (2)$$

at the Earth's surface and g is the acceleration of the Earth's gravity. The elevation ζ of this hypothetical ocean is known as the equilibrium tide. Knowledge of the equilibrium tide over the entire Earth determines completely the tide-generating potential (and hence the tidal forces) at all points within the Earth as well as on its surface. Therefore, when the equilibrium tide is mentioned, it shall be understood that reference to the tide-generating force is also being made.

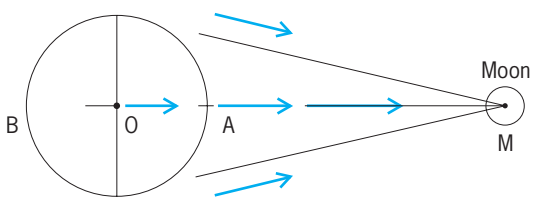


Fig. 1. Schematic diagram of the lunar gravitational force on different points in the Earth.

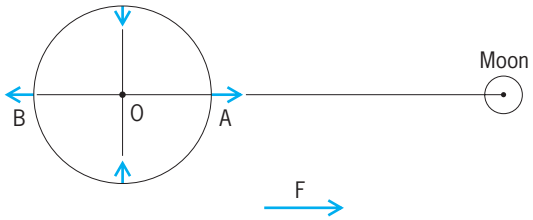


Fig. 2. Schematic diagram of the tide-generating force on different points in the Earth. The vector sum of this tide-generating force and the constant force *F* (which does not vary from point to point) produce the force field indicated in Fig 1. Force *F* is compensated by the centrifugal force of the Earth in its orbital motion.

Harmonic development of the tide. The equilibrium tide as determined from relations (1) and (2) has the form of a prolate spheroid (football-shaped) whose major axis coincides with the Earth-Moon axis. The Earth rotates relative to this equilibrium tidal form so that the nature of the (equilibrium) tidal variation with time at a particular point on the Earth's surface is not immediately obvious. To analyze the character of this variation, it is convenient to express the zenith angle of the Moon in terms of the geographical coordinates θ, ϕ of a point on the Earth's surface (θ is colatitude, ϕ is east longitude) and the declination D and west hour angle reckoned from Greenwich α of the Moon. When this is done, the equilibrium tide can be expressed as the sum of the three terms in Eq. (3), where a is the Earth's radius.

$$\begin{aligned}\bar{\zeta} = \frac{3}{4}\frac{\gamma M}{g}\frac{a^2}{c^3}[&(3\sin^2 D - 1)(\cos^2\theta - {}^1/_3) \\ &+ \sin 2D \sin 2\theta \cos(\alpha + \phi) \\ &+ \cos^2 D \sin^2\theta \cos 2(\alpha + \phi)]\end{aligned} \qquad (3)$$

The first term represents a partial tide which is symmetrical about the Earth's axis, as it is independent of longitude. The only time variation results from the slowly varying lunar declination and distance from Earth. This tide is called the long-period tide. Its actual geographical shape is that of a spheroid whose axis coincides with the Earth's axis and whose oblateness slowly but continuously varies.

The second term of Eq. (3) represents a partial tide having, at any instant, maximum elevations at 45°N and 45°S on opposite sides of the Earth, and two minimum elevations lying at similar, alternate positions on the same great circle passing through the poles. Because of the factor cos $(\alpha + \phi)$ the tide rotates in a westerly direction relative to the Earth, and any geographical position experiences a complete oscillation in a lunar day, the time taken for α to increase by the amount 2π. Consequently, this partial tide is called the diurnal tide. Because of the factor sin $2D$, the diurnal equilibrium tide is zero at the instant the Moon crosses the Equator; because of the factor sin 2θ, there is no diurnal equilibrium tidal fluctuation at the Equator or at the poles.

The third term of Eq. (3) is a partial tide having, at any instant, two maximum elevations on the Equator at opposite ends of the Earth, separated alternately by two minima also on the Equator. This whole form also rotates westward relative to the Earth, making a complete revolution in a lunar day. But any geographic position on the Earth will experience two cycles during this time because of the factor cos $2(\alpha + \phi)$. Consequently, this tide is called the semidiurnal tide. Because of the factor $\sin^2 \theta$, there is no semidiurnal equilibrium tidal fluctuation at the poles, while the fluctuation is strongest at the Equator.

It has been found very convenient to consider the equilibrium tide as the sum of a number of terms, called constituents, which have a simple geographical shape and vary harmonically in time. This is the basis of the harmonic development of the tide. A great number of tidal phenomena can be adequately described by a linear law; that is, the effect of each harmonic constituent can be superimposed on the effects of the others. Herein is the great advantage of the harmonic method in dealing with tidal problems. The three terms of Eq. (3) do not vary with time in a purely harmonic manner. The parameters c and D themselves vary, and the rapidly increasing α does not do so at a constant rate owing to ellipticity and other irregularities of the Moon's orbit. Actually, each of the three partial tides can be separated into an entire species of harmonic constituents. The constituents of any one of the three species have the same geographical shape, but different periods, amplitudes, and epochs.

The solar tide is developed in the same way. As before, the three species of constituents arise: long-period, diurnal, and semidiurnal. The equilibrium tide at any place is the sum of both the lunar and solar tides. When the Sun and Moon are nearly in the same apparent position in the sky (new Moon) or are nearly at opposite positions (full Moon), the lunar and solar effects reinforce each other. This condition is called the spring tide. During the spring tide the principal lunar and solar constituents are in phase. At quadrature the solar effect somewhat cancels the lunar effect, the principal lunar and solar constituents being out of phase. This condition is known as the neap tide.

The entire equilibrium tide can now be expressed by Eq. (4), where $H = 3\gamma Ma^2/gc^3 = 54$ cm, and

$$\overline{\zeta} = H\Big[{}^1\!/_2(1 - 3\cos^2\theta)\sum_L f_i C_i \cos A_i + \sin 2\theta \sum_D f_i C_i \cos(A_i + \phi) + \sin^2\theta \sum_S f_i C_i \cos(A_i + 2\phi)\Big] \qquad (4)$$

$1/c$ represents the mean (in time) value of $1/c$. Each term in the above series represents a constituent. Terms of higher powers of the Moon's parallax (a/c) are not included in Eq. (4) because of their different latitude dependence, but they are of relatively small importance. The subscripts L, D, and S indicate summation over the long-period, diurnal, and simidiurnal constituents, respectively. The C's are the constituent coefficients and are constant for each constituent. They account for the relative strength of all lunar and solar constituents. In a purely harmonic development, such as carried out by A. T. Doodson in 1921, the A parts of the arguments increase linearly with time, and the node factors f are all unity. In George Darwin's "almost harmonic" development of 1882, the constituents undergo a slow change in amplitude and epoch with the 19-year nodal cycle of the Moon. The node factors f take this slow variation into account. The A's increase almost linearly with time. Tables in U.S. Coast and Geodetic Survey Spec. Publ. 98 enable one to compute the phase of the argument of any of Darwin's constituents at any time, and values of the node factors for each year are given.

In spite of the many advantages of the purely harmonic development, Darwin's method is still used by most agencies engaged in tidal work. In Darwin's classification, each constituent is represented by a symbol with a numerical subscript, 0, 1, or 2, which designates whether the constituent is long-period, diurnal, or semidiurnal. Some of the most important of Darwin's constituents are listed in the **table**.

The periods of all the semidiurnal constituents are grouped about 12 h, and the diurnal periods about 24 h. This results from the fact that the Earth rotates much faster than the revolution of the Moon about the Earth or of the Earth about the Sun. The principal lunar semidiurnal constituent M_2 beats against the others giving rise to a modulated semidiurnal

Darwin's constituents

Constituent	Speed, deg/h	Coefficient
Long-period		
Mf, lunar fortnightly	1.098	0.157
Ssa, solar semiannual	0.082	0.073
Diurnal		
K_1, lunisolar	15.041	0.530
O_1, larger lunar	13.943	0.377
P_1, larger solar	14.959	0.176
Semidiurnal		
M_2, principal lunar	28.984	0.908
S_2, principal solar	30.000	0.423
N_2, larger lunar elliptic	28.440	0.176
K_2, lunisolar	30.082	0.115

waveform whose amplitude varies with the Moon's phase (the spring-neap effect), distance, and so on. Similarly, the amplitude of the modulated diurnal wave varies with the varying lunar declination, solar declination, and lunar phase. For example, the spring tide at full Moon or new Moon is manifested by constituents M_2 and S_2 being in phase, thus reinforcing each other. During the neap tide when the Moon is at quadrature, the constituents M_2 and S_2 are out of phase, and tend to cancel each other. The other variations in the intensity of the tide are similarly reflected in the "beating" of other groups of constituents.

Tides in the ocean. The tide in the ocean deviates markedly from the equilibrium tide, which is not surprising if one recalls that the equilibrium tide is based on neglect of the inertial forces. These forces are appreciable unless the periods of all free oscillations in the ocean are small compared with those of the tidal forces. Actually, there are free oscillations in the ocean (ordinary gravity seiches) having periods of the order of a large fraction of a day, and there may be others (planetary modes) having periods of the order of several days. For the long-period constituents the observed tide should behave like the equilibrium tide, but this is difficult to show because of their small amplitude in the presence of relatively large meteorological effects.

At most places in the ocean and along the coasts, sea level rises and falls in a regular manner. The highest level usually occurs twice in any lunar day, the times bearing a constant relationship with the Moon's meridional passage. The time between the Moon's meridional passage and the next high tide is called the lunitidal interval. The difference in level between successive high and low tides, called the range of the tide, is generally greatest near the time of full or new Moon, and smallest near the times of quadrature. This results from the spring-neap variation in the equilibrium tide. Tide range usually exhibits a secondary variation, being greater near the time of perigee (when the Moon is closest to the Earth) and smaller at apogee (when it is farthest away).

The above situation is observed at places where the tide is predominantly semidiurnal. At many other places, it is observed that one of the two maxima in any lunar day is higher than the other. This effect is known as the diurnal inequality and represents the presence of an appreciable diurnal variation. At these places, the tide is said to be of the "mixed" type. At a few places, the diurnal tide actually predominates, there generally being only one high and low tide during the lunar day.

Both observation and theory indicate that the ocean tide can generally be considered linear. As a result of this fact, the effect in the ocean of each constituent of the series in Eq. (4) can be considered by itself. Each equilibrium constituent causes a reaction in the ocean. The tide in the ocean is the sum total of all the reactions of the individual constituents. Furthermore, each constituent of the ocean tide is harmonic (sinusoidal) in time. If the amplitude of an equilibrium constituent varies with the nodal cycle of the Moon, the amplitude of the oceanic constituent varies proportionately.

As a consequence of the above, the tidal elevation in the ocean can be expressed by Eq. (5),

$$\zeta = \sum f_i h_i \cos (A_i - G_i) \tag{5}$$

where $h_i(\theta,\phi)$ is called the amplitude and $G_i(\theta,\phi)$ the Greenwich epoch of each constituent. The summation in Eq. (5) extends over all constituents of all species. The f's and the A's have the same meaning as in Eq. (4) for the equilibrium tide and are determined from astronomic data.

To specify completely the tidal elevation over the entire surface of the ocean for all time, one would need ocean-wide charts of $h(\theta,\phi)$, called corange charts, and of $G(\theta,\phi)$, called cotidal charts, for each important constituent. Construction of these charts would solve the ultimate problem in tidal prediction. Many attempts have been made to construct cotidal charts. These attempts have been based on a little theory and far too few observations.

Figures 3 and **4** show Dietrich's cotidal chart for M_2. Each curve passes through points having high water at the same time, time being indicated as phase of the M_2 equilibrium argument. A characteristic feature of cotidal charts is the occurrence of points through which all cotidal curves pass. These are called amphidromic points. Here the amplitude of the constituent under consideration must be zero. The existence of such amphidromic points has been borne out by theoretical studies of tides in ocean basins of simple geometric shape. The mechanism which gives rise to amphidromic points is intimately related to the rotation of the Earth and the Coriolis force.

The amplitude of a constituent, $h(\theta,\phi)$, is generally high in some large regions of the oceans and low in others, but in addition there are small-scale erratic variations, at least along the coastline. Perhaps this is partly an illusion caused by the placement of some tide gages near the open coast and the placement of others up rivers and estuaries. It is well known that the phase and amplitude of the tide change rapidly as the tidal wave progresses up a river. *See* RIVER TIDES.

The range of the ocean tide varies between wide limits. The highest range is encountered in the Bay of Fundy, where values exceeding 50 ft (15 m) have been observed. In some places in the Mediterranean, South Pacific, and Arctic, the tidal range never exceeds 2 ft (0.6 m).

The tide may be considerably different in small adjacent seas than in the nearby ocean, and here resonance phenomena frequently occur. The periods of free oscillation of a body of water are determined by their boundary and depth configurations. If one of these free periods is near that of a large tidal constituent, the latter may be amplified considerably in the small sea. The large tidal range in the Bay of Fundy is an example of this effect. Here the resonance period is nearly 12 h, and it is the semidiurnal constituents that are large. The diurnal constituents are

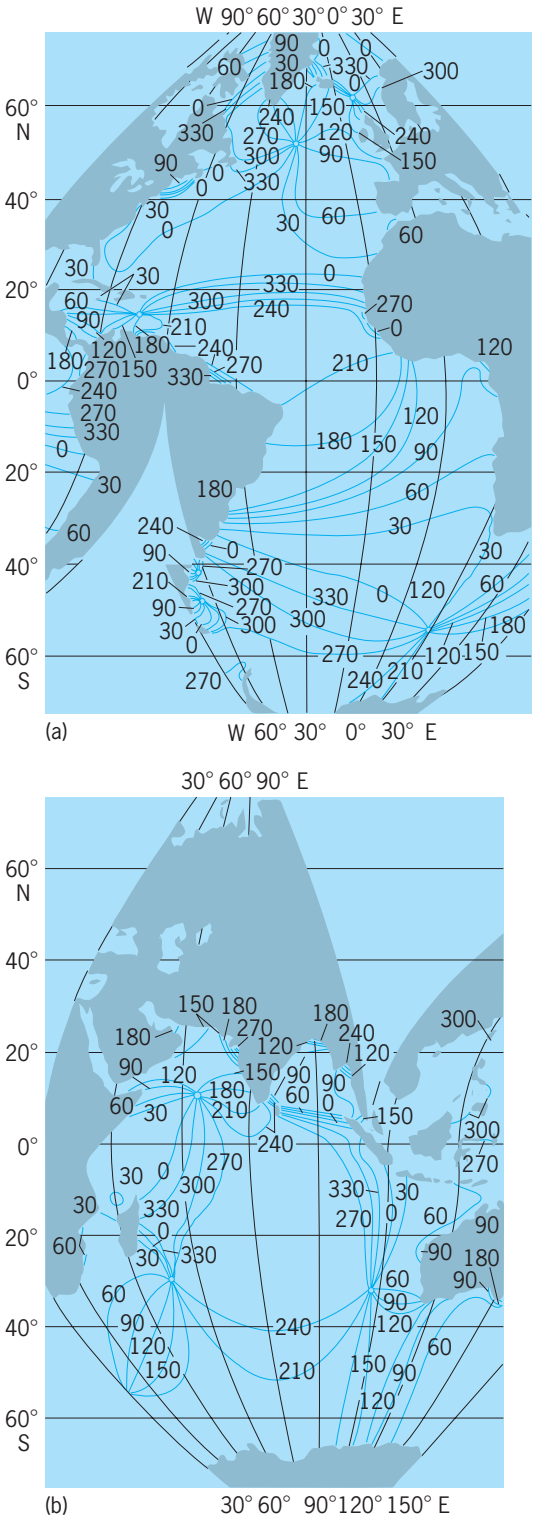


Fig. 3. Cotidal chart for M_2. (*a*) Atlantic Ocean. (*b*) Indian Ocean. (*After G. Dietrich, Veroeff. Inst. Meeresk., n.s. A, Geogr.- naturwiss. Reihe, no. 41, 1944*)

not extremely greater in the Bay of Fundy than in the nearby ocean. *See* SEICHE.

In lakes and other completely enclosed bodies of water the periods of free oscillation are usually much smaller than those of the tidal constituents. Therefore the tide in these places obeys the principles of statics. Since there is no tidal variation in the total volume of water in lakes the mean surface elevation does not change with the tide. The surface slope is determined by the slope of the equilibrium tide, and the related changes in elevation are usually very small, of the order of a fraction of a millimeter for small lakes.

Tidal currents. The south and east components of the tidal current can be developed in the same way as the tidal elevation since they also depend linearly on the tidal forces. Consequently, the same analysis and prediction methods can be used. Expressions similar to Eq. (5) represent the current components, each constituent having its own amplitude and phase at each geographic point. It should be emphasized that the current speed or direction cannot be developed in this way since these are not linearly related to the tidal forces.

Only in special cases are the two tidal current components exactly in or out of phase, and so the tidal current in the ocean is generally rotatory. A drogue or other floating object describes a trajectory similar in form to a Lissajous figure. In a narrow channel only the component along its axis is of interest. Where shipping is important through such a channel or port entrance, current predictions, as well as tidal height predictions, are sometimes prepared.

Owing to the rotation of the Earth, there is a gyroscopic, or Coriolis, force acting perpendicularly to the motion of any water particle in motion. In the Northern Hemisphere this force is to the right of the current vector. The horizontal, or tractive, component of the tidal force generally rotates in the clockwise sense in the Northern Hemisphere. As a result of both these influences, the tidal currents in the open ocean generally rotate in the clockwise sense in the Northern Hemisphere, and in the counterclockwise sense in the Southern Hemisphere. There are exceptions, however, and the complete dynamics should be taken into account. *See* CORIOLIS ACCELERATION.

The variation of the tidal current with depth is not well known. It is generally agreed that the current would be constant from top to bottom were it not for stratification of the water and bottom friction. The variation of velocity with depth due to the stratification of the water is associated with internal wave motion. Serial observations made from anchored or drifting ships have disclosed prominent tidal periodicities in the vertical thermal structure of the water.

Dynamics of ocean tide. The theoretical methods for studying tidal dynamics in the oceans were put forth by Laplace in the eighteenth century. The following assumptions are introduced: (1) The water is homogeneous; (2) vertical displacements and velocities of the water particles are small in comparison to the horizontal displacements and velocities; (3) the water pressure at any point in the water is given adequately by the hydrostatic law, that is, it is equal to the head of water above the given point; (4) all dissipative forces are neglected; (5) the ocean basins are assumed rigid (as if there were no bodily tide), and the gravitational potential of the

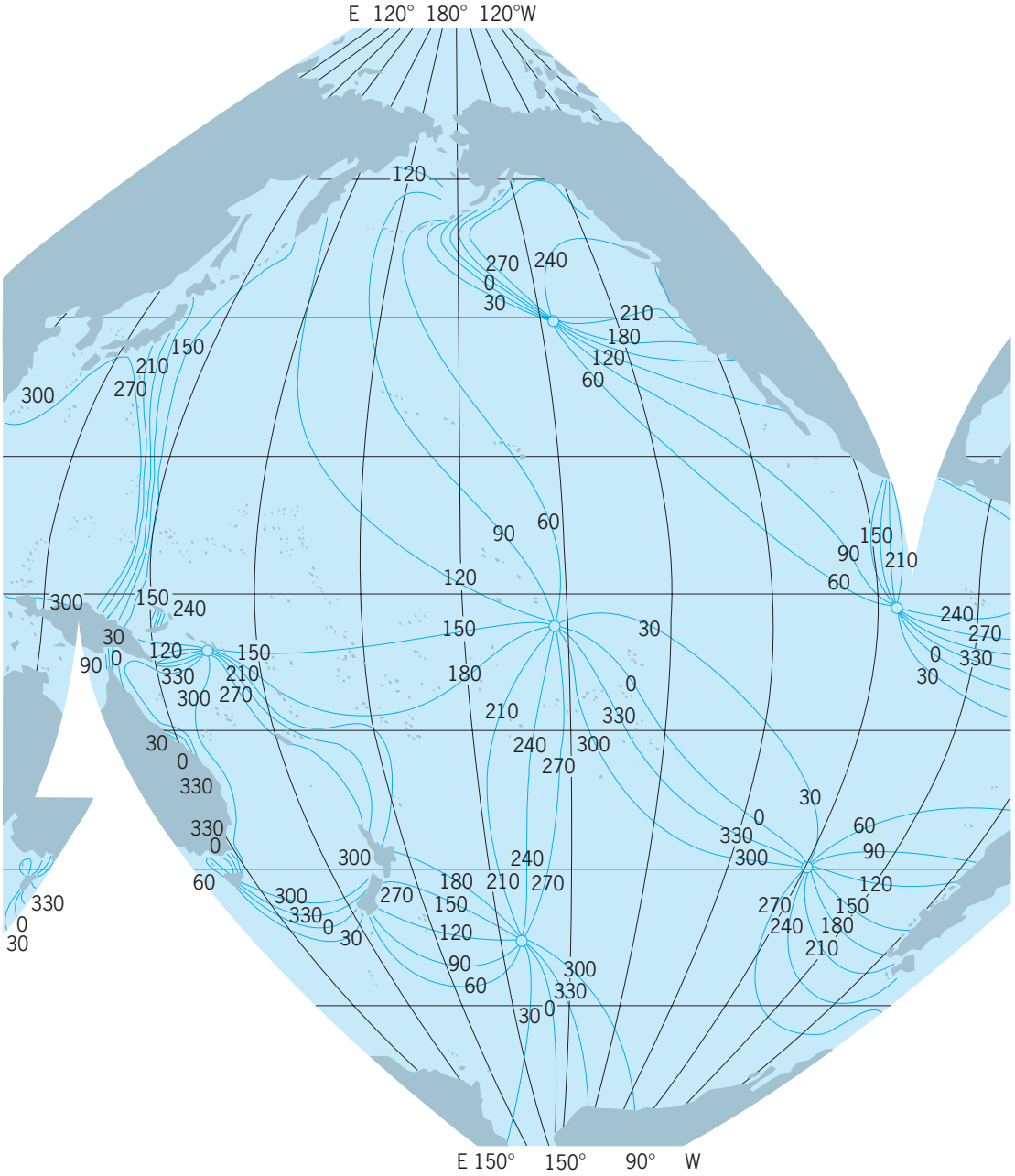


Fig. 4. Pacific Ocean cotidal chart for M_2. (*After G. Dietrich, Veroeff. Inst. Meeresk., n.s. A, Geogr.-naturwiss. Reihe, no. 41, 1944*)

tidally displaced masses is neglected; and (6) the tidal elevation is small compared with the water depth.

If assumptions (1) and (3) are valid, it can readily be shown that the tidal currents are uniform with depth. This is a conclusion which is not in complete harmony with observations, and there are internal wave modes thus left out of Laplace's theory. Nevertheless the main features of the tide are probably contained in the equations.

The water motion in the oceans is, in theory, determined by knowledge of the shape of the ocean basins and the tide-generating force (or equilibrium tide) at every point in the oceans for all time. The theory makes use of two relations: (1) the equation of continuity, which states that the rate of change of water mass in any vertical column in the ocean is equal to the rate at which water is flowing into the column; and (2) the equations of motion, which state that the total acceleration of a water "particle" (relative to an inertial system, thus taking into account the rotation of the Earth) is equal to the total force per unit mass acting on that particle. Under the above assumptions, the equation of continuity takes the form of Eq. (6), where $d(\theta,\phi)$ is the water

$$\frac{\partial \zeta}{\partial t} = -\frac{1}{a \sin\theta}\left[\frac{\partial}{\partial \theta}(ud \sin\theta) + \frac{\partial}{\partial \phi}(vd)\right] \quad (6)$$

depth. The equations of motion in the southward

and eastward directions, respectively, are given by Eqs. (7), where ω designates the angular rate of ro-

$$\begin{aligned} \frac{\partial u}{\partial t} - 2\omega v \cos\theta &= \frac{g}{a}\frac{\partial}{\partial \theta}(\zeta - \overline{\zeta}) \\ \frac{\partial v}{\partial t} + 2\omega u \cos\theta &= -\frac{g}{a}\csc\theta\frac{\partial}{\partial \phi}(\zeta - \overline{\zeta}) \end{aligned} \tag{7}$$

tation of the Earth, and u and v the south and east components of the tidal current. All other quantities are as previously defined.

It is probable that exact mathematical solutions to Eqs. (6) and (7), taking even approximately into account the complicated shape of the ocean basins, will never be obtained. However, the equations have certain features which serve to give us some insight into the nature of ocean tides. For instance, it is evident that if many equilibrium tides are acting simultaneously on the ocean, then the ocean tide will be the sum of the individual reactions. This linearity results directly from assumption (6). In certain shallow regions of the ocean the tides are noticeably distorted, as would be expected if assumption (6) were violated. This distortion is usually considered as resulting from the presence of so-called shallow-water constituents having frequencies equal to harmonics and to beat frequencies of the equilibrium constituents. These must be considered, at some places, or there will be large discrepancies between prediction and observation. Certain mathematical solutions to Eqs. (6) and (7) have been obtained for hypothetical ocean basins of simple geometric shape. Laplace solved them for an ocean of constant depth covering the entire Earth. Several solutions have been obtained for an ocean of constant depth bounded by two meridians. The result of one of the solutions obtained by J. Proudman and A. Doodson is shown in **Fig. 5**, which represents a cotidal chart of the K_2 tide in an ocean of depth 14,520 ft (4426 m) bounded by meridians 70° apart. The K_2 tide was calculated because of mathematical simplifications, but the M_2 tide should be quite similar. Comparison of Fig. 5 with the Atlantic Ocean in Fig. 3 discloses no striking similarities except for the general occurrence of amphidromic systems.

Bodily tide. The solid part of the Earth suffers periodic deformation resulting from the tide-generating forces just as the oceans do. *See* EARTH TIDES.

The gravest known modes of free oscillation of the solid Earth have periods of the order of an hour, much shorter than those of the principal tidal constituents. Therefore, the principles of statics can be used to describe the bodily tide, in contrast to tides in the oceans and atmosphere, where the inertial effect is important.

Associated with the bodily tide are periodic changes in gravity, manifesting themselves as (1) a variation of the vertical, or plumb line, with respect to any solid structure embedded in the Earth's crust; and (2) a variation in the magnitude of the acceleration of gravity at any point. These effects arise from the gravitational attraction of the tidally displaced matter of the Earth (solid, ocean, and atmosphere)

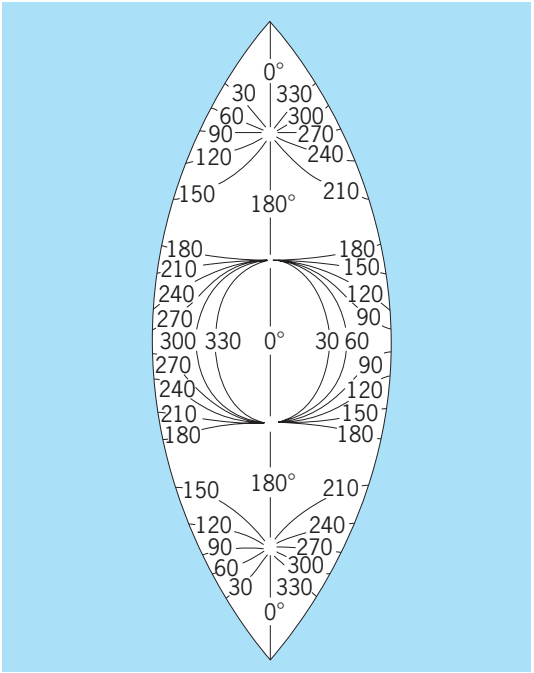


Fig. 5. Cotidal chart for K_2 in a hypothetical ocean of constant depth bounded by meridians 70° apart. (*After A. T. Doodson and H. D. Warburg, Admiralty Manual of Tides, London, 1941*)

as well as directly from the tide-generating forces. The magnitude of the former factor is of the order of several tens of microgals (1 gal = 1 cm/s^2).

Atmospheric tides. Since air, as other matter, is subject to gravitational influence, there are tides in the atmosphere possessing many features of similarity with those in the ocean. One of the characteristics of these tides is a small oscillatory variation in the atmospheric pressure at any place. This fluctuation of pressure, as in the case of the ocean tide, may be considered as the sum of the usual tidal constituents, and standard tidal analysis and prediction methods may be used. The principal lunar semidiurnal constituent M_2 of the pressure variation has been determined for a number of places, and found to have an amplitude of the order of 0.03 millibar (3 pascals). The dynamical theory of these tides has been the subject of considerable study. The equations which have been considered have the same general form as those for ocean tides. The S_2 constituent shows a much larger oscillation with an amplitude of the order of 1 millibar (10^2 Pa), but here diurnal heating dominates the gravitational effects. If diurnal heating were the whole story one would expect an even larger S_1 effect, and the fact that S_2 is larger is attributed to an atmospheric resonance near 12 h. *See* ATMOSPHERIC TIDES; UPPER-ATMOSPHERE DYNAMICS.

Tidal analysis and prediction. The distribution in space and time of the tidal forces within the Earth is precisely known from astronomic data. The effects of these forces on the oceans cannot, by present methods, be described in detail on a worldwide basis because of the difficult nature of the dynamical relationships and the complicated shape of the ocean basins. Practical prediction methods make use

of past observations at the place under consideration.

The procedure is the same for prediction of any tidal variable—such as the atmospheric pressure, component displacements of the solid Earth, components of the tidal current, and so on—which depends linearly on the tidal forces. In the harmonic method the frequencies, or periods, of the tidal constituents are determined by the astronomic data, and the harmonic constants (amplitudes and epochs) are obtained from the observations. Equation (5) then represents the tide at all past and future times for the place under consideration, where the values of b are the amplitudes of whatever tidal variable is being predicted. In this discussion the sea-level elevation will be used as an example, since it is the variable for which predictions are most commonly made. The procedure is basically the same for each constituent, but is most easily described for the series of constituents, $S_1, S_2, S_3, \ldots$, whose periods are submultiples of 24 h.

Suppose that the tidal elevation at 1:00 A.M. is averaged for all the days of the tide record, and similarly for 2:00 A.M., 3:00 A.M., and for each hour of the day. The 24 values thus obtained represent the average diurnal variation during the entire record. Any constituent whose period is not a submultiple of 24 h will contribute very little to the average of all the 1:00 A.M. values since its phase will be different from one day to the next, and its average value at 1:00 A.M. will be very close to zero for a long record. The same is true for each hour of the day, and so its average diurnal variation is small. The longer the record the freer will be the average diurnal oscillation from the effects of the other constituents. The diurnal oscillation is then analyzed by the well-known methods of harmonic analysis to determine the amplitudes and phases of all the harmonics of the 24-h oscillation. *See* FOURIER SERIES AND TRANSFORMS.

The same procedure is used for each other constituent; that is, the tide record is divided into consecutive constituent days, each equal to the period (or double the period in the case of the semidiurnal constituents) of the constituent. If the tide record is tabulated each solar hour, there is a slight complication because the constituent hours do not coincide with the solar hours. This difficulty is overcome by substituting the tabulated value nearest the required time and later compensating the consistent error introduced by an augmenting factor.

Since the record length is always finite, the harmonic constants of a constituent determined by this method are somewhat contaminated by the effects of other constituents. A first-order correction of these effects can be made by an elimination procedure. In general, it is more efficient to take the record length equal to the synodic (beat) period of two or more of the principal constituents. The longer the record, the better. Standard analyses consist of 29 days, 58 days, 369 days, and so on.

It is not practical to determine the harmonic constants of the lesser constituents in this way if errors or uncertainties of the data are of the same order of magnitude as their amplitudes. If tidal oscillations in the oceans were far from resonance then the amplitude H of each constituent should be expected to be approximately proportional to its theoretical coefficient C, and the local epochs G all to be near the same value. For the semidiurnal constituent X, Eqs. (8) should hold. Here X is referred to M_2 for

$$\frac{H(X)}{C(X)} = \frac{H(M_2)}{C(M_2)} \qquad G(X) = G(M_2) \qquad (8)$$

the reason that the latter is one of the principal constituents whose harmonic constants can be determined with best accuracy. Any other important constituent could be used. Inferring the harmonic constants of the lesser constituents by means of Eqs. (8) is sometimes preferable to direct means. It should be borne in mind that a constituent of one species cannot be inferred from one of another species because their equilibrium counterparts have different geographic shapes and no general relationship such as Eqs. (8) exists.

Once the harmonic constants are determined, the tide is synthesized according to Eq. (5), usually with the help of a special tide-predicting machine, although any means of computation could be used. Usually only the times and heights of high and low water are published in the predictions.

Tidal friction. The dissipation of energy by the tide is important in the study of planetary motion because it is a mechanism whereby angular momentum can be transferred from one type of motion to another. An appreciable amount of tidal dissipation takes place in the ocean, and possibly also in the solid Earth. In 1952 Sir Harold Jeffreys estimated that about half the tidal energy present in the ocean at any time is dissipated each day. A large part of this dissipation takes place by friction of tidal currents along the bottom of shallow seas and shelves and along the coasts. The rate of dissipation is so large that there should be a noticeable effect on the tide in the oceans.

If the planet's speed of rotation is greater than its satellite's speed of revolution about it, as is the case in the Earth-Moon system, then tidal dissipation always tends to decelerate the planet's rotation, with the satellite's speed of revolution changing to conserve angular momentum of the entire system. The Moon's attraction on the irregularly shaped tidal bulge on the Earth exerts on it a decelerating torque. Thus tidal friction tends to increase the length of day, to increase the distance between Earth and Moon, and to increase the lunar month, but these increases are infinitesimal. The day may have lengthened by 1 s during the last 120,000 years because of tidal friction and other factors.

Gordon W. Groves

Bibliography. D. E. Cartwright, *Tides: A Scientific History*, 1998; P. Crean, T. S. Murty, and J. A. Stronach, *Mathematical Modeling of Tides and Estuarine Circulation*, 1988; H. Lamb, *Hydrodynamics*, 6th ed., 1945; G. I. Marchuk and B. A. Kagan, *Dynamics of Ocean Tides*, 1989; D. T. Pugh, *Tides, Surges and Mean Sea Level*, 1987.

Tie rod

A rod or bar, usually circular in cross section, used in structural parts of machines to tie together or brace connected members, or, in moving parts of machines or mechanisms, used to connect arms or parts to transmit motion. In the first use the rod ends are usually a threaded fastening, while in the latter they are usually forged into an eye for a pin connection.

In steering systems of automotive vehicles, the rod connects the arms of steering knuckles of each wheel. The connection between the rod and arms is a ball and socket joint. *See* AUTOMOTIVE STEERING.

In pressure piping, large forces are produced between connected parts. The pipes or parts are constrained by tie rods that may be rectangular in cross section, with pinned ends. *See* PIPELINE.

Paul H. Black

Tile

As a structural material, a burned clay product in which the coring exceeds 25% of the gross volume; as a facing material, any thin, usually flat, square product. Structural tile used for load bearing may or may not be glazed; it may be cored horizontally or vertically. Two principal grades are manufactured: one for exposed masonry construction, and the other for unexposed construction. Among the forms of exposure is frost; tile for unexposed construction where temperatures drop below freezing is placed within the vapor barrier or otherwise projected by a facing in contrast to roof tile.

Structural tile with a ceramic glaze is used for facing. The same clay material that is molded and fired into structural tile is also made into pipe, glazed for sewer lines, or unglazed for drain tile.

As a facing, clay products are formed into thin flat, curved, or embossed pieces, which are then glazed and burned. Commonly used on surfaces that are subject to water splash or that require frequent cleaning, such vitreous glazed wall tile is fireproof. Unglazed tile is laid as bathroom floor. By extension, any material formed into a size comparable to clay tile is called tile. Among the materials formed into tile are asphalt, cork, linoleum, vinyl, and porcelain. *See* CLAY, COMMERCIAL.

Frank H. Rockett

Till

Sediment deposited directly from glacier ice. Till is characteristically nonsorted and nonstratified and is deposited by lodgment or melt-out beneath a glacier or by melt-out on the surface of a glacier. The texture of till varies greatly (see **illus.**), and while all tills are characterized by a wide range of particle sizes, some are predominantly fine-grained (clayey or silty), while others are medium-grained (silty or sandy) or coarse-grained (gravelly or stony). Till contains a variety of rock and mineral fragments which reflect the source material over which the glacier flowed. The particles in the deposit usually show a preferred orientation related to the nature and direction of the ice flow. The overall character of the till reflects the source material, position and distance of transport, nature and position of deposition, and postdepositional changes.

Exposure of glacial till at the Black Rocks near Llandudno, Wales. Heterogeneous debris, ranging in size from large boulders to fine powder, displays no sorting or stratification. (*Photograph by K. F. Mather*)

Material released through melting on the surface of a glacier undergoes secondary modification by melt water or through mass movement. The former gives rise to various types of stratified glaciofluvial and glaciolacustrine deposits; the latter often gives rise to viscous debris (mud) flows, the deposits of which are lithologically similar to till. Such deposits have been considered a type of till and are called flow till. However, on a strict genetic basis, these deposits are no longer till; resedimentation has taken place, and they are debris flow deposits.

Till is a common surficial deposit in middle- and high-latitude areas that were glaciated during the Quaternary Period, and is the parent material for some of the best agricultural soils in the world. *See* GLACIATED TERRAIN.

W. Hilton Johnson

Bibliography. J. Ehlers et al. (eds.), *Glacial Deposits in North-East Europe*, 1995; D. E. Lawson, *Sedimentological Analysis of the Western Terminus Region of the Matanuska Glacier, Alaska*, Cold Reg. Res. Eng. Lab. Rep. 79–9, Hanover, NH, 1979; R. F. Legget (ed.), *Glacial Till*, Roy. Soc. Can. Spec. Pub. 12, 1976.

Tillodontia

An extinct order of early Cenozoic (about 65 to 40 million years ago) quadrupedal eutherian land mammals, represented by nine known genera, from the late Paleocene to middle Eocene of North America (*Esthonyx* [*Azygonyx*], *Megalesthonyx*, *Trogosus*, and *Tillodon*), early Paleocene to late Eocene of China (*Lofochaius*, *Meiostylodon*, *Adapidium*, *Trogosus* [*Kuanchuanius*]), middle Eocene of Pakistan (*Basalina*), and the early Eocene of Europe (*Plesiesthonyx*). *Anchippodus*, the first named tillodont genus (1868), based on a single left

lower molar found in middle Eocene rocks of New Jersey, is a nomen dubium (it is not certain which species of tillodont is represented by this tooth) and may represent either *Trogosus* or *Tillodon*. An indeterminate tillodont incisor fragment found on Ellesmere Island demonstrates that tillodonts lived in what is now the Canadian High Arctic during the Eocene under much warmer conditions than exist there today.

The tillodonts left no known descendants and were probably most closely related to the extinct order Pantodonta (another group of extinct ungulate-like mammals from the Paleocene and Eocene that in turn may be related to the early eutherian mammals known as arctocyonids).

Tillodonts were medium- to large-sized mammals (their skulls range in length from 5 to 37 cm or 2 to 15 in.) that probably fed primarily on roots and tubers in warm temperate to subtropical habitats. Tillodonts were most common in the early Eocene faunas of North America. They developed large second incisors that became rodentlike, relatively long snouts, massive skeletons, and moderately large claws. In some respects the adaptive morphology of the tillodonts converged on that of the taeniodonts, and members of the two groups may have competed for similar resources. In the past, some tillodont specimens have been misidentified as taeniodonts; the Pakistan tillodont *Basalina*, for instance, was originally referred to the Taeniodonta. *See* ARCHAIC UNGULATE; MAMMALIA; TOOTH.

Robert M. Schoch

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, W. H. Freeman, 1988; C. L. Gazin, *The Tillodontia: An Early Tertiary Order of Mammals*, Smithsonian Miscellaneous Collections, 1953; C. M. Janis, K. M. Scott, and L. L. Jacobs (eds.), *Tillodontia, Evolution of Tertiary Mammals of North America*, vol. 1: *Terrestrial Carnivores, Ungulates, and Ungulatelike Mammals*, Cambridge University Press, 1988.

Time

The dimension of the physical universe that orders the sequence of events at a given place; also, a designated instant in this sequence, such as the time of day, technically known as an epoch, or sometimes as an instant.

Measurement. Time measurement consists of counting the repetitions of any recurring phenomenon and possibly subdividing the interval between repetitions. Two aspects to be considered in the measurement of time are frequency, or the rate at which the recurring phenomena occur, and epoch, or the designation to be applied to each instant.

A determination of time is equivalent to the establishment of an epoch or the correction that should be applied to the reading of a clock at a specified epoch. A time interval may be measured as the duration between two known epochs or by counting from an arbitrary starting point, as is done with a stopwatch. Time units are the intervals between successive recurrences of phenomena, such as the period of rotation of the Earth or a specified number of periods of radiation derived from an atomic energy-level transition. Other units are arbitrary multiples and subdivisions of these intervals, such as the hour being 1/24 of a day, and the minute being 1/60 of an hour. *See* DAY; MONTH; TIME-INTERVAL MEASUREMENT; YEAR.

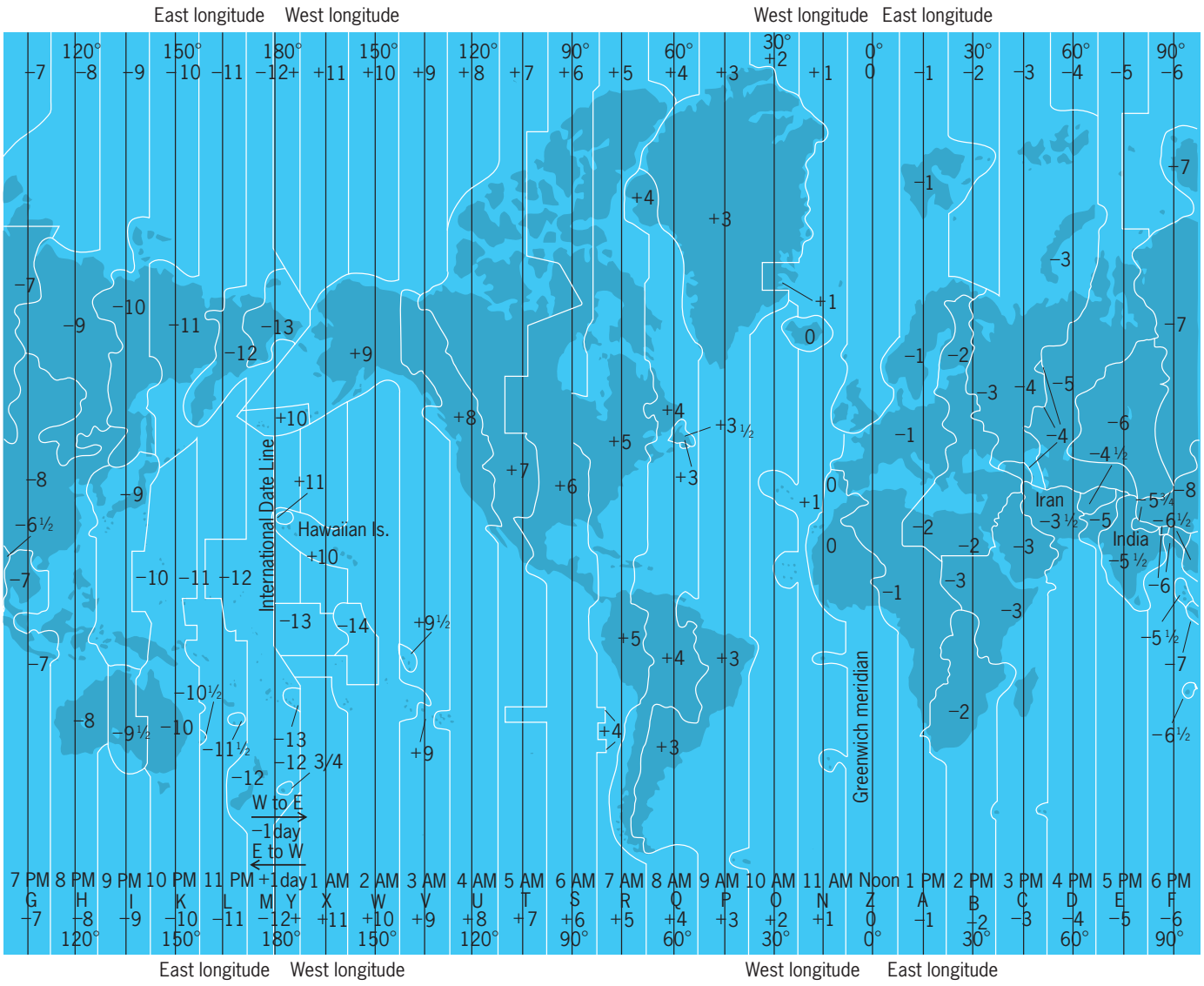
Bases. Several phenomena are used as bases with which to determine time. The phenomenon traditionally used has been the rotation of the Earth, where the counting is by days. Days are measured by observing the meridian passages of the Sun or stars and are subdivided with the aid of precision clocks. The day, however, is subject to variations in duration because of the variable rotation rate of the Earth. Thus, when a more uniform time scale is required, other bases for time must be used.

Sidereal time. The angle measured along the celestial equator between the observer's local meridian and the vernal equinox is the measure of sidereal time. In practice, a conventionally adopted mathematical expression provides this time as a function of civil time. It is reckoned from 0 to 24 hours, each hour being subdivided into 60 sidereal minutes and the minutes into 60 sidereal seconds. Sidereal clocks are used for convenience in many astronomical observatories because a star or other object outside the solar system comes to the same place in the sky at virtually the same sidereal time.

Solar time. The angle measured along the celestial equator between the observer's local meridian and the Sun is the apparent solar time. The only true indicator of local apparent solar time is a sundial. Mean solar time has been devised to eliminate the irregularities in apparent solar time that arise from the inclination of the Earth's orbit to the plane of the Sun's motion and the varying speed of the Earth in its orbit. In practice it is defined by a conventionally adopted mathematical expression. Intervals of sidereal time can be converted into intervals of mean solar time by dividing by 1.002 737 909 35. Both sidereal and solar time depend on the rotation of the Earth for their time base. *See* EQUATION OF TIME.

Universal Time (UT). Historically, the mean solar time determined for the meridian of 0° longitude using astronomical observations was referred to as UT1. Currently UT1 is used only as an angle expressed in time units that depends on the Earth's rotation with respect to the celestial reference system. It is defined by a conventional mathematical expression and continuing astronomical observations. These are made at a number of observatories around the world. The International Earth Rotation and Reference System Service (IERS) receives these data and provides daily values of the difference between UT1 and civil time. *See* EARTH ROTATION AND ORBITAL MOTION.

Because the Earth has a nonuniform rate of rotation and a uniform time scale is required for many timing applications, a different definition of a second was adopted in 1967. The international agreement calls for the second to be defined as 9,192,631,770 periods of the radiation derived from an energy-level



Division of the world into 24 time zones, progressively differing from Greenwich by 1 hour. Some countries use half-hour intervals or fractional hours. Numerical designations indicate number of hours by which zone time must be increased or decreased to obtain Coordinated Universal Time. Longitudes of standard meridians, letter designations, and the times in the zones when it is noon at Greenwich are also shown. (*Updated with data from www.worldtimezone.com and other Web sites*)

transition in the cesium atom. This second is referred to as the international or SI (International System) second and is independent of astronomical observations. International Atomic Time (TAI) is maintained by the International Bureau of Weights and Measures (BIPM) from data contributed by time-keeping laboratories around the world. *See* ATOMIC TIME.

Coordinated Universal Time (UTC) uses the SI second as its time base. However, the designation of the epoch may be changed at certain times so that UTC does not differ from UT1 by more than 0.9 s. UTC forms the basis for civil time in most countries and may sometimes be referred to unofficially as Greenwich Mean Time. The adjustments to UTC to bring this time scale into closer accord with UT1 consist of the insertion or deletion of integral seconds. These "leap seconds" may be applied preferably at 23 h 59 m 59 s of June 30 or December 31 of each year according to decisions made by the IERS. UTC differs from TAI by an integral number of atomic seconds.

Dynamical time. Dynamical time is based on the apparent orbital motion of the Sun, Moon, and planets. It is the time inferred in the ephemerides of the positions of these objects, and from its inception in 1952 until 1984 was referred to as Ephemeris Time. Barycentric Dynamical Time (TDB) refers to ephemerides that have been computed by using the barycenter of the solar system as a reference. Terrestrial Dynamical Time (TDT) is the practical realization of dynamical time and is defined as being equal to TAI + 32.184 seconds. In 1991, the International Astronomical Union recommended that TDT be renamed Terrestrial Time (TT), that Geocentric Coordinate Time (TCG) be the time coordinate for the geocenter, and that Barycentric Coordinate Time (TCB) be the time coordinate for the barycenter of the solar system. These times are related by the appropriate relativistic transformations. *See* DYNAMICAL TIME.

Civil and standard times. Because rotational time scales are local angular measures, at any instant they vary from place to place on the Earth. When the mean solar time is 12 noon at Greenwich, the mean solar time for all places west of Greenwich is earlier than noon and for all places east of Greenwich later than noon, the difference being 1 hour for each 15° of longitude. Thus, at the same instant at short distances east of the 180th meridian the mean solar time is 12:01 A.M., and at a short distance west of the same meridian it is 11:59 P.M. of the same day. Thus persons traveling westward around the Earth must advance their time 1 day, and those traveling eastward must retard their time 1 day in order to be in agreement with their neighbors when they return home. The International Date Line is the name given to a line where the change of date is made. It follows approximately the 180th meridian but avoids inhabited land. To avoid the inconvenience of the continuous change of mean solar time with longitude, zone time or civil time is generally used. The Earth is divided into 24 time zones, each approximately 15° wide and centered on standard longitudes of 0°, 15°, 30°, and so on (see **illustration**). Within each of these zones the time kept is related to the mean solar time of the standard meridian. *See* INTERNATIONAL DATE LINE.

Zone time is reckoned from 0 to 24 hours for most official purposes, the time in hours and minutes being expressed by a four-figure group followed by the zone designation. For example, "1009 zone plus five" refers to the zone 75° west of Greenwich, where zone time must be increased by 5 hours to obtain UTC. The various zones are sometimes designated by letters, especially the Greenwich zone which is Z, "1509 Z" meaning 1509 UTC. The zone centered on the 180th meridian is divided into two parts, the one east of the date line being designated plus 12 and the other minus 12. The time July 2,2400 is identical with July 3,0000.

In civil life the designations A.M. and P.M. are often used, usually with punctuation between hours and minutes. Thus 1009 may be written as 10:09 A.M. and 1509 as 3:09 P.M. The designations for noon and midnight, however, are often confused, and it is better to write 12:00 noon and July 2–3, 12:00 midnight, in order to avoid ambiguity. In some occupations where time is of special importance, there is a rule against using 12:00 at all, 11:59 or 12:01 being substituted. The time 1 minute after midnight is 12:01 A.M. and 1 minute after noon is 12:01 P.M.

The illustration shows the designations of the various time zones, the longitudes of the standard meridians, and the letter designations and the times in the various zones when it is noon at Greenwich. In the United States the boundaries of the time zones are fixed by the Department of Transportation. Frequently the actual boundaries depart considerably from the meridians exactly midway between the standard meridians. Ships at sea and transoceanic planes usually use UTC for navigation and communication, but for regulating daily activities onboard they use any convenient approximation to zone time, avoiding frequent changes during daylight hours.

Many countries, including the United States, advance their time 1 hour, particularly during the summer months, into "daylight saving time." For example, 6 A.M. is redesignated as 7 A.M. Such a practice effectively transfers an hour of little-used early morning light to the evening.

Time scales are coordinated internationally by the BIPM. Most countries maintain local time standards to provide accurate time within their borders by radio, telephone, and TV services. These national time scales are often intercompared by using the Global Positioning System (GPS) or time signals transferred by artificial Earth satellites. *See* SATELLITE NAVIGATION SYSTEMS.

Dennis D. McCarthy

Bibliography. D. W. Allan, N. Ashby, and C. C. Hodge, *The Science of Timekeeping*, Hewlett Packard Appl. Note 1289, 1997; D. D. McCarthy, Astronomical time, *Proc. IEEE*, 79:915–920, 1991; R. A. Nelson et al., The leap second: Its history and possible future, *Metrologia*, 38:509–529, 2001; G. M. R. Winkler, Timekeeping and its applications, in L. Marton (ed.), *Advances in Electronics and Electron Physics*, vol. 44, pp. 33–39,1997.

Time, arrow of

The uniform and unique direction associated with the apparent inevitable flow of time into the future. There appears to be a fundamental asymmetry in the universe. Yet herein lies a paradox, for all the laws of physics, whether they are the equations of classical mechanics, classical electromagnetism, general relativity, or quantum mechanics, are time-reversible in the sense that they admit solutions in either direction of time. This reversibility raises the question of how these fundamentally time-symmetrical equations can result in the perceived asymmetry of temporally ordered events.

Fundamental time asymmetries. The symmetry breaking of temporal order has not yet been fully explained. There are certain indications that an intrinsic asymmetry exists in temporal evolution. Thus it may be that the fundamental laws of physics are not really time-symmetric and that the currently known laws are only symmetrized approximations to the truth. Indeed, the decay of the K^0 meson is not time-reversible. However, it is not at all clear how such a rare and exotic instance of time asymmetry could emerge into the world of essentially macroscopic, electromagnetic phenomena as an everyday observable. *See* TIME REVERSAL INVARIANCE.

Another, more ubiquitous example of a fundamentally time-asymmetric process is the expansion of the universe. It has been speculated that this expansion is the true basis of time asymmetry, and that a resolution of the problem will come when quantum theory (the best available set of equations for mechanics) and general relativity (which deals with the structure of space-time) are combined into a single global theory. The existence of a single dimension

of time is also related to the question of time asymmetry. Because time has only one dimension, there is no analog of spatial rotation, and an entity that is evolving forward in time cannot reverse its direction in time in the same way that it can rotate in space and retrace its steps. Possibly, the inception of the universe at the big bang resulted in progress along time in a particular direction (the one to which the name "forward" is given), and it is now impossible to rotate evolution into the opposite direction. Even if the universe were to collapse again, there is no need to suppose that events will run in reverse, for the possibility is open for a final singularity to be far more elaborate than the initial singularity; so a cosmic asymmetry may exist even in a closed universe (**Fig. 1**). *See* BIG BANG THEORY; COSMOLOGY.

Statistical arguments. An alternative point of view is that time's arrow would exist even in the absence of these fundamental asymmetries. This view is essentially based on the statistical interpretation of the second law of thermodynamics, which identifies the direction of spontaneous change with the increase in disorder of the universe. The first serious attempt to relate the increase in entropy, the statistical measure of disorder, to the underlying time-symmetrical physical laws was made by L. Boltzmann. In his $\mathcal{H}$ theorem, he purported to demonstrate that a property which he called $\mathcal{H}$, and which is an integral over the positions and momenta of particles in a system, invariably decreased with time. Then, by identifying $\mathcal{H}$ with the negative of the entropy, he considered that he had proven that entropy increases with time, and thereby demonstrated the presence of time's arrow in statistically large assemblies of particles. However, Boltzmann's proof was invalid: he had imposed the equivalent of time asymmetry at one stage in the argument. The crucial step was to suppose that although the positions and momenta of two particles were not correlated before they collided, they became correlated by virtue of the collision. Thus, the collision imposed temporal asymmetry. *See* ENTROPY; STATISTICAL MECHANICS; THERMODYNAMIC PRINCIPLES.

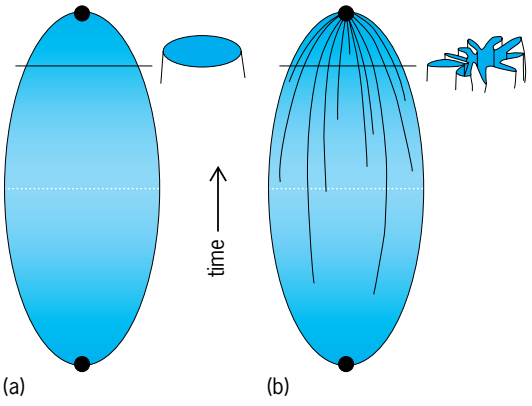


Fig. 1. Depiction of the possible course of evolution of a closed universe (one that starts with a big bang and ends in a big crunch). (*a*) Time-symmetric universe in which the final singularity is no more complex than the initial singularity. (*b*) Time-asymmetric universe in which the final universe is vastly more complex than the initial singularity.

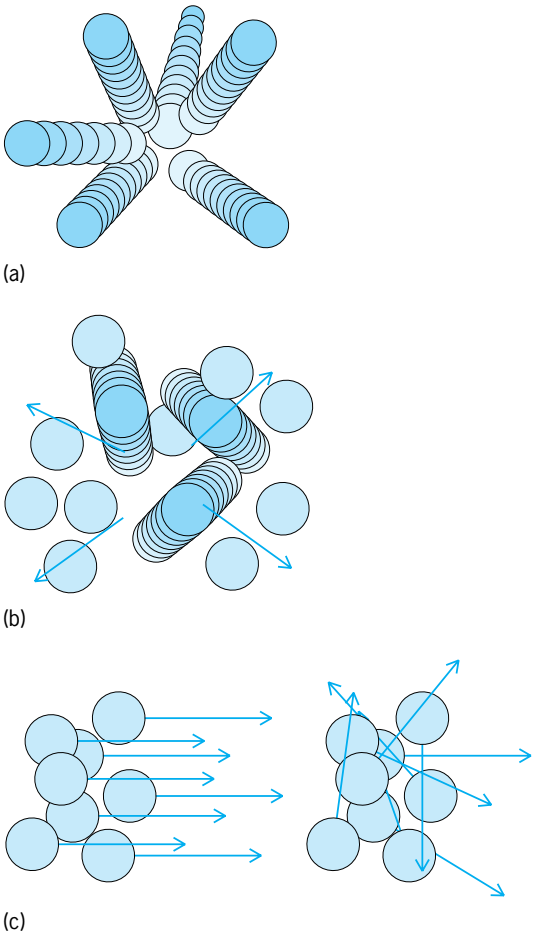


Fig. 2. The three basic irreversibilities of nature which ensure that events are irreversible: (*a*) dispersion of matter, (*b*) dispersion of energy, and (*c*) disorganization of orderly motion.

Consciousness of time asymmetry. It is appropriate to distinguish human consciousness of time asymmetry from an objective asymmetry in the evolution of events. Human consciousness of the unidirectional flow of time stems from the accumulation of memories, which are stored in an as yet unknown format in the brain. However, there can be no doubt that the mechanism of recording is chemical, and that memory is a neurochemical process. Therefore, any irreversibility of chemical reactions in the brain will result in storage of a memory, and hence in building a personality. The effective irreversibility of chemical reactions, neurochemical or otherwise, is well understood. Personality can evolve only as memories are added to previously existing memories, and this accumulation of memory is at the root of human consciousness of the passage of time. If time sometimes stood still and sometimes reversed its direction (there is not the slightest evidence for either event, and they would be difficult to reconcile with special relativity), then there would be no way of detecting the interruptions to time's forward flow, as all physiological changes would be suspended or reversed before resuming again.

Dispersal of energy and matter. The apparent irreversibility of phenomena is understood in terms of the dispersal of energy and matter that accompanies chemical change and the extreme unlikelihood that that dispersal will run in reverse spontaneously (**Fig. 2**). Thus, when a gas is released into a larger volume it spreads throughout the container because of the chaotic motion of its molecules. There is a vanishingly small probability that all the molecules will accumulate simultaneously and spontaneously back into the initial region that they occupied. They might accumulate there, but there is a serious chance of this happening only after such long intervals that the system will have been changed beyond recognition by mundane changes, astronomical disasters, the expansion of the universe, or baryon decay. The same is true of another basic irreversibility, that is, the flow of energy from high temperature to low. This takes place by a similar dispersal mechanism of the thermal motion of molecules, and it is no more reversible than is the dispersal of particles. These basic irreversibilities underpin chemical irreversibility, and hence lie beneath the irreversibility of all material change.

There may be fundamental reasons relating to the structure of space-time that account for the perceived asymmetry of time despite the current formulation of basic physical laws. Alternatively, even a time-symmetrical universe will have a statistical behavior in which configurations of molecules and localizations of energy have significant probabilities of recurring only after enormously long time intervals. Indeed, such time intervals are longer than the times required for the ceaseless expansion of the universe and the evolution of its component particles. Time's arrow is destined, either by the nature of space-time or the statistics of large assemblies, to fly into the future. P. W. Atkins

Bibliography. P. Coveney and R. Highfield, *The Arrow of Time*, 1990; R. Flood and M. Lockwood (eds.), *The Nature of Time*, 1988; J. J. Halliwell, J. Perez-Mercader, and W. Zurek, *The Physical Origins of Time Asymmetry*, 1994; P. Horwich, *Asymmetries in Time*, 1987; G. J. Whitrow, *The Natural Philosophy of Time*, 1980.

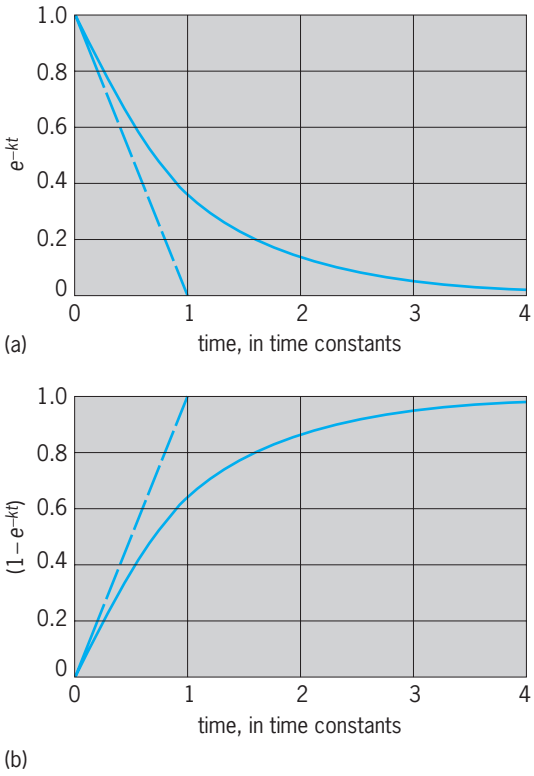
Time constant

A characteristic time that governs the approach of an exponential function to a steady-state value. When a physical quantity is varying as a decreasing exponential function of time as in Eq. (1), or as an increasing exponential function as in Eq. (2) [see **illus.**], the approach to the steady-state value achieved after a long time is governed by a characteristic time T as given in Eq. (3). This time T is called the time constant.

$$f(t) = e^{-kt} \tag{1}$$

$$f(t) = 1 - e^{-kt} \tag{2}$$

$$t = \frac{1}{k} = T \tag{3}$$



Universal time-constant curve indicated (*a*) for the decreasing function and (*b*) for the increasing function.

When time t is zero, $f(t)$ in Eq. (1) has the magnitude 1, and when t equals T the magnitude is $1/e$. Here e is the transcendental number whose value is approximately 2.71828, and the change in magnitude is $1 - (1/e) = 0.63212$. The function has moved 63.2% of the way to its final value. The same factor also holds for Eq. (2). *See* E (MATHEMATICS).

The initial rate of change of both the increasing and decreasing functions is equal to the maximum amplitude of the function divided by the time constant. Parts *a* and *b* of the illustration are universal in that the plotted function is of unit height and the time scale is given in terms of time constants. To use these curves for a specific problem, the values in the ordinate axis are multiplied by the maximum amplitude of the quantity occurring in the problem, and the values in the abscissa axis are multiplied by the numerical value of the corresponding time constant.

The concept of time constant is useful when evaluating the presence of transient phenomena. The relative amplitude of a transient after an elapsed time of a certain number of time constants is readily computed:

Elapsed time, time constants	*Transient completed, %*
1	63.2
2	86.5
3	95.0
4	98.2
5	99.3
10	99.996

Usually a transient can be considered as being over after a period of 4–5 time constants.

For electric circuits, the coefficient k and thus the time constant T is determined from the parameters of the circuit. For a circuit with resistance R and capacitance C, the time constant T is the product RC. When the circuit consists of inductance L and resistance R, the time constant is L/R. *See* ELECTRIC TRANSIENT.

The concept of time constant can be applied to the transient envelope of an ac signal; however, it is more common to describe the change in amplitude in terms of logarithmic decrement. For further discussion *See* DAMPING. Robert L. Ramey

Time-interval measurement

A determination of the duration between two instants of time (epochs). Time intervals are measured with high precision with a digital display counter. An electronic oscillator generates pulses; the count begins with a start signal and ends with a second signal. For an oscillator frequency of 100 MHz, for example, a direct reading is correct to 10 nanoseconds (1 ns = 10^{-9} s). Two atomic clocks, however, can be compared in epoch to 1 picosecond (1 ps = 10^{-12} s) by electronic interpolation. *See* ATOMIC CLOCK; DIGITAL COUNTER; OSCILLATOR; OSCILLOSCOPE.

Rapid motions can be studied at short intervals by means of a large variety of high-speed cameras, including stroboscopic, rotating film-drum, rotating mirror, streak, and image converter cameras. In one camera a helium turbine rotates a mirror at 20,000 revolutions per second to form 130 frames at 25,000,000 frames per second for 5.2 microseconds (1 μs = 10^{-6} s). The framing interval is 40 ns. An electronic streak camera can separate two pulses 1 ps apart. *See* PHOTOGRAPHY; STROBOSCOPIC PHOTOGRAPHY.

Ultrashort laser pulses are used to study rapid processes caused by the interaction of photons with an atom or molecule. The duration of interaction is $\tau = L/c$, where L is the pulse length and c the speed of light. One technique splits the pulse into two pulses. One pulse excites a reaction, and the other, optically delayed, probes the reaction. The probe pulse can be split into several pulses, staggered in time, with an echelon. Differences in path length give differences in time. Pulses as short as three wavelengths of 620-nm light, with $\tau = 6$ femtoseconds (1 fs = 10^{-15} s), have been formed. *See* LASER; LASER PHOTOCHEMISTRY; OPTICAL PULSES; ULTRAFAST MOLECULAR PROCESSES.

Radioactive decay is used to measure long time intervals, to about 5×10^9 years, concerning human history, the Earth, and the solar system. *See* GEOCHRONOMETRY; RADIOCARBON DATING. William Markowitz

Bibliography. J. C. Diels and W. Rudolp, *Ultrashort Laser Pulse Phenomena*, 2d ed., 2006; Eastman Kodak Co., *Encyclopedia of Practical Photography*, 1978; H. E. Edgerton, *Electronic Flash, Strobe*, 3d ed., 1987; L. Stroebel and R. D. Zakia (eds.), *The Focal Encyclopedia of Photography*, 3d ed., 1996.

Time-of-flight spectrometers

A general class of instruments in which the speed of a particle is determined directly by measuring the time that it takes to travel a measured distance. By knowing the particle's mass, its energy can be calculated. If the particles are uncharged (for example, neutrons), difficulties arise because standard methods of measurement (such as deflection in electric and magnetic fields) are not possible. The time-of-flight method is a powerful alternative, suitable for both uncharged and charged particles, that involves the measurement of the time t that a particle takes to travel a distance l. If the rest mass of the particle is m_0, its kinetic energy E_T can be calculated from its measured speed, $v = l/t$, using the equation below, where c is the speed of light.

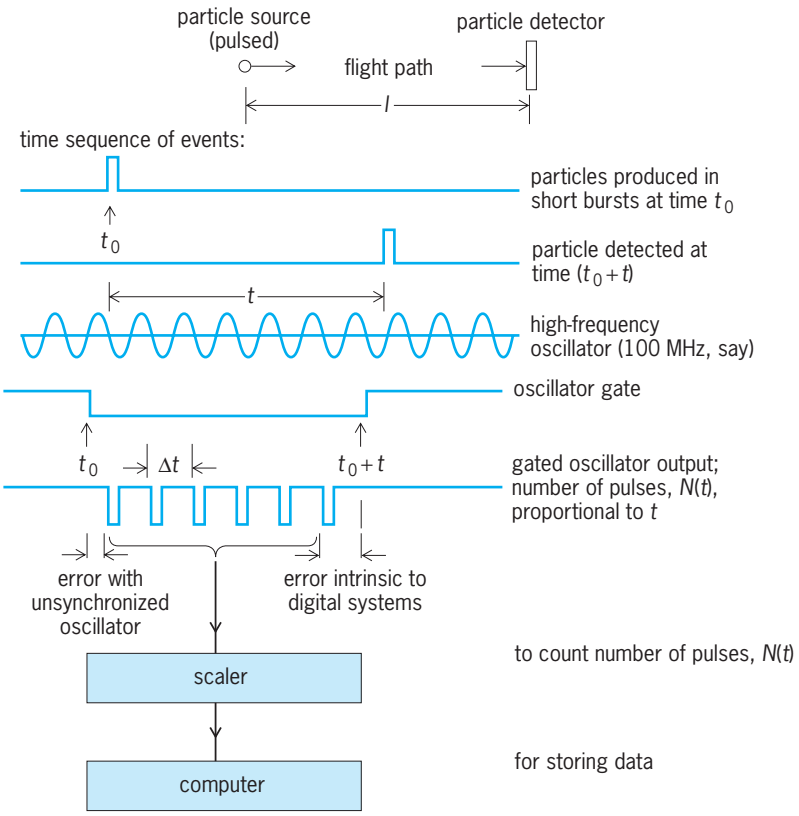
$$E_T = m_0c^2\left\{\left[1-\left(\frac{v}{c}\right)^2\right]^{-1/2}-1\right\}$$

$$\approx \frac{m_0v^2}{2} \qquad \text{if } v$$

Some idea of the time scales involved in measuring the energies of nuclear particles can be gained by noting that a slow neutron of kinetic energy $E_T =$ 1 eV takes 72.3 microseconds to travel 1 m. Its flight time along a 10-m path (typical of those found in practice) is therefore 723 μs, whereas a 4-MeV neutron takes only 361.5 nanoseconds.

The time intervals are best measured by counting the number of oscillations of a stable oscillator that occur between the instants that the particle begins and ends its journey (see **illus.**). Oscillators operating at 100 MHz are in common use. If the particles from a pulsed source have different energies, those with the highest energies arrive at the detector first. Digital information from the "gated" oscillator consists of a series of pulses whose number $N(t)$ is proportional to the time-of-flight t. These pulses can be counted and stored in an on-line computer that provides many thousands of sequential "time channels," $t_0, t_0 + \Delta t, t_0 + 2\Delta t, t_0 + 3\Delta t, \ldots$, where t_0 is the time at which the particles are produced and Δt is the period of the oscillator. To store an event in channel $N(t)$, the contents of memory address $N(t)$ are updated by "adding 1."

Time-of-flight spectrometers have been used for energy measurements of uncharged and charged elementary particles, electrons, atoms, and molecules. The popularity of these instruments is due to the broad energy range that can be covered, their high resolution ($\Delta E_T/E_T \approx 2\Delta t/t$, where ΔE_T and Δt are the uncertainties in the energy and time measurements, respectively), their adaptability for studying different kinds of particles, and their relative



Schematic diagram of a time-of-flight spectrometer.

simplicity. *See* MASS SPECTROSCOPE; NEUTRON SPECTROMETRY.

Frank W. K. Firk

Bibliography. R. J. Cotter (ed.), *Time-of-Flight Mass Spectroscopy*, 1994; E. W. Schlag (ed.), *Time-of-Flight Mass Spectroscopy and Its Applications*, 1994.

Time-projection chamber

An advanced particle detector for the study of ultrahigh-energy collisions of positrons and electrons that was developed originally at the Lawrence Berkeley Laboratory. The underlying physics of the scattering process can be studied through precise measurements of the momenta, directions, particle species, and correlations of the collision products. The time-projection chamber (TPC) provides a unique combination of capabilities for these studies and other problems in elementary particle physics by offering particle identification over a wide momentum range, and by offering high resolution of intrinsically three-dimensional spatial information for accurate reconstruction of events.

The time-projection chamber concept is based on the maximum utilization of ionization information, which is deposited by high-energy charged particles traversing a gas. The ionization trail, a precise image of the particle trajectory, also contains information about the particle velocity. A strong, uniform magnetic field and a uniform electric field are generated within the time-projection chamber active volume in an exactly parallel orientation. The parallel configuration of the fields permits electrons, products of the ionization processes, to drift through the time-projection chamber gas over great distances without distortion; the parallel configuration offers a further advantage in that the diffusion of the electrons during drift can be greatly suppressed by the magnetic field, thus preserving the quality of track information. In practice, the track images are drifted on the order of 100 cm (40 in.) or more, yet with measurement precision typically better than ±0.02 cm.

At the end of the drift volume the ionization electrons are multiplied by an avalanche process on an array of several hundred wires acting as proportional amplifiers. A highly segmented cathode plane just behind the wire array detects the avalanches and provides two-dimensional spatial coordinates in the plane of the array. The drift time provides the trajectory coordinate perpendicular to the plane of the array, hence suggesting the name time-projection chamber. The ionization density, measured precisely by the wire plane signals, offers the means to determine the particle velocity with resolution sufficient to establish the particle mass by a comparison of velocity and momentum.

Several large time-projection chambers are in operation or under construction at the premier storage ring facilities in the United States, Europe, and Japan. *See* ELECTRICAL BREAKDOWN; PARTICLE DETECTOR.

David R. Nygren

Bibliography. J. A. Macdonald (ed.), *The Time Projection Chamber*, AIP Conf. Proc. 108, 1984; J. N. Marx and D. R. Nygren, The time projection chamber, *Phys. Today*, 31:46–53, October 1978.

Time reversal invariance

A symmetry of the fundamental (microscopic) equations of motion of a system; if it holds, the time reversal of any motion of the system is also a motion of the system. To date, only two phenomena have shown evidence (at least indirect) for violation of time reversal invariance. One is the violation of *CP* invariance observed in the decays of the neutral mesons K_L and B^0, $\bar{B}^0$. The other is the baryon asymmetry of the universe.

Time reversal invariance is not evident from casual observation of everyday phenomena. If a movie is taken of a phenomenon, the corresponding time-reversed motion can be exhibited by running the movie backward. The result is usually strange. For instance, water in the ground is not ordinarily observed to collect itself into drops and shoot up into the air. However, if the system is sufficiently well observed, the direction of time is not obvious. For instance, a movie which showed the motion of the planets (in which each of the objects that make up the system can be observed individually), would look just as right run backward or forward. The apparent irreversibility of everyday phenomena results from the combination of imprecise observation and starting from an improbable situation (a state of low entropy, to use the terminology of statistical mechanics). *See* ENTROPY; STATISTICAL MECHANICS.

The known fundamental equations of motion are all time-reversal invariant. For instance, suppose that $\vec{r}_j(t)$ and $\vec{p}_j(t)$, the coordinates and momenta of charged particles, and $\vec{E}(\vec{r}, t)$ and $\vec{B}(\vec{r}, t)$, the electric and magnetic fields, satisfy the equations of motion of the particles and the fields (Newton's and Maxwell's equations); that is, these functions of time describe a motion of the system. Then the functions given in Eqs. (1), which describe the time reversal of the original motion, also satisfy the equations of motion. Thus the system is time-reversible. *See* MAXWELL'S EQUATIONS; NEWTON'S LAWS OF MOTION.

$$\begin{aligned} \vec{r}_{jREV}(t) &= \vec{r}_j(-t) \\ \vec{p}_{jREV}(t) &= -\vec{p}_j(-t) \\ \vec{E}_{REV}(\vec{r}, t) &= \vec{E}(\vec{r}, -t) \\ \vec{B}_{REV}(\vec{r}, t) &= -\vec{B}(\vec{r}, -t) \end{aligned} \tag{1}$$

Quantum mechanics. In quantum mechanics, if the hamiltonian operator H is independent of time (energy-conserving) and real, and if the wave function $\psi(t)$ is a solution of the Schrödinger equation (2), where $\hbar$ is Planck's constant divided by 2π and ∂_t is partial differentiation with respect to time, then the function given in Eq. (3),

$$i\hbar\partial_t\psi = H\psi \tag{2}$$

$$\psi_{REV}(t) = \psi^*(-t) \tag{3}$$

where the asterisk indicates complex conjugation, is also a solution of the Schrödinger equation. The function ψ_{REV} is the wave function of the time reversal of the motion described by ψ; hence the Schrödinger equation with a constant and real Hamiltonian is time reversal–invariant. In the time-reversed motion, other kinds of amplitudes, for example, transition amplitudes, are also complex-conjugated. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; SCHRÖDINGER'S WAVE EQUATION.

Tests. If time reversal invariance holds, no particle (a physical system with a definite mass and spin) can have an electric dipole moment $\vec{d}$, that is, an interaction energy of the form $-\vec{d}\cdot\vec{E}$, where $\vec{E}$ is the applied electric field. This is because the only intrinsic vector quantity of such a system (that is, the only vector quantity with nonvanishing expectation value in the rest frame) is the spin $\vec{S}$; thus an electric dipole moment $\vec{d}$ would have to be a multiple of $\vec{S}$, resulting in an interaction energy of the form $c\vec{S}\cdot\vec{E}$, where c is a constant. This energy changes sign under time reversal, because $\vec{S}$ does but $\vec{E}$ does not. (Since spin is an axial vector, an electric dipole moment of a particle would also violate space inversion symmetry or parity.) Although a polar body, for example, a water (H_2O) molecule, has an electric dipole moment, this shows up only in transition matrix elements between its eigenstates of energy (and spin). No particle has been observed to have an electric dipole moment; for instance, the present experimental upper limit on the electric moments of the electron and the neutron are approximately 10^{-27} cm times e and 10^{-25} cm times e, respectively, where e is the charge of the proton. *See* DIPOLE MOMENT; ELECTRON; NEUTRON; PARITY (QUANTUM MECHANICS); POLAR MOLECULE; SPIN (QUANTUM MECHANICS).

Another test of time reversal invariance is to compare the cross sections for reactions which are inverse to one another, for example, the reactions $^{16}O + d \leftrightarrow {}^{14}N + \alpha$. The present experimental upper limit on the relative size of the time reversal invariance–violating amplitude of such reactions is approximately 3×10^{-3}; unfortunately, this is far larger than any expected violation. *See* NUCLEAR REACTION.

A class of tests involves looking at the relative phases of amplitudes. For instance, if a nuclear transition emits both electric quadrupole and magnetic dipole electromagnetic radiation (gamma rays), certain interference terms in the angular distribution of the radiation cannot occur, because of the relative phase of the two amplitudes imposed by time reversibility. Experiments looking for such effects put an upper limit of approximately 10^{-3} on the relative size of a time reversal invariance–violating amplitude. *See* MULTIPOLE RADIATION.

Evidence for violation. A consequence of the *CPT* theorem is that violation of *T* (time reversal

invariance) is equivalent to violation of *CP*, that is, invariance of the fundamental equations under the combined operations of charge conjugation *C* and space inversion *P*. Hence *CP* violation observed in the decay of the long-lived neutral *K* meson (K_L) and in the decay of the neutral *B* mesons is evidence for *T* violation. *See* CPT THEOREM; MESON.

The K^0 and $\bar{K}^0$ mesons differ only in their strangeness, +1 and −1, respectively. The charged current weak interaction (the exchange of a charged weak boson) changes quark flavors, including strangeness, and so in second order (that is, acting twice) it can turn a K^0 into a $\bar{K}^0$ or vice versa. Consequently a K^0 or $\bar{K}^0$ is not a mass eigenstate but a coherent mixture of two mass eigenstates; conversely, the two mass (and decay rate) eigenstates are coherent mixtures of K^0 and $\bar{K}^0$. If *CP* were conserved, these mass eigenstates would be *CP* eigenstates, the more rapidly decaying one (called K_S, *S* for short-lived) having $CP = +1$, the *CP* value of the dominant decay mode of a neutral *K*, namely two pions, and the other (called K_L, *L* for long-lived) having $CP = -1$, thus unable to decay to two pions. But in fact the K_L has a small but nonvanishing branching ratio, ∼0.3%, for decay into two pions; this was the first observation of *CP* violation. Other evidences of *CP* violation are that the K_L has unequal branching ratios for decay into the *CP* conjugate states $\pi^+ e\bar{\nu}_e$ (a decay mode of $\bar{K}^0$) and $\pi^- \bar{e}\nu_e$ (a decay mode of K^0), and that the probabilities at short decay times for an original K^0 to decay to $\pi^+ e\bar{\nu}_e$ and for an original K^0 to $\pi^- \bar{e}\nu_e$, are unequal. *See* ELECTROWEAK INTERACTION; FLAVOR; QUARKS; STANDARD MODEL; WEAK NUCLEAR INTERACTIONS.

Similarly, B^0 and $\bar{B}^0$ mesons are mixtures of mass eigenstates; but no one decay mode dominates, so the lifetimes of the two mass eigenstates are not very different, and there is no simple way of experimentally distinguishing them. The observed *CP* violation is equivalent to saying that a $\bar{B}^0$ is about twice as likely to decay to $J/\psi\ K_S$ as is a B^0. This is not directly observable because *B* factories create (*B* and *B*) mesons in equal numbers, using the resonance reaction $e\bar{e} \rightarrow \Upsilon(4S) \rightarrow B^0\bar{B}^0$, which creates B^0 and $\bar{B}^0$ mesons in equal numbers. But if one of the two mesons created together is observed to decay into a negatively (positively) charged lepton, *e* or μ, plus other particles, it was a $\bar{B}^0(B^0)$. Hence at that time the other meson was a $B^0(\bar{B}^0)$, and by comparing its observed decay rate into $J/\psi\ K_S$ as a function of time (relative to the time of the semileptonic decay) to the theoretical formula, a *CP* violation parameter is deduced. Qualitatively described: In events in which one of the members of the $B^0\bar{B}^0$ pair decays semileptonically and the other decays into the hadronic mode $J/\psi\ K_S$, the *CP* violation is the observation that if the lepton of the semileptonic decay is positive, that decay is more likely to precede the $J/\psi\ K_S$ decay, whereas if negative, it is more likely to follow. *See* PARTICLE ACCELERATOR; UPSILON PARTICLES.

According to the standard model, *CP* violation arises from the CKM (Cabibbo-Kobayashi-Maskawa) quark-mixing matrix. When a down-type quark (*d*, *s*, or *b*) absorbs a W^+ or emits a W^- boson, it becomes a linear combination of up-type quarks (*u*, *c*, *t*); the coefficients of these three linear combinations is called the CKM matrix. (The same matrix also describes the linear combinations of down-type quarks resulting from absorption of a W^- or emission of a W^+ by an up-type quark.) It is a 3×3 unitary matrix with determinant 1. The unitarity of the CKM matrix expresses a universality of the weak interaction. For instance, a strange quark, *s*, becomes a *u*, *c*, or *t* quark with relative probabilities of 5%, 95%, and 0.2%, respectively. Because of energy conservation, in the decay of a strange particle such as K^+, the *c* or the *t* could only be virtual. The *u* could be real, that is, a constituent of a decay product; the coefficient 0.22 gives a K^+ decay rate about 5% of that predicted from the strength of the weak interaction as seen in ordinary β decay, where the quark transition is $|d> \rightarrow 0.975|u> + \cdots$. In other words, the CKM matrix gives the relative couplings of the charged weak boson to quarks. In general (that is, for a generic CKM matrix), all the couplings cannot be made real by any choice of the phases of the quark states; this results in violation of *T*. (The original quark-mixing matrix was a 4×4 matrix, proposed by N. Cabibbo, before the discovery of the third weak doublet, *t*,*b*; a 4×4 quark-mixing matrix never implies *T* violation. M. Kobayashi and T. Maskawa noted that the extension to a 3×3 matrix could give *T* violation in a natural way.) The CKM matrix consistently describes all weak-decay observations, including the *CP* violation seen in neutral *B* decay. *See* MATRIX THEORY.

An indirect but very prominent evidence for *T* violation is the baryon asymmetry of the universe, that is, the fact that ordinary matter contains baryons, not antibaryons. Put more quantitatively: The observed fact that there is roughly 1 baryon for every 10^9 cosmic blackbody photons means that early in the history of the universe, when the value of *kT* (the product of Boltzmann's constant and the thermodynamic temperature) was larger than 1 GeV (and hence the number of both baryons and antibaryons was roughly the same as the number of photons), there was an excess of baryons over antibaryons of the order of one part in 10^9. It is conceivable that this rather small baryon-number asymmetry developed in an originally baryon-number symmetric universe. This would require interactions, acting in the early universe, which violate both baryon number and *T*. The existence of the former would not be surprising, since particle interactions that do not conserve baryon number are always present in grand unified theories—gauge theories in which the strong and electroweak interactions, and likewise quarks and leptons, are unified. *See* COSMIC BACKGROUND RADIATION; COSMOLOGY; ELEMENTARY PARTICLE; GRAND UNIFICATION THEORIES; SYMMETRY LAWS (PHYSICS).

Charles J. Goebel

Bibliography. I. I. Bigi and A. I. Sanda, *CP Violation*, Cambridge University Press, 2000; W. M. Gibson and B. R. Pollard, *Symmetry Principles in Elementary Particle Physics*, 1976, paper 1980; E. M.

Henley, Parity and time-reversal invariance in nuclear physics, *Annu. Rev. Nucl. Sci.*, 19:367–432, 1969; I. B. Khriplovich and S. K. Lamoreaux, *CP Violation without Strangeness*, Springer-Verlag, 1997; K. Kleinknecht, *Uncovering CP Violation*, Springer-Verlag, 2003; D. Park, *Introduction to the Quantum Theory*, 3d ed., McGraw-Hill, 1992, reprint, Dover, 2005; R. G. Sachs, *The Physics of Time Reversal*, University of Chicago Press, 1987; M. Skalsey et al. (eds.), *Time Reversal*, AIP Press, 1993; L. Wolfenstein (ed.), *CP Violation*, Elsevier Science, 1990.

Time-reversed signal processing

A means for improving the performance of remote sensing and communication systems that rely on electromagnetic- or acoustic-wave propagation but must contend with wave reflections, diffraction, and scattering. In particular, for applications of remote sensing—including radar, sonar, biomedical imaging, and nondestructive evaluation—the main intent is the detection, localization, and identification of distant objects or features that either scatter or generate electromagnetic or acoustic waves. Although the disparity in objectives between military-radar and ultrasonic-imaging systems may be considerable, they have in common that random wave scattering and diffraction tend to limit the accuracy and confidence with which such systems can be used and the distances over which such systems can operate. For radar and sonar, turbulence and wave motions in the ocean or atmosphere, or rough terrestrial or ocean surfaces, may cause such random wave scattering and diffraction. In ultrasonic remote sensing, the random scattering and diffraction may be caused by grain structure within metals or by variations between and within tissues. *See* REMOTE SENSING.

Applications. Exploitation of time reversal in wave propagation problems dates back to the early 1960s, when similar concepts arose in the fields of geoacoustics, radio waves, and underwater sound propagation. In 1965, Antares Parvulescu and Clarence Clay published a study in which acoustic signals were recorded with a single transducer 20 nautical miles (37 km) from a sound source in water that was approximately 1 nautical mile (1.85 km) deep. The recorded signals were time-reversed, and retransmitted through the ocean. The signal received after the second transmission was much clearer and far less distorted than the signal received after the first transmission. Since that time, the sophistication of transmitters, receivers, and signal processing algorithms has allowed time reversal concepts to permeate nearly every application of remote sensing that relies on wave propagation to or from an object or feature of interest. Components for aircraft engines have been inspected with time-reversing ultrasonic arrays. The U.S. Navy is considering time-reversal concepts for new active sonar systems and underwater communication links. Time-reversal concepts are now being applied in biomedical diagnostic ultrasound to increase the clarity of images and in therapeutic ultrasound to better target tumors and kidney stones. Feasibility studies are getting underway for radar systems that incorporate time reversal to enhance operating distances and detection of targets under foliage, and the future of time-reversed signal processing is likely to include structural monitoring of buildings and machines. *See* ACOUSTIC SIGNAL PROCESSING; BIOMEDICAL ULTRASONICS; NONDESTRUCTIVE EVALUATION; RADAR; SONAR; ULTRASONICS; UNDERWATER SOUND.

Time-reversal process. The basic process of time reversal can be described by four steps (**Fig. 1**). First, waves are generated or scattered by an object or feature of interest and travel forward through the environment to an antenna or array of transducers. The paths that waves follow between the object and the array may be complicated and unknown. Second, the array records the signal in the usual manner. These recordings will include signal distortion from echoes, scattering, and diffraction in the environment. Third, the signal recorded by each element in the array is retransmitted from that element with the direction of time inverted; the end of the signal is transmitted first, and the start of the signal is transmitted last. In the final step, these array-transmitted time-reversed waves travel backward through the environment, retracing their paths to converge at the location where they originated. Although specific applications of time reversal typically involve more steps and greater processing of the array-received signals, all are based, directly or indirectly, on these simple steps.

When time reversal is working properly, the waves that return to the location of their origin focus tightly and are undistorted, even though they may have passed through a complicated environment that generates echoes and causes random scattering and diffraction. In fact, the size of the focal region may be much smaller when the environment is complicated than when it is uniform in all directions. Both the tight focusing and distortion-removal capabilities of time reversal are of interest in remote sensing.

Distortion removal and focusing. The time-reversed waves are able to remove distortion and focus well when the transducer array is many wavelengths long, the absorption of wave energy by the environment is weak, background noise is low, and the environment changes little (or not at all) between forward and backward wave travel. The process by which time reversal accomplishes distortion removal may be understood by considering an environment containing three travel paths—fast, middle, and slow—with different travel times between the object and the transducer array. In this case, a signal that starts at the object will be received at the array as three possibly overlapping signals. This phenomenon is called multipath distortion, and it occurs when there are well-defined wave-travel paths and when there is wave scattering from particles, turbulence, or other fluctuations between the object and the transducer array. In the three-path environment, the time-reversed broadcast will launch the signal on the slow

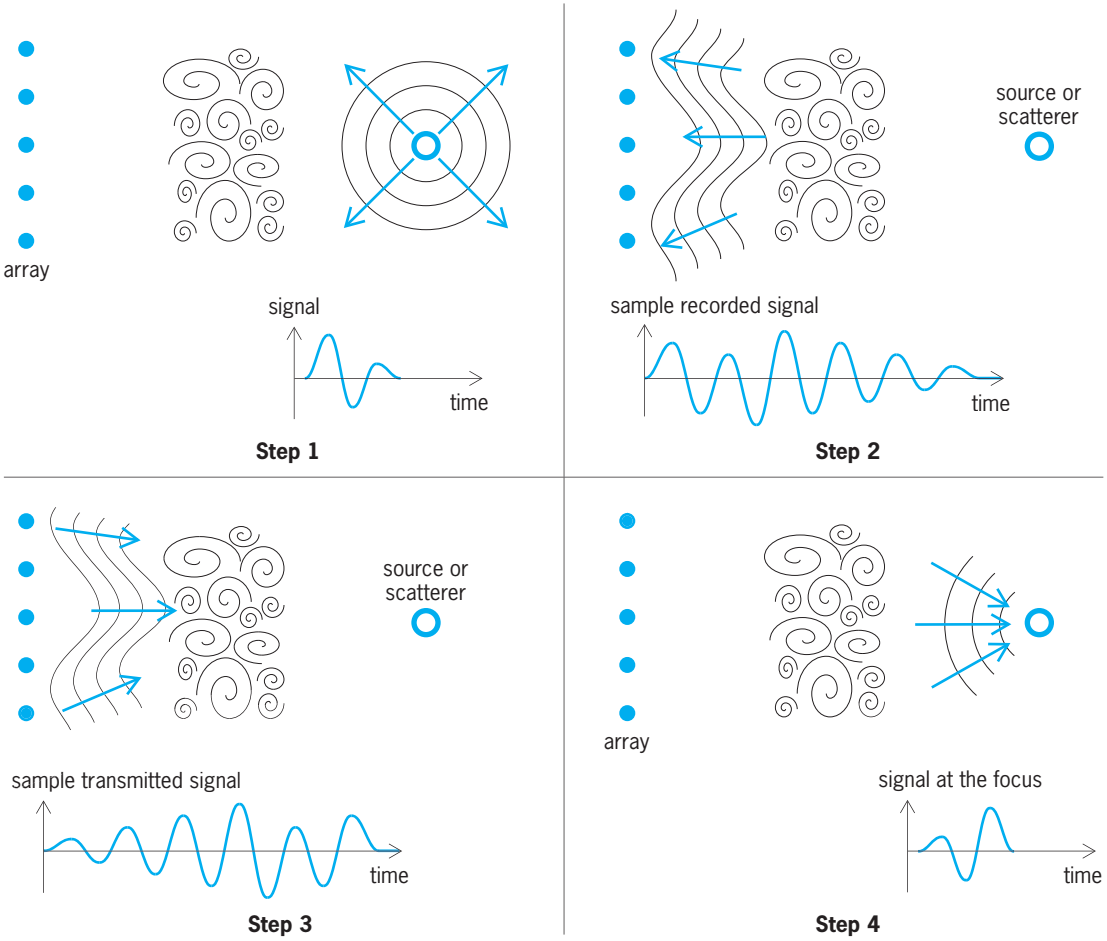


Fig. 1. Focusing and distortion compensation of waves that travel through an unknown random medium with time-reversed signal processing. Step 1: signal generation. Step 2: ordinary recording. Step 3: time-reverse playback. Step 4: focus formation.

path first, the middle path second, and the fast path last. Here, the timing of the signal launches will exactly undo the multipath distortion, and the signal will arrive undistorted back at its place of origin.

The superior focusing characteristics of time-reversed waves come from their ability to exploit reflection, scattering, and diffraction within the environment. The range of angles through which waves converge determines how tightly the converging waves focus. If all the time-reversed waves come from the same direction, the focus is larger than if the waves converge from above, from below, and from either side. When the environment is uniform, there is only one travel path between the object (a source or scatterer) and each element of the transducer array, so the focus size of the time-reversed waves is determined by the array's angular aperture. When the environment produces echoes, random scattering, and diffraction, there may be many travel paths between the object and the transducer array, and some of these paths can be used by the time-reversed waves to increase the range of convergence angles occurring during backward travel to decrease the focus size compared to what would occur in a uniform environment (**Fig. 2**). However, when the environment changes between forward and backward wave travel, the advantages of time reversal may be degraded or even lost.

Techniques. There are several signal processing techniques that have been used to convert time reversal from an alluring oddity of wave physics into a useful remote sensing tool. These techniques generally fall into two categories: those that require an active transducer array that can transmit and receive waves, and those that merely require a passive receiving array.

Active techniques. The most popular of the active techniques, referred to by the French acronym DORT (décomposition de l'operateur de retournement temporel; decomposition of the time-reversal operator), was developed by Claire Prada and Mathias Fink for nondestructive evaluation and biomedical ultrasound applications. This method can be used to detect and separately illuminate distinct scatterers. In its simplest implementation, DORT requires the measurement of the object-scattered signal at every array element when the object is separately illuminated by each array element. This carefully constructed measurement matrix can then be mathematically analyzed to determine the number or scattering objects, and the directions toward the objects in an environment with some random scattering and

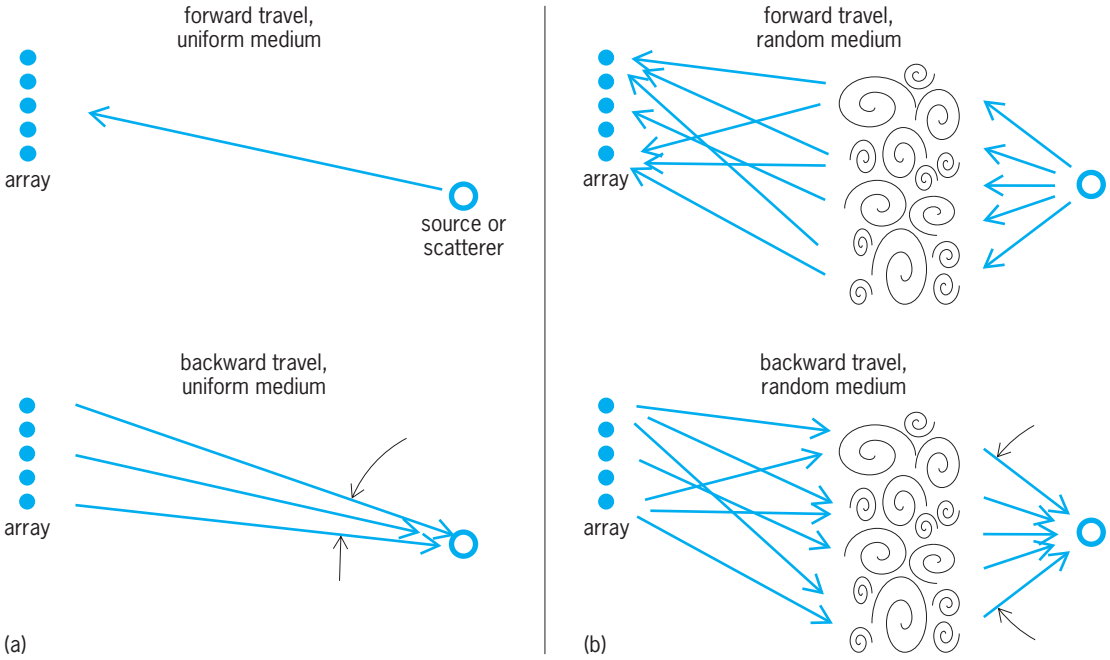


Fig. 2. Time-reversed focusing. (*a*) Without random wave scattering. (*b*) With random wave scattering.

diffraction but no strong echoes. In such an environment, an image of the scattering objects may be formed using DORT.

A second active technique may be combined with DORT and involves identifying and selectively time-reversing segments of the received signal that may contain scattered wave energy from the object or feature of interest. When the correct signal segment is time-reversed and transmitted, the waves that return to the transducer array will concentrate or peak at one particular time when an item of interest is present. When the rebroadcast signal segment corresponds to an ordinary section of the medium, the returning waves will not concentrate at any particular time. This technique has been used to detect the presence of internal flaws in titanium, an important aerospace material.

Other active array techniques have been proposed for underwater acoustic communication, security barriers, and reverberation reduction in active sonar systems.

Passive techniques. The passive array signal processing techniques based on time reversal are also important. The best known of these is matched-field processing (MFP). Here, the array-received signals are delivered to a computer program that can simulate the backward wave travel. The computed location where the backward traveling waves converge is the presumed location of the source or scatterer. Thus, this technique can be used by the computer operator to locate remote sources or scatterers based on the array reception alone; a transmission by the array is unnecessary. Although matched-field processing can theoretically locate or image objects with subwavelength accuracy, this technique cannot be implemented successfully without enough environmental information to ensure the accuracy of the computer model. *See* MATCHED-FIELD PROCESSING.

There are also passive time-reversed signal processing techniques that do not require any knowledge of the environment. One of these, passive phase conjugation (PPC), may be used for underwater communication between a cooperating remote source and a receiving array. If the source wishes to send a coded information stream to the array, it first sends a single pulse to characterize the multipath distortion of the environment. When the source starts its coded-information broadcast, the receiving array uses the measurements of the distorted single-pulse signal to unravel the distortion put into the coded message by the environment. Similar methods are used by modems to correct for variations in telephone lines. Another promising passive array technique, artificial time reversal, is similar to PPC but does not require the initial single-pulse broadcast.

David R. Dowling

Bibliography. J. Berryman et al., Statistically stable ultrasonic imaging in random media, *J. Acous. Soc. Amer.*, 112:1509–1522, 2002; D. R. Jackson and D. R. Dowling, Phase-conjugation in underwater acoustics, *J. Acous. Soc. Amer.*, 89:171–181, 1991; E. Kerbrat et al., Imaging in the presence of grain noise using the decomposition of the time reversal operator, *J. Acous. Soc. Amer.*, 113:1230–1240, 2003; W. A. Kuperman et al., Phase-conjugation in the ocean: Experimental demonstration of an acoustic time reversal mirror, *J. Acous. Soc. Amer.*, 103: 25–40, 1998; N. Mordant, C. Prada, and M. Fink, Highly resolved detection and selective focusing in a waveguide using the D.O.R.T. method, *J. Acous. Soc. Amer.*, 105:2634–2642, 1999; A. Parvulescu and C. S. Clay, Reproducibility of signal transmissions in the ocean, *Radio Electr. Eng.*, 29:223–228, 1965.

Timothy

A plant, *Phleum pratense*, of the order Cyperales, long the most important hay grass for the cooler temperate humid regions. It is easily established and managed, produces seed abundantly, and grows well in mixtures with alfalfa and clover. It is a short-lived perennial, makes a loose sod, and has moderately leafy stems 2–4 ft (0.6–1.2 m) tall and a dense cylindrical inflorescence (see **illus.**). Timothy responds

Timothy (*Phleum pratense*).

to fertile soils with high yield and nutritive content. Cutting promptly after heading improves the feed quality. Timothy-legume mixtures still predominate in hay and pasture seedings for crop rotations in the northern half of the United States, but orchard grass and bromegrass have increasingly replaced timothy in such mixtures in many areas. *See* ALFALFA; CLOVER; GRASS CROPS; INFLORESCENCE.

Howard B. Sprague

Tin

A chemical element, symbol Sn, atomic number 50, atomic weight 118.69. Tin forms tin(II) or stannous (Sn^{2+}), and tin(IV) or stannic (Sn^{4+}) compounds, as well as complex salts of the stannite (M_2SnX_4) and stannate (M_2SnX_6) types. *See* PERIODIC TABLE.

Tin melts at a low temperature, is highly fluid when molten, and has a high boiling point. It is soft and pliable and is corrosion-resistant to many media. An important use of tin has been for tin-coated steel containers (tin cans) used for preserving foods and beverages. Other important uses are solder alloys, bearing metals, bronzes, pewter, and miscellaneous industrial alloys. Tin chemicals, both inorganic and organic, find extensive use in the electroplating, ceramic, plastic, and agricultural industries.

1																	18
1 H	2											13	14	15	16	17	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

The most important tin-bearing mineral is cassiterite, SnO_2. No high-grade deposits of this mineral are known. The bulk of the world's tin ore is obtained from low-grade alluvial deposits. *See* CASSITERITE.

Two allotropic forms of tin exist: white (β) and gray (α) tin. Tin reacts with both strong acids and strong bases, but it is relatively resistant to solutions that are nearly neutral. In a wide variety of corrosive conditions, hydrogen gas is not evolved from tin and the rate of corrosion becomes controlled by the supply of oxygen or other oxidizing agents. In their absence, corrosion is negligible. A thin film of stannic oxide forms on tin upon exposure to air and provides surface protection. Salts that have an acid reaction in solution, such as aluminum chloride and ferric chloride, attack tin in the presence of oxidizers or air. Most nonaqueous liquids, such as oils, alcohols, or chlorinated hydrocarbons, have slight or no obvious effect on tin. Tin metal and the simple inorganic salts of tin are nontoxic. Some forms of organotin compounds, on the other hand, are toxic. Some important physical constants for tin are shown in the **table**.

Stannous oxide, SnO, is a blue-black, crystalline product which is soluble in common acids and strong alkalies. It is used in making stannous salts for plating and glass manufacture. Stannic oxide, SnO_2, is a white powder, insoluble in acids and alkalies. It is an excellent glaze opacifier, a component of pink, yellow, and maroon ceramic stains and of dielectric and refractory bodies. It is an important polishing agent for marble and decorative stones.

Stannous chloride, $SnCl_2$, is the major ingredient in the acid electrotinning electrolyte and is an intermediate for tin chemicals. Stannic chloride, $SnCl_4$, in the pentahydrate form is a white solid. It is used in the preparation of organotin compounds and chemicals to weight silk and to stabilize perfume and colors in soap. Stannous fluoride, SnF_2, a white water-soluble compound, is a toothpaste additive.

Organotin compounds are those compounds in which at least one tin-carbon bond exists, the tin usually being present in the + IV oxidation state. Organotin compounds that find applications in industry are the compounds with the general formula R_4Sn, R_3SnX, R_2SnX_2, and $RSnX_3$. R is an organic group, often methyl, butyl, octyl, or phenyl, while X is an inorganic substituent, commonly chloride, fluoride,

Properties of tin*

Property	Value
Melting point	231.9°C (449.4°F)
Boiling point	2270°C (4118°F)
Specific gravity,	
α-form (gray tin)	5.77
β form (white tin)	7.29
Liquid at melting point	6.97
Transformation temperature, °C	13.2
Specific heat, cal/g,	
white tin at 25°C	0.053
gray tin at 10°C	0.049
Latent heat of fusion, cal/g	14.2
Latent heat of vaporization, cal/g	520 ± 20
Heat of transformation, cal/g	4.2
Thermal conductivity,	
cal/(cm)(cm^2)°C(s), white tin at °C	0.150
Coefficient of linear expansion at 0°C	19.9×10^{-6}
Shrinkage on solidification, %	2.8
Resistivity of white tin, microhms/cm^3	
at 0°C	11.0
at 100°C	15.5
Brinell hardness, 10 kg/(5 mm)(180 s)	
at 20°C	3.9
at 220°C	0.7
Tensile strength as cast, lb/in.^{2}	
at 15°C	2100
at 200°C	650
at −40°C	2900
at −120°C	12,700

*1 cal = 4.184 J, 1 lb/in.^{2} = 6.895 kPa, °F = (°C × 1.8) + 32.

oxide, hydroxide, carboxylate, or thiolate. *See* TIN ALLOYS.

Joseph B. Long

Bibliography. P. W. Atkins et al., *Inorganic Chemistry*, 4th ed., 2006; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., 1999; D. R. Lide, *CRC Handbook of Chemistry and Physics*, 85th ed., 2004.

Tin alloys

Solid solutions of tin and some other metal or metals. Alloys cover a wide composition range and many applications because tin alloys readily with nearly all metals. *See* ALLOY.

Soft solders constitute one of the most widely used and indispensable series of tin-containing alloys. Common solder is an alloy of tin and lead, usually containing 20–70% tin. It is made easily by melting the two metals together. With 63% tin, a eutectic alloy melting sharply at 361°F (169°C) is formed. This is much used in the electrical industry. A more general-purpose solder, containing equal parts of tin and lead, has a melting range of 56°F (31°C). With less tin, the melting range is increased further, and wiping joints such as plumbers make can be produced. Lead-free solders for special uses include tin containing up to 5% of either silver or antimony for use at temperatures somewhat higher than those for tin-lead solders, and tin-zinc base solders often used in soldering aluminum. *See* SOLDERING.

Bronzes are among the most ancient of alloys and still form an important group of structural metals. Of the true copper-tin bronzes, up to 10% tin is used in wrought phosphor bronzes, and from 5 to 10% tin in the most common cast bronzes. Many brasses, which are basically copper-zinc alloys, contain 0.75–1.0% tin for additional corrosion resistance in such wrought alloys as Admiralty Metal and Naval brass, and up to 4% tin in cast leaded brasses. Among special cast bronzes are bell metal, historically 20–24% tin for best tonal quality, and speculum, a white bronze containing 33% tin that gained fame for high reflectivity before glass mirrors were invented. *See* BRONZE; COPPER ALLOYS.

Babbitt or bearing metal for forming or lining a sleeve bearing is one of the most useful tin alloys. It is tin containing 4–8% each of copper and antimony to give compressive strength and a structure desired for good bearing properties. An advantage of this alloy is the ease with which castings can be made or bearing shells relined with simple equipment and under emergency conditions. Aluminum-tin alloys are used in bearing applications that require higher loads than can be handled with conventional babbitt alloys. *See* ANTIFRICTION BEARING.

Pewter is an easily formed tin-base alloy that originally contained considerable lead. Thus, because Colonial pewter darkened and because of potential toxicity effects, its use was discouraged. Modern pewter is lead-free. The most favorable composition, Britannia Metal, contains about 7% antimony and 2% copper. This has desired hardness and luster retention, yet it can be readily cast, spun, and hammered. *See* PEWTER.

Type metals are lead-base alloys containing 3–15% tin and a somewhat larger proportion of antimony. As with most tin-bearing alloys, these are used and remelted repeatedly with little loss of constituents. Tin adds fluidity, reduces brittleness, and gives a structure that reproduces fine detail.

Flake and nodular gray iron castings are improved by adding 0.1% tin to give a fully pearlitic matrix with attendant higher hardness, heat stability, and improved strength and machinability.

Tin is commonly an ingredient in costume jewelry, consisting of pewterlike alloys and bearing-metal compositions often cast in rubber molds; in die castings hardened with antimony and copper for applications requiring close tolerances, thin walls, and bearing or nontoxic properties; and in low-melting alloys for safety appliances. The most common dental amalgam for filling teeth contains 12% tin. *See* TIN; TIN METALLURGY.

Bruce W. Gonser

Bibliography. American Society for Testing and Materials, *Annual Book of ASTM Standards*, vol. 02.04: *Nonferrous Metals*, 1993; B. T. K. Barry and C. G. Thwaites, *Tin and Its Alloys and Compounds*, 1983; *Conference on Tin Consumption*, International Tin Council, London, 1972.

Tin metallurgy

The extraction of tin from its ores and its subsequent refining and preparation for use. Most tin concentrates are primarily cassiterite (SnO_2), the naturally

occurring oxide of tin. These are comparatively easy to reduce by using carbon at high temperatures. However, this operation differs from the smelting of most common metals because retreatment of the slag is necessary to obtain efficient metal recovery. *See* CASSITERITE.

In primary smelting, carbon monoxide (CO) formed during heat-up reacts with the solid cassiterite particles to produce tin (Sn) and carbon dioxide (CO_2). As the temperature increases, silica (present in nearly all concentrates) also reacts under reducing conditions with the SnO_2 to give stannous silicate. Iron, also present as an impurity in all concentrates, reacts with the silica to form ferrous silicate ($FeSiO_3$). These silicates fuse with the added fluxes to form a liquid slag, at which point unreacted carbon from the fuel becomes the predominant reductant in reducing both stannous silicate to tin and ferrous silicate to iron. The metallic iron then reduces tin from stannous silicate, as shown in the reaction below.

$$SnSiO_3 + Fe \rightleftharpoons FeSiO_3 + Sn$$

Primary smelting can be effected in a reverberatory, rotary, or electric furnace with the choice being more dependent on economic than technical considerations. In the Far East, for example, reverberatory furnaces fired with anthracite coal are widely used. Both Malaysia and Singapore have added electric furnaces to improve smelting efficiencies. Indonesia and Singapore use slow-speed rotary furnaces. Reverberatory and rotary furnaces are also used in Indonesia. On the other hand, the smelters in Zaire and Rwanda as well as those in South Africa, which are far away from coal sources, use electric furnaces because of the availability of electric power. In the case of Bolivia which has complex concentrates that may range as low as 15% tin from lode mining, roasting may be needed as a pretreatment before smelting in order to remove such undesirable impurities as sulfur and arsenic plus some lead, antimony, and bismuth. *See* ELECTRIC FURNACE.

One of the greatest contributions to modern tin smelting has been the fuming of tin slags. Stimulated by the need for better metal recoveries, this process relies on the formation and volatilization of tin as stannic oxide (SnO_2) in a type of blast furnace. The process requires the addition of pyrites to the tin-rich slag, where it reacts to produce $FeSiO_3$ and stannous sulfide (SnS). The SnS vapor oxidizes to SnO_2 and is carried out in the furnace exhaust gases, from which it is collected and recycled.

Fuming is an alternative to roasting in the smelting of low-grade concentrates (5–25% tin). A tin oxide dust, free of iron, is obtained which is fed back to a conventional smelting furnace.

The crude tin from slags and smelted concentrates is further refined by heat treatment (that is, liquidation or boiling) or sometimes electrolytic processes. In liquidation, tin is heated on a sloping hearth to just above its melting point. The tin runs into a poling kettle, while metals with higher melting points remain in the dross. Most of the iron is removed in this manner as well as part of the arsenic, antimony, and copper. Lead and bismuth remain. In the final refining step, the molten tin is agitated with steam, compressed air, or poles of green wood which produce steam. The remaining traces of impurities form a scum which is removed and recirculated through the smelting cycle. The pure tin is cast in iron molds to ingots of about 99 lb (45 kg). Purity is guaranteed to exceed 99.8%.

Iron, copper, arsenic, and antimony can be readily removed by the above processes or variations on these. However, for removing large amounts of lead or bismuth, electrolysis or a vacuum-refining process is used.

Secondary tin from metal scrap amounts to about one-quarter of the total tin consumed in the United States. Most of this comes from tin-bearing alloys, and secondary smelters rework them into alloys and chemicals. However, additional tin of high purity is recovered from the detinning of tinplate scrap. *See* ELECTROCHEMICAL PROCESS; ELECTROMETALLURGY; HEAT TREATMENT (METALLURGY); PYROMETALLURGY, NONFERROUS; TIN.

Daniel Maykuth

Bibliography. American Society for Testing and Materials, *Annual Book of ASTM Standards*, vol. 2.04: *Nonferrous Metals*, 1993; T. S. Mackey, Review of recent developments in tin—1981, *J. Metals*, pp. 72–75, April 1982; P. A. Wright, *Extractive Metallurgy of Tin*, 1982.

Tintinnida

An order of the Spirotrichia whose members are conical or trumpet-shaped pelagic forms bearing shells (loricae). These protozoa are planktonic ciliates

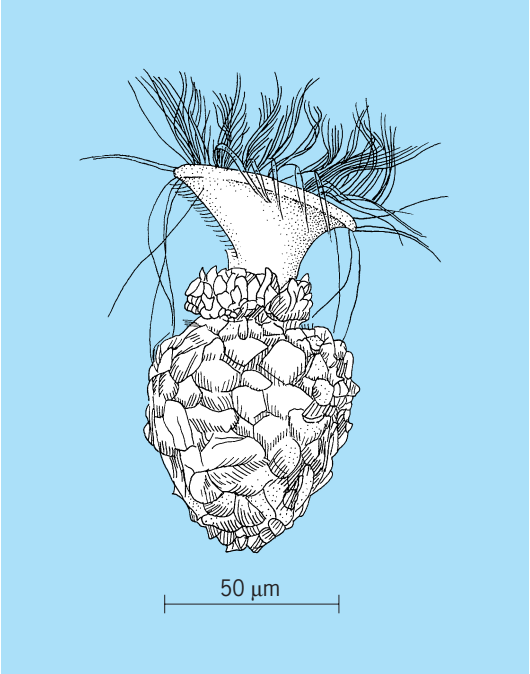


Fig. 1. ***Tintinnopsis*****, a living specimen shown protruding from its lorica, or shell.**

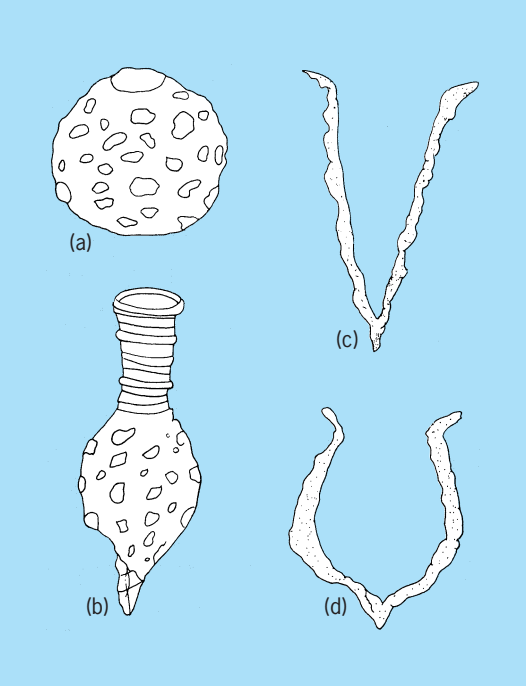


Fig. 2. Fossil and modern Tintinnida. (*a*) *Tintinnopsis*, Jurassic to Recent. (*b*) *Codonellopsis*, Recent. (*c*) *Amphorellina* (section), Lower Cretaceous. (*d*) *Calpionella* (section), Recent.

and are especially abundant in oceans, notably the Pacific. The lorica is composed of a resistant organic compound in which various foreign mineral grains are embedded; its shape may range from trumpet- or bell-form to cylindrical or subspherical, and its size from 50 to 200 micrometers. The exact structure, often quite elaborate, and the dimensions of the lorica are so recognizably different among the hundreds of known genera that the taxonomic arrangement of forms within the order has been based solely on characteristics and properties of this secreted "house." The adoral zone of membranelles (**Fig. 1**) is prominent, while the other ciliature is greatly reduced.

Fossil tintinnids, representing practically the only fossilized species of ciliate protozoa known to science, are identified on the basis of the shape of the lorica in cross section as seen in randomly oriented thin sections of the rocks in which they are found. Twelve genera of fossil tintinnids have been described from limestones and cherts of the Jurassic and Cretaceous.

Common genera are illustrated. Figure 1 shows a present-day member of a species of the genus *Tintinnopsis*. **Figure 2** shows drawings of loricae, whole and in section, including some fossil forms. *See* CILIOPHORA; PROTOZOA; SPIROTRICHIA.

John O. Corliss; Daniel J. Jones

Tire

A continuous pneumatic rubber and fabric cushion encircling and fitting onto the rim of a wheel. Sizes range from only a few inches in diameter up to 12 ft (3.7 m) and 12,500 lb (5700 kg).

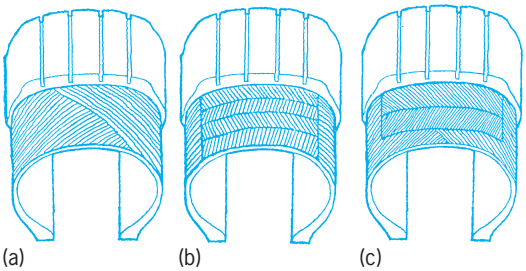
In modern tire building, rubber (both natural and synthetic) and fabric remain the basic ingredients. Chemicals are compounded into the rubber to help it withstand wear, heat, and aging and to produce desired changes in its characteristics. Fabric (rayon, nylon, or polyester) is used to give the tire body strength and resilience. In belted tires, additional layers of fabric (rayon, fiber glass, finely drawn steel, or aramid) are placed just under the tread rubber to increase mileage and handling. Steel wire is used in the bead that holds the tire to the rim.

A tire is made up of two basic parts: the tread, or road-contacting part, which must provide traction and resist wear and abrasion, and the body or carcass, consisting of rubberized fabric that gives the tire strength and flexibility. In compounding the rubber, large amounts of carbon black are mixed with it to improve abrasion resistance. Other substances, such as sulfur, are added to enable satisfactory processing and vulcanization. *See* RUBBER.

Manufacture. The basic part of a tire-building machine is a collapsible cylinder shaped like a wide drum that turns under power and is controlled by the tire builder. First, plies of rubberized fabric are wrapped around the drum. The beads then are placed in position, and the sides of the plies are wrapped around them. If the tire is of belted bias or radial construction, belts of rubberized material are centered on the plies. Other narrow strips of material are placed near the beads for further strength when the tire flexes and rubs against the rim. Next, the tread and sidewall rubber is wrapped around the drum over the fabric. All of the components, which can number more than 20 in radial auto tires, are pressed together with rollers. The drum is collapsed and the tire is removed, ready to be molded and vulcanized. At this point the tire looks like a barrel with both ends open.

The tire still needs shaping and curing (or vulcanizing) to gain its final shape and strength characteristics. Besides molding the tread design and equalizing the stresses within the tire body, the vulcanization changes the rubber compound into a tough, highly elastic material and bonds the parts of a tire into one integral unit. When the tire emerges from the curing press, the building process is complete.

Types. There are three types of tires: bias-ply, radial, and belted bias (see **illus.**). For bias tires, cords in the plies extend diagonally across the tire



Tire construction. (*a*) Bias-ply. (*b*) Radial. (*c*) Belted bias. (*Goodyear Tire and Rubber Co.*)

from bead to bead. The cords run in opposite directions in each successive ply, resulting in a crisscross pattern. For radial tires, cords in the plies extend transversely from bead to bead, substantially perpendicular to the direction of travel. Belts are placed circumferentially around the tire. For belted bias tires, plies are placed in a manner similar to that used in the bias-ply tire, with belts of material placed circumferentially around the tire between the plies and the tread rubber.

Developments. Although most tire improvements appear gradually, a number of important developments have marked great advancements.

Fabric. Rayon was introduced in the late 1930s as a replacement for cotton. Nylon followed in the late 1940s, and it remains the basic material in truck, earthmover, and aircraft tires. Polyester, combining the best features of rayon and nylon, was first used in auto tires in the early 1960s; it became the most used tire cord and is in virtually all auto tires.

Tubeless tires. Prior to the mid-1950s all tires had to have inner tubes to contain the air pressure. The development of the tubeless tire brought increased puncture resistance and less heat buildup.

Belted tires. The belted bias tire was developed in the late 1960s to increase tread life and tire performance. Then in the 1970s the radial tire, long popular in Europe, won acceptance in the United States and became the most popular form of auto tire construction. Fiber glass, steel, and finally aramid were developed as materials for the belts.

Rubber. Early tires were totally dependent on natural rubber, which was often poor in quality. During World War II synthetic rubber was developed and now accounts for about 80% of the rubber used by the tire industry. Other compounds have allowed for greatly improved traction on ice, even without metal studs. Compound development also has led to lower rolling resistance, improved gasoline mileage, and longer tread life. *See* RUBBER. David B. Harrison

Bibliography. J. C. Dixon, *Tires, Suspension, and Handling*, 2d ed., 1996; H. B. Pacejka, *Tire and Vehicle Dynamics*, 2d ed., 2005; Society Of Automotive Engineers, *Tire and Wheel Technology*, 1999.

Tissue

An aggregation of cells more or less similar morphologically and functionally. The animal body is composed of four primary tissues, namely, epithelium, connective tissue (including bone, cartilage, and blood), muscle, and nervous tissue. The process of differentiation and maturation of tissues is called histogenesis. *See* CONNECTIVE TISSUE; EPITHELIUM; PLANT TISSUE SYSTEMS. Charles B. Curtin

Tissue culture

The branch of biology in which tissues or cells of higher animals and plants are grown artificially in a controlled environment. Such studies were undertaken in the hope that the behavior of various body components could be studied and their potentialities more readily analyzed under the simpler and more readily manipulated conditions possible in the test tube.

Study of the growth and interaction of animal cells with physical and chemical environments outside the body began about 1900. During the first decade it was demonstrated that cell multiplication from chick tissue transplanted into glass vessels could be maintained indefinitely, if suitable physical conditions for cell attachment to a solid substrate were provided, and if the necessary complex nutrient medium was replenished as fast as it was depleted by the cells' activities. From early beginnings, tissue culture has developed in many directions (**Fig. 1**).

Early methods of tissue culture were successful in promoting cell multiplication only when large numbers of cells were seeded together in a community. Growth of such large populations permitted many kinds of important experimentation on the multiplication process, but it also left many questions unsolved. For example, it was impossible to determine which fraction of the cells of any population had retained the ability to multiply. It also was difficult to determine the specific conditions which the individual cell requires in order to be able to initiate its reproductive process. A major advance along these lines was made by a group of scientists who succeeded in providing conditions permitting growth of single cells when individually sealed in capillary tubes. These cells were later grown into huge populations so that it was demonstrated that at least some cells which had originated in the mammalian body maintain their ability to multiply indefinitely in isolation, just like independent bacteria.

In early tissue culture, growth of cells was successful only when they were attached to a solid substrate

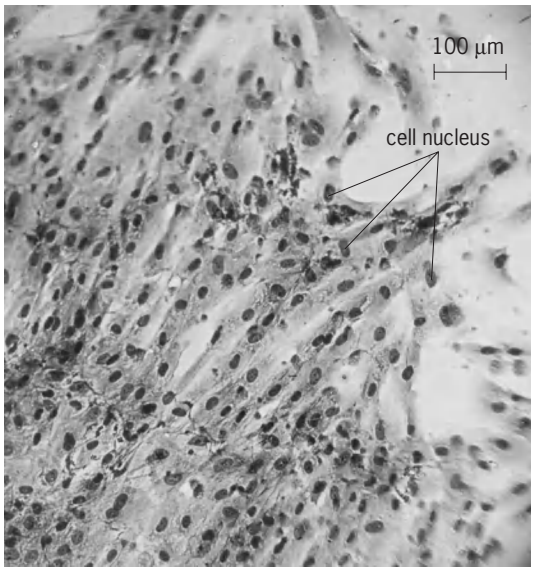


Fig. 1. Human spleen cells grown in a glass vessel containing a nutrient medium.

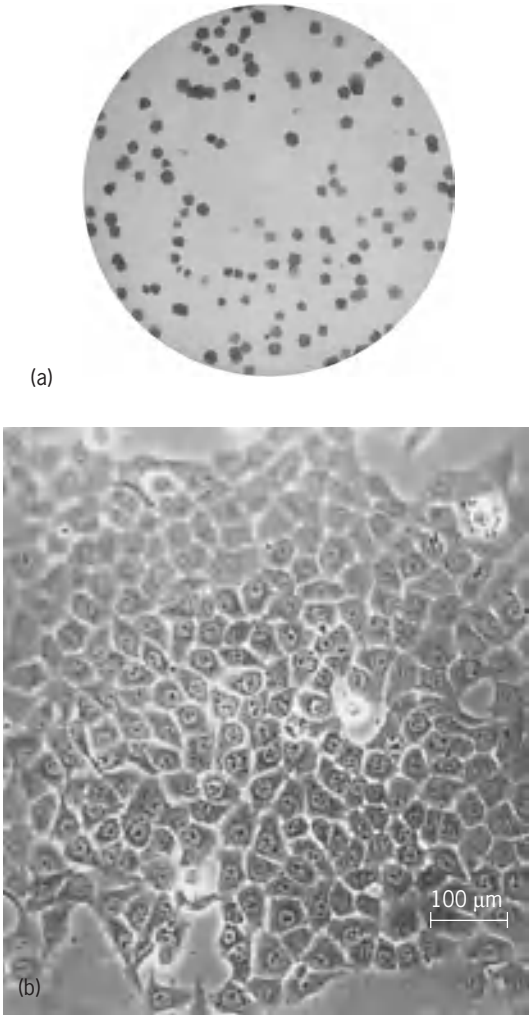


Fig. 2. Colonies developed from single human cancer (HeLa) cells. (*a*) Colonies grown on a glass dish. (*b*) Photomicrograph of a typical colony.

like glass or cellophane. In 1954 it became possible to grow cells in liquid suspension as well, a technique that permits many new operations, such as continuous farming of cells in the same vessel for indefinite periods. In addition, it became possible to simplify greatly the medium required for cell multiplication so as to eliminate, at least in some cases, the need for animal serum. Definition of the chemical requirements for mammalian cell growth in test tubes proceeded in a variety of laboratories and reached the stage wherein massive populations can be grown for long periods in a completely molecularly defined medium. Single cells can be reliably grown into discrete colonies in a medium containing completely defined, small-molecular weight constituents and a purified protein fraction obtained from blood. The completely defined small-molecular weight constituents are salts, amino acids, glucose, choline and inositol, and the vitamins biotin, pantothenic acid, folic acid, niacinamide, pyridoxine, riboflavin, thiamine, and B_{12}. Such advances have been tremendously important in establishing the specific nutrient requirements for different types of mammalian cells, and for elucidation of the metabolic pathways taken by the different molecules, in both healthy and diseased subjects.

Quantitation. In research undertaken to permit quantitative measurement of cell growth, means were found by which animal cells grown in tissue culture could be dispersed singly, then added to a glass dish, under conditions wherein every single cell would reproduce in isolation to form a discrete colony (**Fig. 2**). This aim was, at first, successfully achieved for cancer cells and then was also achieved for cells from normal human and animal tissues. This method of "plating" single mammalian cells made possible many more precise kinds of experiments. The effects of different physical and chemical agents on growth of cells could now be measured with much greater accuracy, since the number of cells able to reproduce under the required conditions could be precisely determined by a simple colony count. It also became possible to measure accurately and conveniently the growth rate of such single cells in different media (**Fig. 3**).

Study of hereditary mechanisms. Development of single-cell techniques afforded tissue-culture means for study of hereditary mechanisms in animal cells.

Cells grown in tissue culture by older techniques have demonstrated changes in their chromosome numbers and structures occurring with the passage of time. Since these bodies contain the genes which determine the hereditary potentialities of the cells, the genetic constitution of such cells was uncertain. Inferences drawn from the behavior of such cultures and then applied to interpretation of functions and potentialities of cells in the body, where chromosomal integrity is rarely altered, were often of doubtful significance. Various investigators then turned attention to these problems and developed new methods for study of the chromosomal constitution of mammalian cells in tissue culture.

One of the results of such advances in technique was the production of methods for regulating cell growth through extended periods of cultivation, so that the chromosomal integrity was maintained as reliably as it is in the body. It thus became possible to make many kinds of biochemical and genetic studies

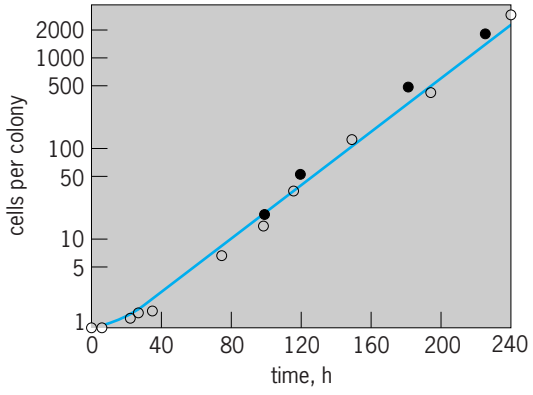


Fig. 3. Typical growth curve of single cells plated in a complete, nutrient medium. Cells begin to reproduce after initial delay of about 18 h and continue to double every 20 h as long as medium is not exhausted.

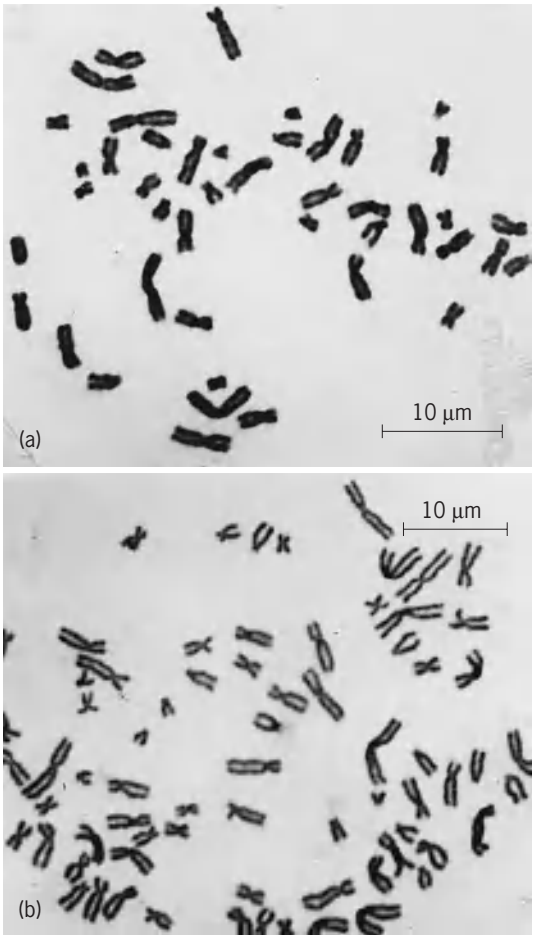


Fig. 4. Chromosome constitution. (*a*) Cells taken from normal human male and grown in tissue culture; 46 chromosomes, constant in number. (*b*) Human cancer; the chromosome number and structure are changed and variable from cell to cell.

on such cells with considerably more confidence in the applicability of results to an understanding of cell functions in the intact animal.

Advances in methods of taking cell specimens for culture reached the point where it became possible to obtain with ease a cell sample from a minute amount of skin taken from any individual and to cultivate these cells stably in the test tube.

Another fundamental result arising from chromosomal studies in tissue culture was the demonstration that the chromosome constitution of humans, which had been accepted for approximately 30 years, was in error. In 1956, tissue-culture studies demonstrated that the normal cells of humans contain 46 instead of 48 chromosomes. These results were verified in many laboratories by study of cells drawn from a variety of tissues of many human subjects. However, the occasional occurrence of a chromosome number other than 46 has been demonstrated in human subjects with genetic disease.

Diseases like ovarian dysgenesis (Turner's syndrome) have been found to be accompanied by a chromosome number of 45 instead of the normal 46. The missing one is a sex chromosome, so that such individuals have the X O sex chromosomal constitution. Many hitherto mysterious human diseases are now known to be due to specific chromosomal abnormalities. Thus, Down syndrome, a birth defect which results in a subnormal intelligence and certain physical signs, is now known to be caused by an extra chromosome of pair number 21. *See* DOWN SYNDROME; HUMAN GENETICS.

The universally used system for identification and classification of the human chromosomes was devised in 1959 by an international group and is known as the Denver System. Each one of the human chromosomes was identified and its structure delineated. In contrast to a human cell with its normal complement of chromosomes, the karyotype of human cancer cells has been studied and many of these have been found to possess abnormally high

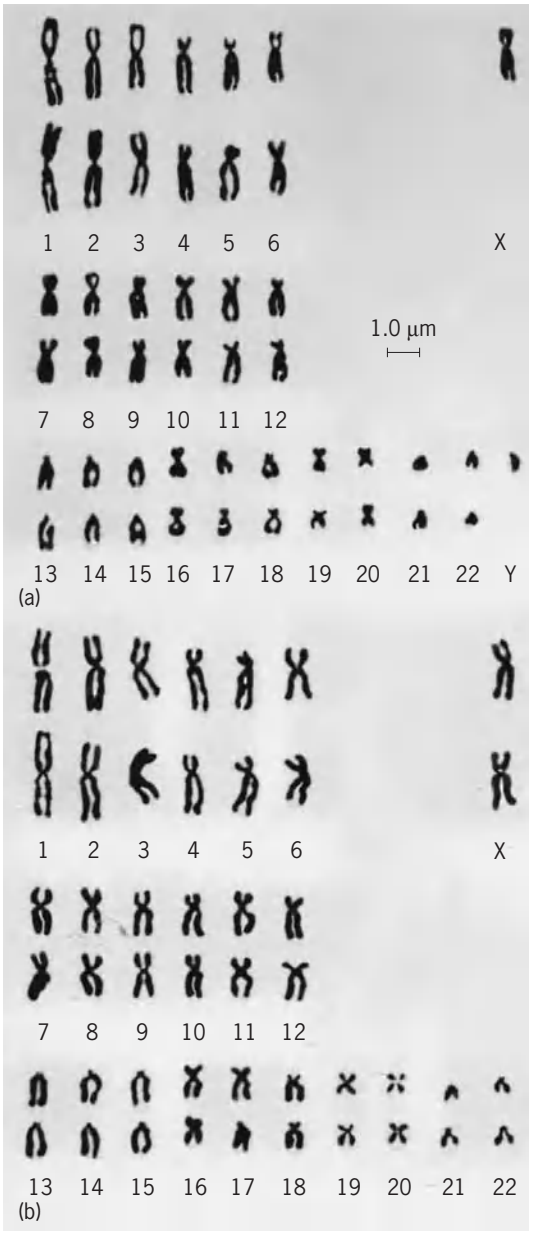


Fig. 5. The normal human chromosome constitution. (*a*) Male cells. (*b*) Female cells. Newer methods of specimen preparation reveal characteristic chromosome bands which also aid in identification of the chromosome.

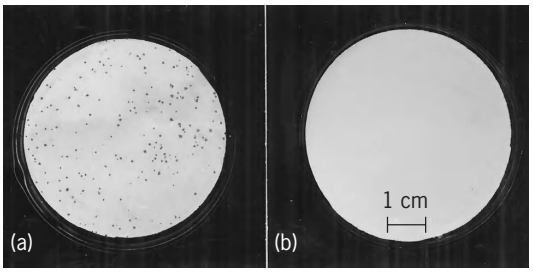


Fig. 6. Behavior of different mutant cells from the same cell population. Both cell types are identical in appearance and general behavior. Two hundred cells of each kind were placed in identical dishes containing the same nutrient medium. (*a*) The original cell reproduced and developed approximately 200 colonies. (*b*) The mutant cell produced no colonies whatever. However, if the amino acid glycine is added to the mutant cell, it produces colonies exactly like the unmutated cell type.

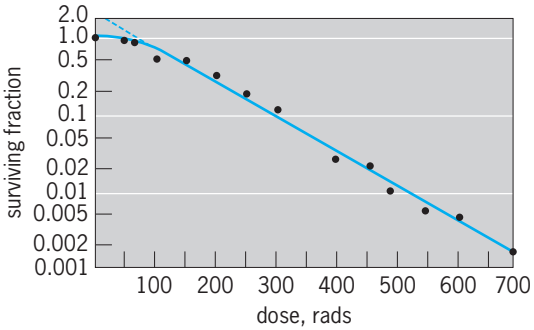


Fig. 7. Survival of reproductive capacity of cells from culture of a human cancer. The average lethal dose obtained from this curve is 96 rads.

chromosome numbers (**Fig. 4**). A karyotype is an idiogram of the chromosome complement characteristic of any individual or group of related organisms. The normal karyotype of human somatic cells is shown in **Fig. 5**. The Y chromosome, possessed only by males, is less than one-third the size of the X chromosome, which occurs singly in male cells but is doubled in female cells. All other chromosomes have been characterized by their sizes, by the position of their centromeres, and by the position of characteristic bands.

In studies of the genetic biochemistry of mammalian cells, methods were developed for the production and isolation of mutant cell cultures, whose biochemistry could be studied in the same fashion that was so successful in elucidating the molecular biology of bacteria. Mutants of cell populations were obtained which differ in their requirements for growth in the test tube (**Fig. 6**), and their underlying biochemistry has been shown. This research makes it possible to measure accurately the power of various physical and chemical agents to produce mutations in mammalian cells. Mutational rates have been measured for several mutagens. *See* MUTAGENS AND CARCINOGENS.

There are methods for locating the genes on their chromosomes, and powerful techniques for rapid and accurate measurement of the enzyme contents of cells with different genetic constitutions. With these tools the links between genes, enzymes, and specific developmental processes in different tissues are demonstrable in mammals.

Radiation studies. Investigations of the effect of high-energy irradiation on single human and other animal cells have accurately defined the dose required to prevent colony formation by reproduction of single cells (**Fig. 7**). This dose is approximately 100 rads for most mammalian cells. Before these studies it had been generally considered that the human cell is many times more resistant to radiation than this figure implies. These measurements made possible interpretation of the effect of high-energy radiation on the human body in a manner that satisfactorily explains many previously obscure aspects of radiation pathology. Study of the mechanisms by which radiation interferes with reproductive processes in mammalian cells has shown that most damage is confined to the cellular genetic apparatus and particularly to the chromosomes. **Figure 8** presents a series of pictures showing normal human chromosomes and those from human cells irradiated with various doses of x-rays. *See* RADIATION INJURY (BIOLOGY).

Use in virus studies. Tissue-culture studies opened a new era in investigations of mammalian viruses. These techniques made it possible to prepare large quantities of viruses for the work of virologists and immunologists. Thus the preparation of vaccines was speeded, as well as the performance of a wide variety of physical, chemical, and biological studies on these infectious agents.

Tissue-culture studies also opened whole new vistas of understanding of the virus-cell interaction. It became possible to adapt to tissue culture the plaque technique, a method in which single virus particles are recognized by the round area of cell destruction they produce (**Fig. 9**); thus an accurate determination can be made of the number of virus particles in the original suspension. This technique allows a much more precise measurement of the effects of viruses under different kinds of controlled physical and chemical situations. Since all the virus progeny of a single plaque area have descended from the same individual, it becomes more readily possible to carry out controlled genetic investigations on viruses. *See* VIRUS.

Mammalian cell molecular biology. Molecular biology was formed from the amalgamation of two previously distinct scientific fields, biochemistry and genetics, and it achieved an explosion of understanding of the simplest living cells. Molecular biological studies with bacteria furnished the first blueprint for the workings of a living cell: The genetic substance is a group of specific molecules of known chemical structure, deoxyribonucleic acid (DNA); these molecules are capable of self-replication; and these molecules which constitute the genes regulate the biosynthesis of the cell proteins from moment to moment throughout the life of the cell. Therefore, since DNA is responsible for construction of the cell machinery, it is ultimately responsible for the whole range of biochemical activity exhibited by any cell.

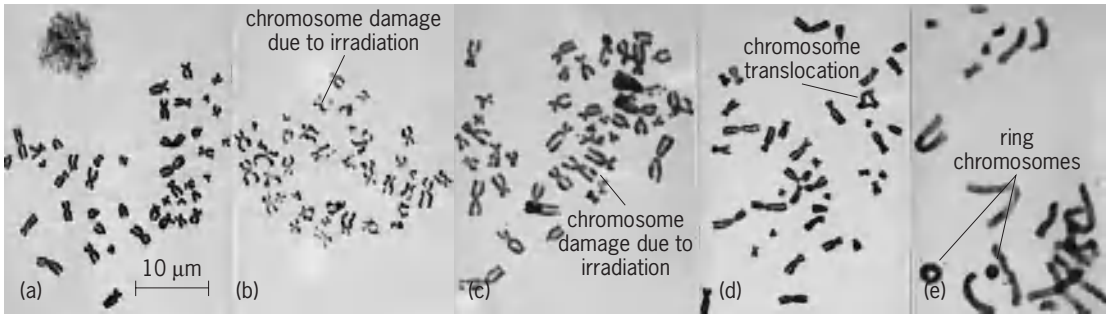


Fig. 8. Typical chromosomal lesions produced in human cells grown in tissue culture and irradiated with various doses of x-rays. (*a*) Normal human chromosome complement, unirradiated cell. (*b*) Chromosomes of cell after irradiation with 50 rads. (*c*) Chromosomes of cell irradiated with 75 rads. (*d*) Chromosomes of cell irradiated with 75 rads. Translocations have appeared as a result of abnormal restitution of the multiple breaks. (*e*) Development of ring chromosomes as a result of irradiation of a normal human cell with 150 rads.

See DEOXYRIBONUCLEIC ACID (DNA); MOLECULAR BIOLOGY.

An attempt to apply a similar approach to humans was not possible before 1956, because human genetics was fragmentary and weak. This situation existed for two reasons. The reproductive period of the simple bacterial cells is 20 min, while the generation time in humans is almost 25 years. Of equal significance was the fact that it is not possible in humans to make those matings which would be most illuminating genetically. Perhaps the greatest triumph of tissue culture in this century was the development of single-cell plating methods, which made possible study of human and mammalian genetics in the test tube by the methods of cloning which had been demonstrated to be so powerful with the simple bacterial cells. *See* GENETIC ENGINEERING.

In addition to study at the chromosome level, the human genetic structure has been examined at the single-gene level. Large numbers of genes have been identified and their protein products isolated. These developments made possible the study of biochemical pathways in human cells with great precision, and have illuminated previously mysterious diseases in which particular biochemical steps are defective. *See* MOLECULAR PATHOLOGY.

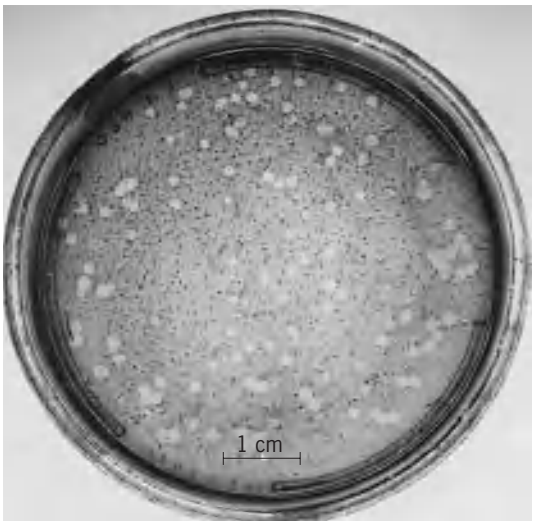


Fig. 9. Plaques or areas of cell destruction produced by the action of the virus of Newcastle disease on chick fibroblasts cultivated in a glass dish.

Another important development was the process of cell hybridization, which in effect makes possible the mating of cells in the test tube. Examining the progeny of such cellular mating processes uncovers details of human genetic structure and function. This procedure made possible determination of dominance and recessiveness in human genes in a very simple and definitive fashion. In addition, it made possible mapping of the human genes on their chromosomes, which is necessary in order to tell which genes are closely linked and which are far apart. Perhaps of even greater importance, however, is the need to understand the mechanism of gene regulation and how it is affected by gene position. Mammalian cells differ from simple bacterial cells in possessing the enormously complex property of differentiation in which different genetic regions are active in cells of the different tissues, permitting each cell grouping to act in its own highly characteristic fashion. In the simple bacterial cells, contiguous genes are often turned on and off together. Elucidation of mechanisms of this kind will provide understanding of the differentiation process and of the many different diseases which are due to defective gene regulation either in the course of earlier embryonic development or later in life. There is strong evidence to indicate that at least some aspects of the aging process and various degenerative diseases are due to failures in gene regulation. *See* CELL DIFFERENTIATION; CELL SENESCENCE; GENETIC MAPPING; SOMATIC CELL GENETICS.

One of the most important developments in this respect has been gene mapping at every level from that of parts of the DNA molecule to the entire chromosome. In addition to providing the fundamental data which should eventually make gene regulation understandable, these techniques provide new methods of diagnosis of genetic disease. These methods are applicable even very early in pregnancy, so that the presence of severe genetic disability can often be determined in time to terminate the pregnancy if the prospective parents choose. The procedure could make possible prevention of many human genetic tragedies.

Mutant cells with respect to regulatory processes involved in critical diseases like those dealing with cholesterol metabolism have been studied both in individuals and in the test tube. Methods were developed for isolating and identifying the proteins which are the characteristic gene products of mammalian cells. Receptor sites which are the targets for the action of hormones have been identified in a variety of tissue culture studies. Along with these developments in genetic biochemistry, tissue culture has permitted elucidation of cell structures which were previously unsuspected. *See* CELL (BIOLOGY).

There are also tissue culture methods for monitoring environmental agents for their ability to cause gene and chromosomal mutations, which underlie human genetic disease. It is now presumed that the same kinds of defects are also responsible for a large proportion of cancers. The use of these methods may also make it possible in a single operation to develop protection against agents which are responsible for a great deal of human genetic disease and cancer.

Other techniques. Many advances were carried out by cinema photomicrography, which portrayed the surprising variety of movements of different kinds of cells growing in tissue culture. It was shown that cells taken from a whole animal and dispersed individually have the ability to reaggregate in tissue culture to form tissues very much like those of their original site. *See* CELL ADHESION.

There are other techniques for mapping with great precision the point at which various biochemical steps occur in the life cycle of a cell. In this way, the specific events of the cell's reproductive cycle appear to be capable of delineation and the point in the life cycle at which various drugs and other agents exert their effects can now be determined. *See* CELL CYCLE.

Cells in tissue culture can be made cancerous by treatment with certain viruses, x-irradiation, or carcinogenic hydrocarbons. They usually acquire specific changes in their patterns of biochemical activity, and they lose the capacity to inhibit one another's reproduction as occurs when noncancerous cells are crowded together in a tissue culture. *See* MUTAGENS AND CARCINOGENS.

A number of techniques produce specific differentiation of cells in tissue culture so that they can carry out the biochemical steps characteristic of various organs of the body. Various hormones are effective in causing nondifferentiated cells to synthesize specific enzymes like those characteristic of normal cells. Under the influence of feeder layers, layers of cells whose reproductive power has been suppressed by x-irradiation, cells from muscle or bone marrow will multiply and, at the same time, develop differentiated functions characteristic of their tissue of origin.

While this discussion has been confined to animal tissue culture, mention should be made that there have been many parallel developments in the field of plant tissue culture. Theodore T. Puck

Bibliography. R. F. Beers, Jr., and E. G. Bassett (eds.), *Cell Fusion: Gene Transfer and Transformation*, 1984; L. Goldstein and D. M. Prescott (eds.), *Cell Biology: A Comprehensive Treatise*, vol. 4, 1981; B. M. Martin, *Tissue Culture Techniques: An Introduction*, 1994; T. T. Puck, *The Mammalian Cell as a Microorganism*, 1972; A. E. Sowers (ed.), *Cell Fusion*, 1987; T. A. Springer (ed.), *Hybridoma Technology in the Biosciences and Medicine*, 1985.

Tissue typing

A procedure involving a test or a series of tests to determine the compatibility of tissues from a prospective donor and a recipient prior to transplantation. The immunological response of a recipient to a transplant from a donor is directed against many cell-surface histocompatibility antigens controlled by genes at many different loci. However, one of these loci, the major histocompatibility complex (MHC), has the greatest genetic complexity and controls antigens that evoke the strongest immunological response. The MHC is a cluster of closely linked gene loci and is conserved in all vertebrate species, including humans. The human MHC is known as the HLA system, which stands for the first (A) Human Leukocyte blood group system discovered. The HLA complex was first described as a gene locus on chromosome 6 that controls the allograft rejection response, which is the rejection of a graft from a donor by a genetically dissimilar recipient of the same species. It has since been found, however, that these genes are also physiologically important in the regulation of the immune response to highly foreign antigens such as bacterial or viral antigens, as well as to self-antigens. *See* CELLULAR IMMUNOLOGY; HISTOCOMPATIBILITY.

The success of transplantation is greatly dependent on the degree of histocompatibility (identity) between the donor and recipient, which is determined by the HLA complex. When the donor and recipient have a low degree of histocompatibility, the organ is said to be mismatched, and the recipient mounts an immune response against the donor antigen. The intensity of this response depends on the haplotype combination of HLA locus alleles, since certain alloantigens are more immunogenic than others. Additionally, previous sensitization of a potential transplant recipient to HLA antigens through pregnancy, blood transfusion, or a prior transplant may result in preexisting anti-HLA antibodies. If these preformed antibodies are specific for the donor HLA antigens, hyperacute rejection of transplanted tissue can occur. Therefore, histocompatibility testing plays an important role in the selection of donors and recipients for organ transplantation. By laboratory testing, the degree of antigenic similarity between the donor and the recipient and the degree of preexisting recipient sensitization to donor antigens can be determined. This is known as cross-matching. The success of transplantation depends largely on the degree of MHC compatibility between donor and recipient. and an absence in the recipient of preformed antibodies to donor antigens.

Phenotyping of HLA-A, -B, and -C (ABC typing) of an individual is determined by reacting that individual's lymphocytes with a large panel of antisera directed against specific HLA antigens. These antisera are generally obtained from women who have had two or more pregnancies and are selected so that, collectively, all significant HLA antigens are likely to be encountered. The procedure is known as complement-mediated cytotoxicity assay. The person's lymphocytes are incubated with the different antisera and complement is added. Killing of the cells being tested indicates that they express the HLA determinants recognized by the particular antiserum being used. Killing of potential donor lymphocytes in the complement-mediated cytotoxicity assay indicates the presence of antibodies specifically directed against HLA antigens, and is a contraindication to transplantation of tissue from that donor. *See* COMPLEMENT; HYPERSENSITIVITY; IMMUNOASSAY.

Assistance in tissue typing is being sought with additional techniques. These include the use of primed lymphocytes (lymphocytes that have been stimulated previously by contact with lymphocytes bearing target HLA specificities), monoclonal antibodies directed against selected specificities, and direct analysis of deoxyribonucleic acid (DNA) either by means of suitable probes and the polymerase chain reaction or by limited enzymatic digestion followed by analysis of the restriction fragment length polymorphism.

In addition to its important role in organ transplantation, determination of the HLA phenotype is useful in paternity testing, forensic medicine, and the investigation of HLA-disease associations. Although the immune system is carefully regulated, it can react against the individual's own tissues (autoimmunity). Several autoimmune disorders, such as ankylosing spondylitis, rheumatoid arthritis, and diabetes, have been linked to a high degree with certain HLA antigens. *See* AUTOIMMUNITY; TRANSPLANTATION BIOLOGY.

M. Wayne Flye; T. Mohanakumar

Bibliography. E. Albert et al., Nomenclature for factors of the HLA system—1977, *Tissue Antigens*, 11:81–86, 1978; E. Albert, M. P. Bauer, and W. R. Mays (eds.), *Histocompatibility Testing, 1984*, 1985; P. Dyer and D. Middleton (eds.), *Histocompatibility Testing: A Practical Approach*, 1993; M. W. Flye (ed.), *Principles of Organ Transplantation*, 1989; G. Opelz, Correlation of HLA matching with kidney graft survival in patients with or without cyclosporine treatment, *Transplantation*, 40:240–243, 1985; R. Patel and P. I. Terasaki, Significance of a positive crossmatch test in kidney transplantation, *N. Engl. J. Med.*, 288:735–736, 1969.

Titanite

A calcium, titanium silicate, $CaTiOSiO_4$, of high titanium content. Titanite is also known as sphene.

Titanite is an orthosilicate (nesosilicate) in which silicate $[(SiO_4)^{-4}]$ tetrahedra do not share any apical oxygens with adjacent tetrahedra but are cross-linked by chains of octahedrally (sixfold) coordinated titanium ions (Ti^{4+}) and calcium ions (Ca^{2+}) that are coordinated to seven oxygens. This yields a monoclinic structure with the space group C2/c. It has a hardness of 5–5$^1/_2$ on the Mohs scale, a distinct cleavage, a specific gravity of 3.4–3.55, and an adamantine to resinous luster. It commonly occurs as distinct wedge-shaped crystals that are usually brown in hand specimens. Titanite may also be gray, green, yellow, or black. *See* COORDINATION CHEMISTRY; CRYSTAL STRUCTURE.

Titanite is a common accessory mineral in many igneous and metamorphic rocks. It may be the principal titanium-bearing silicate mineral, especially in intermediate and alkali-rich intrusive igneous rocks such as nepheline syenites. It occurs in abundance in the Magnet Cove, igneous complex in Arkansas and in the intrusive alkalic-rocks of the Kola Penninsula, Russia. It is also common in metamorphosed mafic rocks (that is, metabasalts) and metamorphosed impure limestones and dolostones. In metamorphic rocks, titanite's stability at high temperature is limited by reactions resulting in the formation of rutile (TiO_2) as the dominant titanium-bearing mineral. Titanite grains may also occur in detrital sediments and sedimentary rocks.

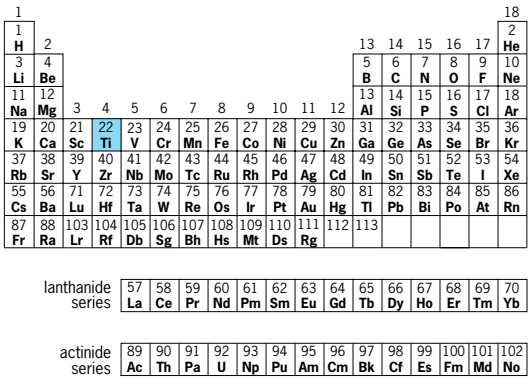
The composition of titanite may diverge from pure $CaTiSiO_4$ because of a variety of chemical substitutions. Calcium ions (Ca^{2+}) can be partially replaced by strontium ions (Sr^{2+}) and rare-earth ions such as thorium (Th^{4+}) and uranium (U^{4+}). Aluminum ions (Al^{3+}), ferric iron (Fe^{3+}), and ferrous iron (Fe^{2+}) may substitute for titanium ions (Ti^{4+}), whereas oxygen ions (O^{2-}) may be replaced by hydroxyl ions (OH^-), fluoride ions (F^-), and chloride ions (Cl^-). Because titanite commonly contains radioactive elements, it has been used for both uranium-lead and fission track methods of dating. *See* DATING METHODS; IGNEOUS ROCKS; METAMORPHIC ROCKS; SILICATE MINERALS; TITANIUM.

John C. Drake

Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *Rock Forming Minerals*, vol. 1: *Ortho- and Ring Silicates*, 1962; *Geology and Resources of Titanium*, U.S. Geol. Survey Prof. Pap. 959-E, 1976; J. A. Hunt and D. M. Kerrick, The stability of sphene: Experimental redetermination and geological implications, *Geochim. Cosmochim. Acta*, 41:279–288, 1977; C. Klein and C. S. Hurlbut, Jr., *Manual of Mineralogy*, 21st ed., rev. 1999; P. H. Ribbe (ed.), *Orthosilicates*, 1982.

Titanium

A chemical element, Ti, atomic number 22, and atomic weight 47.90. It occurs in the fourth group of the periodic table, and its chemistry shows many similarities to that of silicon and zirconium. On the other hand, as a first-row transition element, titanium has an aqueous solution chemistry, especially of the lower oxidation states, showing some resemblances to that of vanadium and chromium. *See* PERIODIC TABLE; TRANSITION ELEMENTS.



The catalytic activity of titanium complexes forms thebasis of the well-known Ziegler process for the polymerization of ethylene. This type of polymerization is of great industrial interest since, with its use, high-molecular-weight polymers can be formed. In some cases, desirable special properties can be obtained by forming isotactic polymers, or polymers in which there is a uniform stereochemical relationship along the chain. *See* POLYOLEFIN RESINS.

The dioxide of titanium, TiO_2, occurs most commonly in a black or brown tetragonal form known as rutile. Less prominent naturally occurring forms are anatase and brookite (rhombohedral). Both rutile and anatase are white when pure. The dioxide may be fused with other metal oxides to yield titanates, for example, K_2TiO_3, $ZnTiO_3$, $PbTiO_3$, and $BaTiO_3$. The black basic oxide, $FeTiO_3$, occurs naturally as the mineral ilmenite; this is a principal commercial source of titanium.

Titanium dioxide is widely used as a white pigment for exterior paints because of its chemical inertness, superior covering power, opacity to damaging ultraviolet light, and self-cleaning ability. The dioxide has also been used as a whitening or opacifying agent in numerous situations, for example as a filler in paper, a coloring agent for rubber and leather products, a pigment in ink, and a component of ceramics. It has found important use as an opacifying agent in porcelain enamels, giving a finish coat of great brilliance, hardness, and acid resistance. Rutile has also been found as brilliant, diamondlike crystals, and some artificial production of it in this form has been achieved. Because of its high dielectric constant, it has found some use in dielectrics.

The alkaline-earth titanates show some remarkable properties. The dielectric constants range from 13 for $MgTiO_3$ to several thousand for solid solutions of $SrTiO_3$ in $BaTiO_3$. Barium titanate itself has a dielectric constant of 10,000 near 120°C (248°F), its Curie point; it has a low dielectric hysteresis. These properties are associated with a stable polarized state of the material analogous to the magnetic condition of a permanent magnet, and such substances are known as ferroelectrics. In addition to the ability to retain a charged condition, barium titanate is piezoelectric and may be used as a transducer for the interconversion of sound and electrical energy. Ceramic transducers containing barium titanate compare favorably with Rochelle salt and quartz, with respect to thermal stability in the first case, and with respect to the strength of the effect and the ability to form the ceramic in various shapes, in the second case. The compound has been used both as a generator for ultrasonic vibrations and as a sound detector. *See* PIEZOELECTRICITY.

Arthur W. Adamson

In addition to important uses in applications such as structural materials, pigments, and industrial catalysis, titanium has a rich coordination chemistry. The formal oxidation of titanium in molecules and ions ranges from −II to +IV. The lower oxidation states of −II and −I occur only in a few complexes containing strongly electron-withdrawing carbon monoxide ligands.

The lower oxidation states of titanium are all strongly reducing. Thus, unless specific precautions are taken, titanium complexes are typically oxidized rapidly to the +IV state. Moreover, many titanium complexes are extremely susceptible to hydrolysis. Consequently, the handling of titanium complexes normally requires oxygen- and water-free conditions. *See* COORDINATION CHEMISTRY.

L. Kieth Woo

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; M. F. Lappert, *Comprehensive Organometallic Chemistry II: Scandium, Yttrium, Lanthanides and Actinides, and Titanium, Zirconium, and Hafnium*, vol. 4, 1995; G. Lütjering and J. C. Williams, *Titanium: Engineering Materials and Processes*, 2003.

Titanium metallurgy

The winning of metallic titanium (Ti) from its ores followed by alloying and processing into forms and shapes that can be used for structural purposes.

History. In 1791 a British clergyman, William Gregor, published his observations and experiments with black sands found on Cornish beaches containing an unknown element. A few years later an Austrian chemist, M. H. Klaproth, identified this as an oxide of a new element which he named titanium.

The first metallic titanium was produced in the United States by M. A. Hunter at Rensselaer Polytechnic Institute and his associates at the General Electric Company. Titanium tetrachloride, $TiCl_4$, was reacted with sodium to produce metal.

In 1932 Wilhelm Kroll, a native of Luxembourg and the recognized father of the modern titanium industry, manufactured metallic titanium by combining titanium tetrachloride with calcium, and he made a few pieces of wire, sheet, and rod. By late 1940 Kroll had switched to reacting titanium tetrachloride with magnesium under an argon atmosphere, the basis for the first practical commercial process for producing titanium metal. By the mid-1950s a number of Japanese and American companies were producing many thousands of tons of metal using the Kroll method, a practice which still dominates the industry.

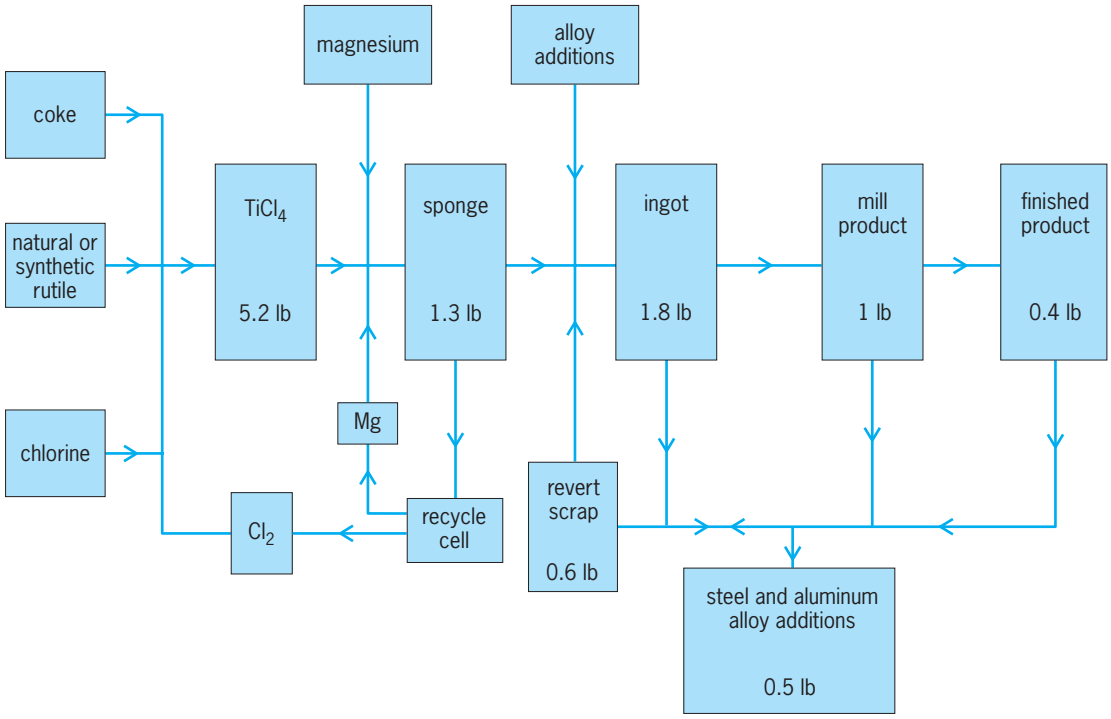


Fig. 1. Material cycle for titanium mill product manufacture. 1 lb = 0.45 kg.

In response to the unique material requirements of light weight and temperature and corrosion resistance for the evolving gas turbine engine, the titanium metal industry emerged in 1950. Titanium's unique properties—density half that of steel, excellent strength retention to 1000°F (538°C), and atmospheric corrosion immunity superior to that of other metals—made it an ideal construction material for both the engines and airframes of the newly developing jet airplanes. *See* AIRCRAFT ENGINE; AIRFRAME.

In the 1960s, discovery of titanium's excellent corrosive performance opened up a host of new applications in the chemical process industry. The first successful examples of industrial applications of titanium included reboilers in nitric acid concentrators, wet chlorine gas coolers in chlor-alkali plants, and chlorine dioxide bleach equipment in pulp/paper plants.

Natural occurrence. A stimulus for the initial interest in metallic titanium was the fact that it is the fourth most abundant structural element on the Earth's crust, exceeded only by aluminum, iron, and magnesium. Titanium deposits are widely scattered throughout the Earth's surface. Two forms dominate: rutile, essentially pure TiO_2 (95%), which usually occurs as black particles in beach sands; and the more abundant ilmenite, a titaniferrous ore, $FeTiO_3$ (50–65% Ti-bearing), which occurs in both alluvial and volcanic formations. *See* ILMENITE.

Of all the titanium minerals mined, only 3–5% are used to produce metal. The remainder is processed to titanium oxide (TiO_2) for use in the pigment industry, which utilizes either a sulfate or chlorination process to recover and purify it. Because of environmental concerns, the chlorination process is the preferred technology. This process requires a feedstock containing greater than 85% TiO_2, thus precluding the direct use of ilmenite. Because of the dwindling supplies of rutile, numerous plants are under construction for processes that can upgrade the more abundant ilmenite ore to greater than 90% TiO_2.

Winning. All commercial titanium metal is produced from titanium tetrachloride ($TiCl_4$), an intermediate compound produced during the chlorination process for titantium oxide pigment. The process (**Fig. 1**) involves chlorination of ore concentrates; reacting TiO_2 with chlorine gas (Cl_2) and coke (carbon; C) in a fluidized-bed reactor forms impure titanium tetrachloride as in reaction (1). For the production of acceptable metal, purification of the raw tetrachloride is required to remove other metal chlorides that would contaminate the virgin titanium. These critical purification steps involve distillation and precipitation of chlorides of vanadium, iron, zirconium, silicon, and magnesium, all of which occur with titanium in the ore.

$$TiO_2 + 2C + 2Cl_2 \rightarrow TiCl_4 + 2CO \quad (1)$$

The purified titanium tetrachloride is delivered as a liquid to the reactor vessel. In these vessels, constructed of carbon or stainless steel, the titanium tetrachloride is reacted with either magnesium (Mg) or sodium (Na), as in reactions (2) and (3), to form

$$TiCl_4 + 2Mg \rightarrow Ti + 2MgCl_2 \quad (2)$$

$$TiCl_4 + 4Na \rightarrow Ti + 4NaCl \quad (3)$$

the pure metal called sponge, because of its porous cellular form. To avoid contamination by oxygen or nitrogen, the reaction is carried out in an argon atmosphere.

The magnesium chloride ($MgCl_2$) or sodium chloride (NaCl) can be recycled to obtain both the metal fraction, magnesium [reaction (4)] or sodium [reaction (5)], and chlorine. The recycling process is a

$$MgCl_2 \rightarrow Mg + Cl_2 \quad (4)$$

$$2NaCl \rightarrow 2Na + Cl_2 \quad (5)$$

conventional production method.

The sponge, removed from the reactor pot by boring, is cleaned by acid leaching. In Russia and Japan, these excess reactants are removed by vacuum distillation.

Consolidation. One of the earliest challenges in the production of ductile titanium metal and its alloys was the development of an economical and technically acceptable method to consolidate the titanium sponge without contamination and embrittlement from atmospheric oxygen and nitrogen. Further complicating the problem was the highly reactive nature of molten titanium metal and its propensity for dissolving the mold materials into which it is cast.

As a result of these considerations, the consumable-electrode arc furnace eventually evolved. A mass of sponge, alloy additions, and scrap are mixed, then compressed into compacts and welded together to form a sponge electrode. This is melted by an electric arc into a water-cooled copper crucible in a vacuum or an atmosphere of purified argon. The arc progressively consumes the sponge electrode to form an ingot. No reaction occurs between the cool copper wall and the molten titanium; the vacuum or inert atmosphere prevents contamination of the molten metal. To attain commercially acceptable uniformity, the first ingot is remelted at least once and sometimes twice in a similar consumable-electrode furnace. Ingots up to 30,000 lb (13,600 kg) are routinely produced by using this consolidation method.

The cold-hearth melting process consists of feeding raw materials into a water-cooled crucible shaped like a bathtub and completely enclosed within a vacuum or inert-gas environment. Either electron beams or plasma gas (argon, helium, or a mixture of both) is used as the heat source. The molten metal flows over the lip into a mold of the appropriate shape. The process offers the advantage of casting rectangular shapes that can be processed easily on plate mills (rolling mills for producing flat products). Also, the process can be designed to remove, by density separation, harmful high-density inclusions.

Forming. The conversion of the titanium ingot into mill products, such as forging billet, plate, sheet, and tubing, is accomplished for the most part on conventional metalworking equipment. Mills designed to roll and shape stainless or alloy steel are used with only slight modifications. For this reason titanium and its structural alloys are produced in most of the same forms and shapes as stainless steel. *See* METAL FORMING; STAINLESS STEEL.

Shaping. Fabricating titanium mill products into finished parts is performed on conventional metalworking equipment with only a few exceptions. During any heating operation it is necessary to minimize the contaminating and embrittling effects of oxygen, nitrogen, and hydrogen. Close control of furnace temperatures and environments for heating prior to forging, forming, or heat treating are critical. During welding operations, the molten and hot metal must be protected from the atmosphere; otherwise a brittle weld will result. However, shielding techniques using argon or helium gas are routinely practiced. *See* EMBRITTLEMENT.

Basic metallurgy. Titanium is a relatively light, silvery-gray metal with a specific gravity of 0.163 lb/in.^{3} (4.51 g/cm^3). Pure titanium has a high melting point, 3035°F (1668°C). Titanium has a lower coefficient of expansion and lower thermal conductivity than either steel or aluminum alloys, and is not magnetic. Its modulus of elasticity, a measure of stiffness, is 1.6×10^7 lb/in.^{2} (1.1×10^{11} pascals), midway between that of steel and aluminum.

Titanium is allotropic. Up to 1625°F (774°C), titanium atoms arrange themselves in a hexagonal close-packed crystal array known as alpha (**Fig. 2***a*). When titanium is heated above the transition temperature (beta transus) of 1625°F (774°C), the atoms rearrange themselves into a body-centered cubic structure known as beta (Fig. 2*b*). The addition of other metals to a titanium base will favor one or the other of the two crystallographic forms. Some common titanium alloy additions are as follows:

Alpha stabilizers	*Beta stabilizers*	*Neutral*
Aluminum	Vanadium	Zirconium
Oxygen	Tantalum	Tin
Nitrogen	Molybdenum	
Carbon	Chromium	
	Iron	
	Nickel	

Whether a particular element favors the alpha or beta phase will raise or lower the beta transus temperature. Aluminum, for example, favors (stabilizes) the

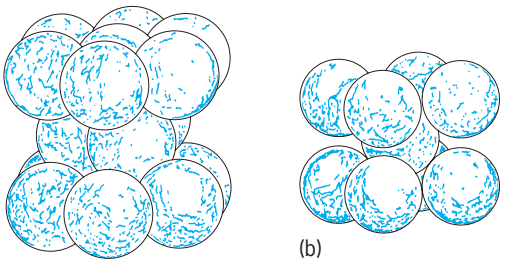


Fig. 2. Crystallographic forms of titanium. (*a*) Hexagonal close-packed alpha phase. (*b*) Body-centered cubic beta phase.

TABLE 1. Titanium corrosion resistance in corrosive environments*

Environment	Concentration, wt %	Temperature, °F (°C)	Corrosion rate, mil/yr† (mm/yr)
Acetic	All	212 (100)	<0.05 (0.001)
Aniline hydrochloride	5–20	122–212 (50–100)	<0.05 (0.001)
Benzene + HCl, no Cl	Vapor + liquid	349 (176)	0.2 (0.005)
Carbon tetrachloride	100	Boiling	<5.0 (0.1)
Chromic acid	25	212 (100)	<0.05 (0.001)
Ethanol	95	Boiling	0.05 (0.01)
Formaldehyde	37	Boiling	<5 (0.1)
Formic acid	10	212 (100)	<0.05 (0.001)
Hydrochloric acid	5	212 (100)	>250 (6.4)
Nitric acid	15	212 (100)	<0.05 (0.001)
	40	212 (100)	>25 (0.64)
Oxalic acid	25	212 (100)	<1000 (25)
Phosphoric acid	10	212 (100)	450 (11)
Sulfuric	5	Boiling	>2000 (50)
Terephthalic acid	75	Boiling	<0.05 (0.001)

*After S. R. Seagle and D. E. Thomas, Status of titanium technology, *Chem. Eng. Prog*., p. 63, June 1986.
†1 mil = 0.001 in.

alpha structure, raising the temperature at which the alpha transforms to beta. Iron favors the beta phase; therefore, iron depresses the beta transus. *See* CRYSTAL STRUCTURE.

Pure titanium is soft, weak, and extremely ductile. However, through appropriate additions of other elements, the titanium metal base is converted to an engineering material having unique characteristics, including high strength and stiffness, corrosion resistance, and usable ductility. The type and quantity of alloy addition determine the mechanical and, to some extent, the physical properties.

Alloys. Titanium alloys are classified into three groups depending on the phases present: alpha, beta, or a combination of the two, alpha-beta alloys. *See* ALLOY.

Alpha alloys. The hexagonal-structure compositions generally possess the highest strength at elevated temperatures (600–1300°F or 316–705°C) and the best weldability. However, in general, these alloys have the lowest room-temperature strength and are not heat-treatable.

Commercially pure titanium alloys, containing minor amounts of oxygen and nitrogen as strengtheners, are alpha alloys, and they find wide use in applications where maximum corrosion resistance and ease of fabrication are necessary. A higher-strength alpha alloy containing 5% aluminum and 2.5% tin is used in aircraft applications.

Alpha-beta alloys. As a group, alpha-beta alloys have higher strength and respond to heat treatment, but are less formable than alpha alloys. This class of titanium alloy accounts for more than half of all titanium metal products used.

The alpha-beta alloys vary widely in their composition, and therefore in their general characteristics, strength, heat treatability, and ductility. By far, the most common alpha-beta alloy is 6% aluminum plus 4% vanadium, the basic titanium alloy for jet engines and airframes.

Through heat treatment, the tensile strength of the 6% aluminum plus 4% vanadium alloy can be varied from 120,000 to 180,000 lb/in.^{2} (827 to 1240 megapascals).

Beta alloys. Titanium can be made to exist entirely in the beta phase at room temperature by adding alloys which inhibit the beta-to-alpha transformation. The alloys are extremely ductile at room temperature and can be heat-treated easily to strengths exceeding 200,000 lb/in.^{2} (1380 MPa). *See* HEAT TREATMENT (METALLURGY).

Corrosion resistance. Titanium's corrosion resistance is based on its highly reactive metal surface. This protective and tenacious oxide film provides excellent corrosion resistance in chloride brines, especially those of a neutral or oxidizing character. In addition, titanium demonstrates good corrosion resistance in most organic media. **Table 1** shows this corrosion resistance in several industrially important environments.

Unfortunately, titanium possesses poor resistance to reducing mineral acids such as sulfuric or hydrochloric acid. However, impurities common to many commercial environments act as powerful inhibitors. In reducing acids, impurities such as iron, copper, chromium, cobalt, and nickel behave as oxidizers, altering the oxidation-reduction characteristics. The environment becomes oxidizing, protecting the oxide surface. This has resulted in the use of titanium in several industries such as metal finishing and hydrometallurgy. Hydrometallurgical applications are particularly attractive, since copper and nickel ores are routinely digested by using hot sulfuric acid in titanium autoclaves. *See* HYDROMETALLURGY.

Additions of noble metal alloys as low as 500 parts per million improve the corrosion resistance of titanium. Platinum and palladium additions lower corrosion rates in reducing acids by several orders of magnitude. This is incorporated in a popular alloy produced as ASTM (American Society for Testing and Materials) Grade 7 and includes 1500 ppm palladium.

A coating of a platinum-group metal oxide attached to a titanium substrate results in a

TABLE 2. Typical titanium applications for various markets

Market	Application
Gas turbine engines	Compressor blades, disks, ducts, cases
Airframes	Landing gear beams, wing structure, hydraulic tubing
General chemical	Heat exchangers, condensers, mixers, piping
Organic/petrochemical	Strippers, reboilers, condensers, reactors
Power plants	Surface condensers, turbine blades
Electrolysis	Anodes for chlorine, chlorate, manganese dioxide; cathodes for copper, manganese; cathodic protection of bridges
Pulp/paper	Bleach tanks, wet chlorine systems, drum washers
Water technology	Flash desalination, heat exchangers for desalination
Metal recovery	Plating of chromium, nickel, silver, gold, zinc and galvanizing; hydrometallurgy of copper
Energy extraction	Logging tools, seals, springs, tubulars for sour gas and geothermal power
Medical	Prosthetic implants, instruments
Environmental	Flue gas desulfurization, wet air oxidation of waste, incinerator stacks, nuclear waste
Marine	Heat exchangers, piping systems, ball valves, sonar masts
High-performance vehicles	Racing valves, springs, retainers, connecting rods

dimensionally stable anode with low chlorine overvoltage. Compared to mercury or graphite, this anode represents a significant improvement in the chlor-alkali industry. The success of the dimensionally stable anode was possible only because of the titanium substrate's exceptional corrosion resistance even under extremely oxidizing chloride conditions. An extension of this technology has been applied to concrete structures to prevent the steel reinforcement bars from corroding.

Titanium has low thermal conductivity but performs well in heat-transfer applications. Several titanium characteristics are combined to produce high-performance, competitive heat exchangers; they include zero corrosion allowance, high strength with low density, near immunity to process upsets, unlimited flow rates, and no corrosion film. The use of high-ductility ASTM Grades 1 or 2 in heat exchangers can in many cases produce a unit one-tenth the size of comparable graphite or copper-nickel units. *See* CORROSION; HEAT EXCHANGER.

Titanium industry. Titanium sponge plants are operating in five countries: Russia, the United States, Japan, the United Kingdom, and the People's Republic of China. Melting and processing capabilities are present in most countries having an aircraft industry.

The United States titanium market has been growing at an annual rate of 4% since 1972, and represents approximately 50% of the total world market excluding Russia. The capacity for producing titanium in Russia is very large, about equal to the world capacity. The United States market is 80% aerospace and 20% nonaerospace. In the remaining non-Russian world, it is 50% for each of these applications.

Because of the decreasing price of titanium mill products and the confidence gained through evaluations conducted during previous years, increasing quantities of titanium mill products are being used outside the aerospace sectors. Primary areas of such applications are chemical processing equipment; dimensionally stable anodes for the production of chlorine; and tubing for electrical-power-generating-plant surface condensers and heat exchangers used in oil refineries, desalination plants, and pollution control devices. Typical applications for titanium markets are shown in **Table 2**.

A significant quantity of titanium metal is used as an alloy addition in steel, aluminum, and nickel-base alloys as a strengthener, a grain refiner, and an improver of corrosion resistance. These applications have become an important adjunct of the titanium metal-producing industry. *See* METALLURGY; TITANIUM.

Ward W. Minkler; Stanley R. Seagle

Bibliography. P. Crowson (ed.), *Minerals Handbook: 1994–95*, 1994; M. J. Donachie, Jr., *Titanium: A Technical Guide*, ASM International, 1988; D. Eylon, *Titanium for Energy and Industrial Applications*, Metallurgical Society AIME, 1987; F. H. Froes and I. Caplan, *Tritium Science and Technology*,1992; Institute of Metals, London, *Designing with Titanium*, 1986; R. W. Schutz, *Process Industries Corrosion*, 1986; S. R. Seagle and D. E. Thomas, Status of titanium technology, *Chem. Eng. Prog.*, pp. 63–68, June 1986; Titanium Development Ass., *Titanium 1986: Products and Applications*, 1986; U.S. National Materials Advisory Board, *Titanium: Past, Present and Future*, NMAB-392, 1983.

Titanium oxides

Chemical compounds of the metal titanium and oxygen. The most commonly found and used titanium oxides are the titanium dioxides, TiO_2; but other oxides are known including the sesquioxide Ti_2O_3, the monoxide TiO, and nonstoichiometric phases TiO_x, with x taking values between 0.7 and 1.3. Titanium dioxide exists in three common crystalline forms under ambient conditions: rutile, anatase (also known as octahedrite), and brookite (see **table**). Each polymorph contains titanium atoms surrounded by a distorted octahedron of oxygen atoms, and each form differs in the way in which the octahedral units are linked by various combinations of edge and corner sharing to give extended network structures (see **illustration**). Rutile is considered to be the most stable form of TiO_2, since at

Some physical properties of TiO_2 polymorphs

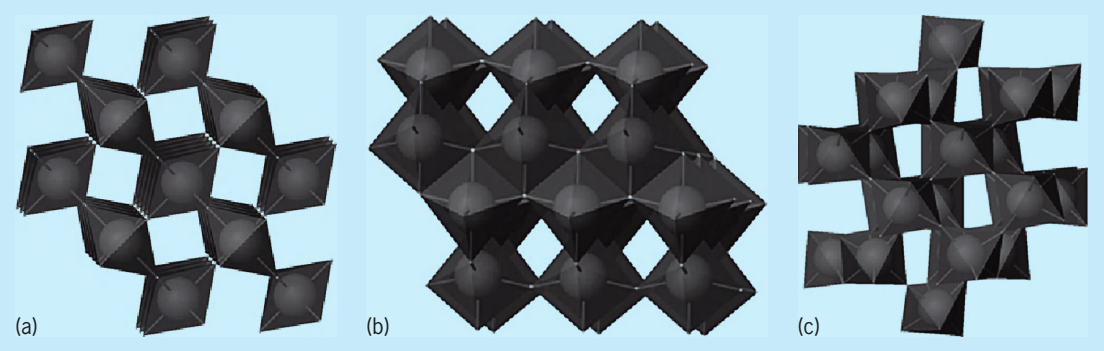
	Rutile	Anatase	Brookite
Crystal system	Tetragonal	Tetragonal	Orthorhombic
Space group	$P4_2/mmm$	$I4_1/amd$	$Pbca$
Unit cell	$a = b = 4.5845$ Å $c = 2.9533$ Å	$a = b = 3.7842$ Å $c = 9.5146$ Å	$a = 9.184$ Å $b = 5.447$ Å $c = 5.145$ Å
Density	4.2743 g cm^{-3}	3.895 g cm^{-3}	4.123 g cm^{-3}
Refractive index	2.60–2.90	2.49–2.55	2.58–2.70

temperatures above 500°C (930°F) both anatase and brookite are converted into rutile. *See* OXIDE; RUTILE; TITANIUM.

TiO_2 production. Both rutile and anatase are found in nature as minerals. There are also other more complex titanium-containing oxides, such as ilmenite ($FeTiO_3$) and leucoxene ($TiO_2 \cdot xFeO \cdot yH_2O$), from which the titanium dioxides may be extracted. These titanium ores are widespread throughout the world but are most readily exploited in Australia, the United States, India, and South Africa. Most of the titanium dioxide for commercial applications is extracted from titanium ores using one of two processes. In the chloride process, titanium ore is chlorinated using chorine gas and charcoal to yield volatile $TiCl_4$, which is separated by distillation from impurities and then oxidized by flame treatment to yield pure TiO_2. In the sulfate process, titanium ore is dissolved in sulfuric acid to produce a suspension of titanium oxyhydroxide from which TiO_2 is produced after filtration and firing. *See* ILMENITE.

Applications. The most widespread use of titanium dioxides is in the area of white pigments. Both anatase and rutile have exceptionally high refractive indices and therefore scatter light very effectively, offering highly effective opacity or hiding power as well as imparting whiteness and brightness to products. Rutile is most widely used since it has superior properties over anatase (a higher refractive index and a higher density, therefore offering greater opacity per gram), and has major pigmentary applications in coatings, paints, paper, inks, plastics, and rubbers, as well as in foods, cosmetics, toothpastes, and ultraviolet (UV) protection products. The particle size of TiO_2 pigments is an important property to control since the most effective light-scattering properties are exhibited by submicrometer crystals. Therefore, an important step in the manufacture of TiO_2 pigments is grinding or milling of powders to give the desired particle characteristics. TiO_2 pigments are not degraded during paint and plastic manufacture, giving them significant advantages over organic white pigments. Titanium dioxide pigments are considered to have almost no health hazards. They are nontoxic by ingestion, inhalation, or skin contact, and they are noncorrosive and noncombustible. For these reasons, TiO_2 white pigments are without commercial competition. Around 4 million metric tons of titanium dioxide pigment are produced worldwide annually. *See* PIGMENT (MATERIAL).

Photocatalysts. Titanium dioxides have been widely investigated as photocatalysts for a variety of applications, including catalyst materials for the production of energy by the decomposition of water into oxygen and hydrogen for solar energy conversion and the degradation of organic pollutants. These applications rely on the fact that TiO_2, both as anatase and rutile forms, is a semiconducting material with a band gap appropriate for the absorption of ultraviolet radiation by the excitation of valence electrons. This light-absorption process results in the formation of electron holes in the valence band and free electrons in the conduction band of TiO_2. Upon reaction with water, this excited form produces reactive hydroxyl and superoxide radicals that are able to initiate the destruction of organic pollutants into carbon dioxide and water. This process has been widely researched as a means of cleaning contaminated water from pollutants such as cyanide. One striking application of TiO_2 films is as coatings for self-cleaning



Representations of the crystal structures of the three polymorphs of titanium dioxide, TiO_2: (*a*) rutile, (*b*) anatase, and (*c*) brookite. The octahedral units represent titanium atoms surrounded by six oxygen atoms.

windows and tiles, where the destruction of organic deposits such as bacteria or odorous chemicals occurs upon the action of sunlight, and coupled with the hydrophilic nature of TiO_2 surfaces, washing is easily achieved by water alone. *See* BAND THEORY OF SOLIDS; ELECTRON-HOLE RECOMBINATION; FREE RADICAL.

Rechargeable batteries. One further use of more complex titanium oxide materials lies in rechargeable battery materials. Lithium/manganese/titanium rechargeable batteries are compact and were developed for watches and as backup power supplies for electronic devices such as pagers and timers. The batteries employ lithium–manganese complex oxide as the cathode material, and lithium–titanium oxide as the anode material. These batteries provide a capacity that is more than 10 times that of capacitors of the same size. *See* BATTERY; CAPACITANCE.

Richard Walton

Bibliography. J. H. Braun, Titanium dioxide: A review, *J. Coat. Technol.*, 69:59–72, May 1997; O. Carp, C. L. Huisman, and A. Reller, Photoinduced reactivity of titanium dioxide, *Prog. Solid State Chem.*, 32:33–177, 2004; A. Fujishima, T. N. Rao, and D. A. Tryk, Titanium dioxide photocatalysts, *J. Photochem. Photobiol. C: Photochem. Rev.*, 1:1–21, 2000; A. B. G. Lansdown and A. Taylor, Zinc and titanium oxides: Promising UV-absorbers but what effect do they have on intact skin?, *Int. J. Cosmetic Sci.*, 19:167–172, 1997.

Titration

A quantitative analytical process that is basically volumetric. However, in high-precision titrimetry the titrant solution is sometimes delivered from a weight buret, so that the volumetric aspect is indirect. Generally, a standard solution, that is, one containing a known concentration of substance X (titrant), is progressively added to a measured volume of a solution of a substance Y (titrand) that will react with the titrant. The addition is continued until the end point is reached. Ideally, this is the same as the equivalence point, at which an excess of neither X nor Y remains. If the stoichiometry or exact ratio in which X and Y react is known, it is possible to calculate the amount of Y in the unknown solution.

The normal requirements for the performance of a titration are: a standard titrant solution; calibrated volumetric apparatus, including burets, pipets, and volumetric flasks; and some means of detecting the end point.

Sometimes the standard titrant solution is prepared by the accurate weighing out of a primary standard, a substance that is stable and readily available in a high state of purity, such as potassium dichromate, anhydrous sodium carbonate, potassium hydrogen phthalate, or silver (which dissolves easily in nitric acid). The weighed material is dissolved, usually in water, and the solution is made up to a fixed volume in a volumetric flask of suitable capacity. The final solution must be homogeneous, a state normally achieved by ample shaking. If better-than-routine accuracy is required, the temperature must be known and constant.

Many valuable titrant solutions cannot be readily and accurately prepared by the direct method. For example, sodium hydroxide solution, much used for titrating acidic substances, is made from a solid that rapidly picks up moisture and so forth from the air. However, a solution of approximately the desired concentration can be prepared. This solution is then standardized by titrating (or being titrated with) another suitable standard solution. Sometimes a known weight of a suitable primary standard is placed directly in the titration vessel, dissolved in a suitable but unmeasured volume of solvent, and then titrated with the solution to be standardized. For example, sodium hydroxide solution is commonly standardized by titrating a known weight of potassium hydrogen phthalate. Analogous procedures are used for other titrant solutions that are made from substances that do not possess primary-standard properties.

Skillful operation may allow the titration to be stopped almost exactly at the end point. However, it is easy to overshoot, or add too much titrant. Provided that the total volume of titrant solution is noted, the titration can often be saved by the process of back titration. In the case mentioned above, a small overshoot of sodium hydroxide could be measured by back titration with standard hydrochloric acid solution. The process of back titration is sometimes used deliberately, especially for slow titration reactions. A known but excessive amount of titrant A is added to the substance to be determined. Then, after a suitable delay to allow for completion of the reaction, the remaining A is back-titrated with a reagent B that reacts rapidly with A. *See* CONCENTRATION SCALES; STOICHIOMETRY.

Classification by Chemical Reaction

For the purposes of titrimetry, chemical reactions can be placed in three general categories: acid-base or neutralization, combination, and oxidation-reduction.

Acid-base reactions. These titrations involve neutralization of an acid by titration with a base, or vice versa. However, the process is often nonspecific; in the titration of a mixture of nitric and hydrochloric acids, only the total acidity can be found without recourse to additional measurements. This arises because the only real reaction involved in these aqueous systems is the formation of water, reaction (1).

$$H_3O^+ + OH^- \rightarrow 2H_2O \qquad (1)$$

A salt derived from a strong base and a very weak acid can often be titrated just as if it were a base. For example, solutions of hydrochloric acid, or of other strong acids, are often standardized by titrating known weights of primary standard sodium carbonate. Sodium hydroxide solutions are often standardized by the use of primary standard potassium

hydrogen phthalate, which is a so-called acid salt. *See* ACID AND BASE.

Combination reactions. In titrimetry, attention is usually focused upon the combination of an ion in the titrant with one of the opposite sign in the titrand solution. Sometimes the combination may involve more than two species, some of which may be nonionic. The combination may result in precipitation. A classic example is the determination of chloride ion by titration with silver nitrate solution, when silver chloride is precipitated. Although a reaction may be known to yield a precipitate that has insignificant solubility and constant composition, this does not guarantee that the reaction will be useful titrimetrically. Any practical titration also requires that the reaction involved must not be unduly slow.

The same limitation also applies to complex-formation titrations, where precipitation may be absent or merely incidental. Except in a few special cases, complex-formation titrations were of little importance until the discovery of ethylenediaminetetraacetic acid (EDTA) and related compounds. These titrants not only are powerful complexing agents that combine with very many cations, but also form a single type of complex. This is in marked contrast to complexing agents such as ammonia or cyanide ion, which may yield a mixture of complexes.

With ubiquity comes lack of specificity. This difficulty with the EDTA family of titrants can sometimes be overcome by pH control and by technique masking (competitive complexation, whereby some species are prevented from taking part in the titration reaction). Then it may be possible to determine separately some or all of several metal ions in a mixture. *See* COORDINATION COMPLEXES; PRECIPITATION (CHEMISTRY).

Oxidation-reduction reactions. In so-called redox titrations the titrant is usually an oxidizing agent, and is used to determine a substance that can be oxidized and hence can act as a reducing agent. Because titrants that have usefully strong reducing properties are themselves attacked by oxygen in the air, the reverse procedure, although possible, is less usual. *See* OXIDATION-REDUCTION.

Coulometric Titration

Faraday's laws of electrolysis indicate that the extent of an electrochemical reaction is proportional to the total amount of electricity that is passed through the system. The passage of a uniform current for a measured period of time can be used to generate a known amount of a product such as a titrant. This fact is the basis of the technique known as coulometric titration. An obvious requirement is that generation shall proceed with a fixed, preferably 100%, current efficiency. The uniform current is then analogous to the concentration of an ordinary titrant solution, while the total time of passage is analogous to the volume of such a solution that would be needed to reach the end point.

Coulometric titration has several attractive features. Unless the sample to be titrated must be measured by volume, no volumetric glassware is needed. Nor are standard solutions; the titrant is generated during the process. This means that titrants of low stability, such as the strong reducing agent chromium(II) ion, can be employed routinely. The titration is started or stopped by the mere closing or opening of a switch, which can be remote from the titration vessel. This feature is obviously useful when titrating highly radioactive materials. The process can be controlled by an operator who is safely screened from the materials in the titration system. *See* ELECTROLYSIS.

Classification by End-Point Techniques

The precision and accuracy with which the end point can be detected is a vital factor in all titrations. Because of its simplicity and versatility, chemical indication is quite common, especially in acid-base titrimetry. That certain natural pigments such as litmus in an acidic solution change color when the solution is made basic has been known for centuries. Thus an acid solution that contains a small amount of litmus can be titrated with sodium hydroxide solution until the initial red color has just changed to blue. Litmus and other natural pigments, which are usually mixtures of various compounds, have been supplanted by synthetic indicators. These not only have sharper color transitions, but also can be made to suit particular applications.

Indicators. An acid-base indicator is a weak acid or a weak base that changes color when it is transformed from the molecular to the ionized form, or vice versa. The color change is normally intense, so that only a low concentration of indicator is needed. Phenolphthalein is an example of an indicator that acts as a weak monobasic acid according to reaction (2).

$$\underset{\text{Acidic form (colorless)}}{\mathrm{HIn}} \rightleftharpoons \underset{\text{Basic form (red)}}{\mathrm{H^+ + In^-}} \tag{2}$$

Methyl orange acts as a weak base according to reaction (3).

$$\underset{\text{Basic form (yellow)}}{\mathrm{InNR_2}} + \mathrm{H^+} \rightleftharpoons \underset{\text{Acidic form (red)}}{\mathrm{InNR_2H^+}} \tag{3}$$

The working range, or visual color change, of a typical acid-base indicator is spread over a hundredfold (~2 pH units) change in hydrogen ion concentration. Available indicators have individual working ranges that together cover the entire range of hydrogen ion concentration (10^{-1} to $10^{-13}M$, or pH 1 to 13) likely to be encountered in general acid-base titration. For example, the working ranges of phenolphthalein and of methyl orange are pH 8.0–9.8 and pH 3.0–4.4, respectively.

The success of a titration may hinge upon a suitable choice of indicator. **Figure 1** shows the titration curves of two monobasic acids (a strong acid, such as hydrochloric acid, and a moderately weak acid,

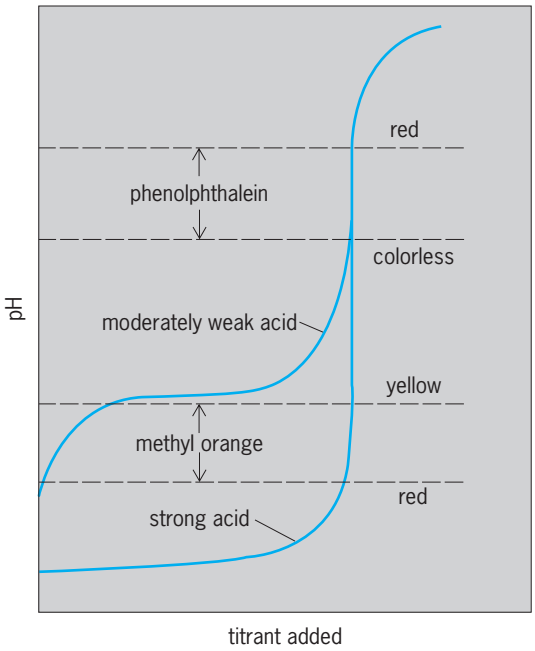


Fig. 1. Titration curves of two monobasic acids (0.1 *M*), with sodium hydroxide as the titrant.

such as acetic acid), each at a concentration of approximately 0.1 *M*, with the strong base sodium hydroxide as titrant. The respective working ranges of methyl orange and of phenolphthalein are also indicated. The curve for the strong acid is almost vertical in the approximate region pH 4–10. This means that a very small addition of titrant when the pH is approximately 4 causes methyl orange to change from orange to yellow. If phenolphthalein is used in place of methyl orange, the colorless indicator begins its change to the red form when the pH is approximately 8. Theoretically the end point with phenolphthalein requires a little more titrant than that required to reach the methyl orange end point. However, under routine conditions, the difference may be less than calibration-plus-operator error.

In the titration of the weak acid with methyl orange as the indicator, the pH rises to the lower limit of the working range before much titrant has been added. The curve rises slowly within this working range, so that the color change red through orange to yellow is both gradual and complete before all of the acid has been titrated. However, the weak acid curve eventually merges into the strong acid curve. A titration indicated by phenolphthalein therefore yields a sharp end point.

Similar reasoning applies to the titration of a base with an acid solution. If both species are strong, any one of a number of common indicators suffices. The successful titration of a weak base such as ammonia with a strong acid requires an indicator like methyl orange, which has a working range that is low on the pH scale. In the titration of a weak acid with a weak base, or the reverse, no simple indicator is suitable because the titration curve is not steep anywhere near the expected end point. The standardization of a solution of ammonia with standard acetic acid solution would be done indirectly. *See* ACID-BASE INDICATOR; HYDROGEN ION; PH.

In the complexometric titration of a cation, M^{n+}, such as that of magnesium, with EDTA or a similar agent, the concentration of M^{n+} falls most rapidly in the immediate region of the end point. The response resembles that shown by the curve of a strong acid in Fig. 1, but pM^{n+} replaces pH. The indicator is itself a suitable complexing agent that changes color when it is combined with M^{n+}. This is the situation during most of the titration. At or very near the expected end point, the titrant reacts with the small amount of cation in the indicator complex. The color then changes to that of the free agent.

When a solution of an oxidizing agent is used to titrate a substance that is readily oxidized, the electrochemical potential rises most rapidly in the region of the end point. The titration curve is generally similar to that shown for the weak acid in Fig. 1, but potential replaces pH. The chemical indicator is here a substance that undergoes a marked change in color when a suitable potential is reached. Unlike the indicators commonly used in acid-base or complex-formation titrations, oxidation-reduction indicators are often irreversible. Thus, if an excess of oxidizing titrant has been added, the indicator color may not revert to the original if back titration is subsequently attempted.

The important oxidizing titrant potassium permanganate forms an intensely purple solution. Its reduction products are essentially colorless. This titrant is thus self-indicating, provided that the titrand and its oxidation products have little or no color. The end point is taken as the first appearance of a pale but permanent pink color.

Sometimes no suitable chemical indicator can be found for a desired titration. Possibly the concentrations involved may be so low that chemical indication functions poorly. Other situations might be the need for high precision or for the automatic arrest of the titration. Recourse is then made to some physical method of end-point detection.

Potentiometric titration. If a pH meter is used, acid-base titration curves like those shown in Fig. 1 can be plotted or recorded. The meter and its associated electrodes are first standardized by use of a buffer solution of known pH. The electrodes are then immersed in the well-stirred solution to be titrated, and the titration is begun. For routine purposes, interest is in rapid and reasonably close end-point location. Titration is first carried out quite quickly until the meter shows signs of rapid response, and then is slowed so that it can be stopped at the pH jump or fall that marks the end point. Obviously, potentiometric titration can be used for a highly colored titrand solution, in which the response of a color-change indicator could not be seen.

There are two approaches to higher precision. In the first the pH at which the desired end point occurs is determined, and the titration of the actual sample is then arrested exactly at, or very close to, this pH value. The other approach is to stop at the point of steepest slope of the titration curve. This can

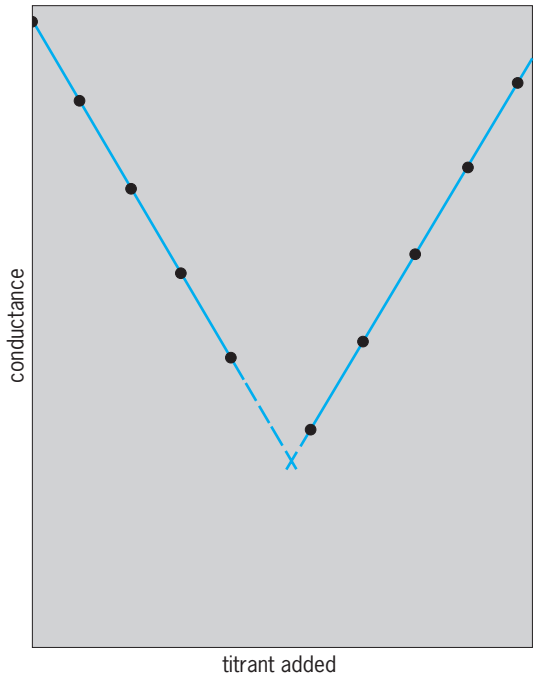


Fig. 2. Conductometric titration curves of a dilute hydrochloric acid solution with sodium hydroxide as the titrant.

be done by plotting (or, with suitable instrumentation, sensing) the first derivative ($\Delta pH/\Delta V$) or the second derivative ($\Delta^2 pH/\Delta V^2$) against the volume V of titrant added. In theory this method is applicable to any potentiometric titration, without the need to predetermine the end-point conditions.

By suitable choice of electrodes, these potentiometric methods can also be applied to combination titrations and to oxidation-reduction titrations. The advent of modern ion-selective electrodes has greatly extended the scope of potentiometric titration and other branches of titrimetry. *See* ELECTRODE POTENTIAL.

Conductometric titration. Several instrumental techniques give rise to titration curves that are essentially linear. Generally, the end point is found graphically from readings taken at a number of points that fall on appropriate linear portions of the curve. One advantage is that poor response in the actual vicinity of the end point does not cause difficulties. Another is good precision, due to the averaging effect of the many individual readings.

The underlying principles of conductometric titration are that the solvent and any molecular species in solution exhibit only negligible conductance; that the conductance of a dilute solution rises as the concentration of ions is increased; and that at a given concentration the hydrogen ion and the hydroxyl ion are much better conductors than any of the other ions.

One example is that of a dilute hydrochloric acid solution titrated with one of sodium hydroxide. The initial conductance, high because the hydrogen ion concentration is high, falls as the titrant is added (**Fig. 2**). When neutralization is complete, the conductance is that of a solution of sodium chloride. The change in conductance is due to progressive replacement of hydrogen ion by sodium ion, which is a poorer conductor. Further addition of titrant causes the conductance to rise again, principally because the highly conducting hydroxyl ion now remains in the system, instead of undergoing mutual destruction with the hydrogen ion. Addition of titrant naturally dilutes the titrand solution, causing distortion of the linear branches of the titration curve. This diluting effect can be compensated for by a simple arithmetic correction or can be rendered negligible by use of a concentrated titrant solution.

Conductometric titration is sometimes successful when chemical indication fails. A typical case is the titration of acetic acid with ammonia (**Fig. 3**). Both reactants are weak electrolytes, but ammonium acetate, the salt produced in the titration, is a strong electrolyte. The rising linear branch is due to the buildup of the ions of this salt. In the region beyond the end point, the addition of more titrant causes very little change in conductance. The concentration of ammonium ion then present forces the added titrant to remain almost entirely in the molecular state. Because of the effects of hydrolysis, the graph is decidedly rounded in the region of the end point. Although an old technique, conductometric titration still finds extensive use in studies of nonaqueous systems.

Certain precipitations can be satisfactorily performed conductometrically. However, conductometry is generally of little use in oxidation-reduction titrations. In fact the technique is difficult to use if the system contains appreciable concentrations of electrolytes other than those involved in the actual

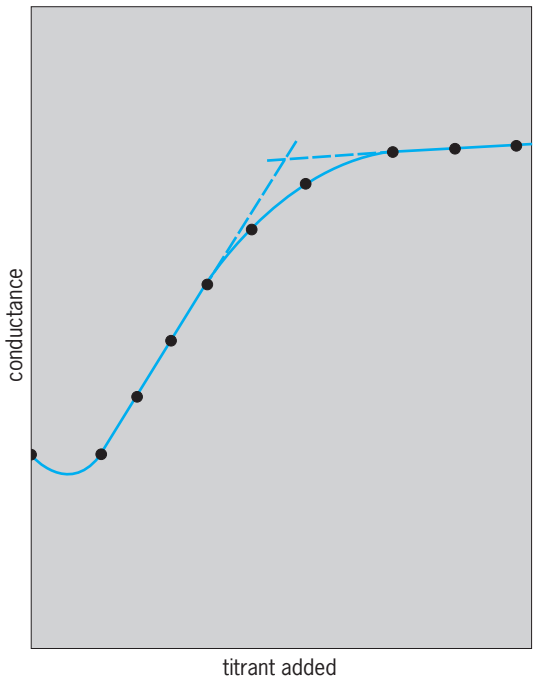


Fig. 3. Conductometric titration curve of acetic acid with ammonia as the titrant.

titration. The observed conductance is a function of the total ionic content, while the precision with which the end point can be determined depends upon the relative change in conductance. For example, the angle between the branches in Fig. 2 is less acute if the titrand solution contains sodium nitrate as well as hydrochloric acid. *See* ELECTROLYTIC CONDUCTANCE.

Conductometric titrations are normally carried out with the aid of audio-frequency alternating current. By use of radio frequencies, the titration can be monitored in a cell that has the electrodes on the outside, so that they are not in contact with the titrand solution. However, the titrand curves are usually less simple than those encountered in normal conductometric titration.

Spectrophotometric titration. The spectrophotometer is an optical device that responds only to radiation within a selected very narrow band of wavelengths in the visual, ultraviolet, or infrared regions of the spectrum. The response can be made both quantitatively and linearly related to the concentration of a species that absorbs radiation within this band. Titrations at wavelengths within the visual region are by far the most common. An example is the titration of iron(II) in dilute sulfuric acid solution with a standard solution of potassium permanganate. The spectrophotometer is adjusted to measure the absorbance of the highly colored titrant. Very little absorbance is observed until the end point is reached; iron(II) and the reaction products are essentially colorless and are not "seen" by the spectrophotometer. However, the absorbance rises as titration is continued beyond the end point because the titrant is no longer being destroyed. The titration curve is shown in **Fig. 4**.

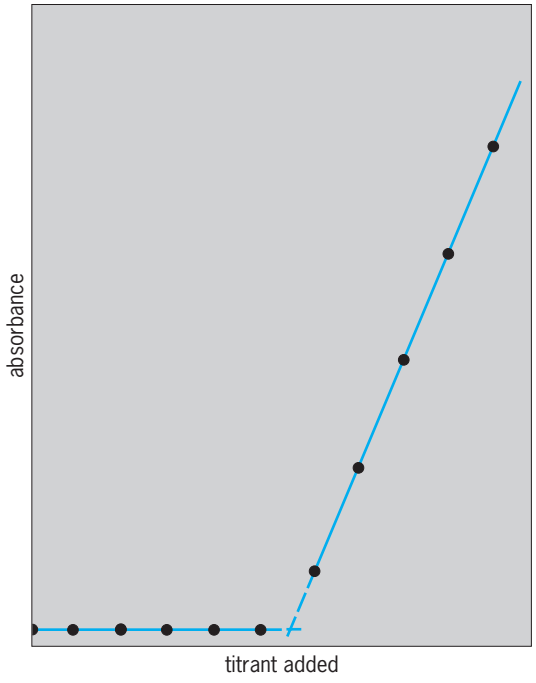


Fig. 4. Spectrophotometric titration curve of iron(II) in dilute sulfuric acid solution with potassium permanganate as the titrant. The curve shows the response when the titrant absorbs beyond the end point.

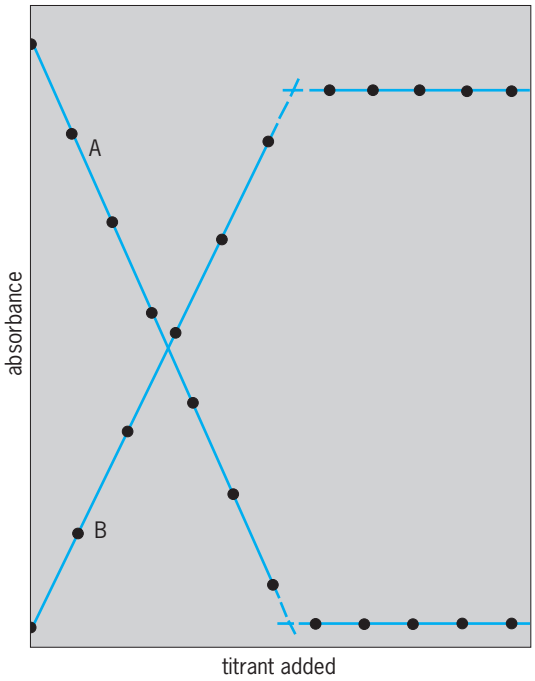


Fig. 5. Spectrophotometric titration curves: A, when the titrand absorbs but the products do not, and B, when neither of the reactants absorbs but the products do.

If the titrant does not absorb at the chosen wavelength, the curve is horizontal in the region beyond the end point. If the titrand absorbs but the products do not, the curve has the form of curve A in **Fig. 5**. Curve B shows the response when neither of the reactants absorb, but the products are strong absorbers.

In some cases, systems that involve no absorbing species can be handled by the addition of an optical indicator. This is a substance that undergoes no change in absorbance until the titrant is present in slight excess over that required for the main reaction.

Amperometric titration. By the use of a dropping-mercury or other suitable microelectrode, it is possible to find a region of applied electromotive force (emf) in which the current is proportional to the concentration of one or both of the reactants in a titration. **Figure 6** shows the separate current-voltage curves, at approximately similar concentrations, of two species X and Y that can react to form nonelectroactive products. For example, X and Y may form an essentially insoluble precipitate. Provided that conditions are suitable, the currents may be anodic or cathodic.

Suppose that the applied emf is fixed within the range *pq* and that Y is being titrated with X. The current remains at the small residual value until near the end point and rises as the titration is carried past the end point. The general shape of the titration curve is thus similar to that shown in Fig. 4. Titration at a fixed emf within the range *rs* gives a V-shaped curve, because both titrant and titrand are electroactive under these conditions.

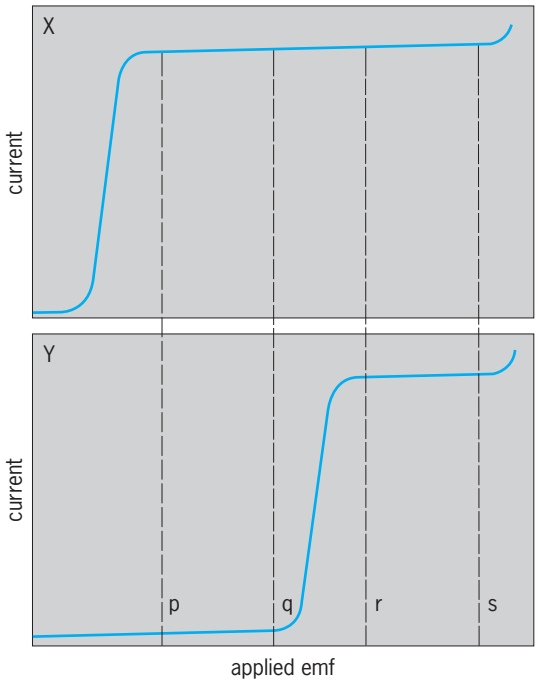


Fig. 6. Current-voltage curves at similar concentrations for two species (X and Y) that can react to form nonelectro active products.

Amperometric titration has been applied to all classes of reactions. Foreign ions may be present (and are often deliberately added to suppress migration currents), provided that they neither interfere with the titration reaction nor are electroactive under the chosen conditions.

Biamperometric titration is a closely related technique. An emf that is usually small is applied across two identical microelectrodes, usually of platinum, that dip into the titrand solution. This arrangement, which involves no liquid-liquid junctions, is obviously valuable in nonaqueous titrations, but also finds much use in aqueous titrimetry. The titration curves are nonlinear, so that direct observation, rather than graphing, is normally used to find the end point. *See* POLAROGRAPHIC ANALYSIS.

Thermometric or enthalpimetric titration. Many chemical reactions proceed with the evolution of heat. If one of these is used as the basis of a titration, the temperature first rises progressively and then remains unchanged as the titration is continued past the end point. If the reaction is endothermic, the temperature falls instead of rising. Thermometric titration is obviously applicable to all classes of reactions.

The total temperature change may be small, so that transfer of heat between the titration vessel and its surroundings must be minimized. Temperature changes are usually measured by means of a thermistor, which has both small heat capacity and a high negative coefficient of electrical resistance. The lower specific heats of solvents other than water indicate that the thermometric titration is particularly suited to nonaqueous titration. *See* THERMOCHEMISTRY.

Nonaqueous titration. Water, the cheapest solvent, may react with, or may not dissolve, the substances to be used in a titration. Recourse must then be made to nonaqueous titrimetry. However, the main use of this technique is to perform titrations that give poor or no end points in water. Although applicable in principle to all classes of reactions, acid-base applications have greatly exceeded all others.

When an acid AH is dissolved in an amphiprotic solvent SH, and equilibrium is established as shown in reaction (4), if SH is strongly basic (protophilic),

$$AH + SH \rightleftharpoons A^- + SH_2^+ \qquad (4)$$

the equilibrium is forced to the right. All acids, whether weak or strong in water, tend to behave as strong acids. This is termed the leveling effect. Compounds such as phenol are too weakly acidic to be titrated in water. However, the titration can be performed in a solvent such as ethylenediamine. A common titrant is tetrabutylammonium hydroxide in the same solvent.

Analogously, a strongly acidic (protogenic) solvent exerts a leveling effect on basic solutes. Pyridine and other bases that are very weak in water can be nicely titrated in anhydrous acetic acid. The usual titrant is perchloric acid in the same solvent. By suitable choice of solvent, the apparent strengths of individuals in a mixture of acids or bases can sometimes be spread out, so that a succession of separate end points is obtained.

The coefficient of thermal expansion of a typical nonaqueous solvent is several times greater than that of water. Titrations should therefore be carried out at, or very near to, the temperature at which the titrant was standardized. Nonaqueous titrations in which the solvent is a molten salt or salt mixture are also possible.

Automatic titration. Automation is particularly valuable in routine titrations, which are usually performed repeatedly. One approach is to record the titration curve and to interpret it later. This requires a buret of the syringe or pump type that has a constant rate of delivery and is coupled to the recorder to provide the volume-of-titrant axis. Another method is to stop titrant addition or generation automatically at, or very near to, the end point. Although a constant-delivery device is desirable, an ordinary buret with an electromagnetically controlled valve is often utilized.

The second method is common in automatic potentiometric titration. Overshoot must be prevented if precise and accurate results are needed in minimum time. Titrant addition, rapid during most of the process, is automatically slowed down when the potential is approaching the end-point value. Microcomputer control permits such refinements as the continuous adjustment of the titrant flow rate during the titration. In some cases it is possible to automate an entire analysis, from the measurement of the sample to the final washout of the titration vessel and the printout of the result of the analysis. *See* ANALYTICAL CHEMISTRY.

John T. Stock

Bibliography. G. D. Christian, *Analytical Chemistry*, 5th ed., 1993; G. W. Ewing, *Instrumental Methods of Chemical Analysis*, 5th ed., 1985; S. B. Khoo, *Analytical Chemistry*, 1990; E. Scholz, *Karl Fischer Titration*, 1984.

Tobacco

The plant genus *Nicotiana*, certain species in the genus, and dried leaves of these plants are all called tobacco. Most often tobacco means a leaf product containing 1–3% of the alkaloid nicotine, which produces a narcotic effect when smoked, chewed, or snuffed. The plant *N. rustica* provides tobacco in parts of Europe, but the tobacco of world commerce is *N. tabacum* (**Fig. 1**).

Tobacco is American in origin. Columbus found West Indians smoking it in a hollow forked stick. Historians do not know who first took tobacco to Europe, but most of them credit Jean Nicot in 1561. *Nicotiana* and nicotine incorporate his name. *See* SOLANALES.

Characteristics. This solanaceous annual, found only in cultivation, probably began as an amphidiploid or fertile hybrid between two species of *Nicotiana* native to Bolivia. It has 24 chromosome pairs, an erect, thick stem 2–9 ft (0.6–2.7 m) high, and alternate, sessile, oval, or lanceolate leaves. The flowers are produced in a panicle. They have a tubular corolla 3.5–5.5 cm long widening midway to a throat and ending in a five-pointed flare. There are five anthers which shed sticky pollen just as the flower opens. Most flowers are self-pollinated. The fruit is a two-section capsule with minute seeds (about 15,000 in a gram). *See* GENETICS; INFLORESCENCE.

Cultivation, harvesting, and curing. The local soil type and climate within a country, state, or even county determine the type of tobacco grown there.

Fig. 1. ***Nicotiana tabacum*, Connecticut cigar wrapper type. (*Connecticut Agriculture Experiment Station*)**

Generally all tobacco does best in a warm, even climate on light, well-drained, carefully fertilized soils that are clean cultivated and receive moisture weekly from rain or irrigation. Special seedbeds, either open or covered with cloth, glass, or plastic, provide transplants for fields. Plants are usually topped (blossom removed) to expand upper leaves. Harvest proceeds by cutting the whole stalk (stalk-cut) or picking leaves successively as they ripen (primed). Primed leaves are supported on strings, wires, or sticks.

Types of tobacco

Type	Major production areas	Use	Curing method
Flue-cured (bright)	United States (Virginia, North Carolina, South Carolina, Georgia-Florida border); Canada (Ontario); Rhodesia; Europe; China; Japan; Australia; India	Cigarettes	Primed leaves heat-cured to a bright yellow color
Light air-cured (Burley, Maryland)	United States (Kentucky, Maryland, Tennessee); Canada; Europe	Blending in cigarettes	Whole plant air-dried in ventilated sheds
Oriental (Turkish)	Greece (Macedonia); Turkey (Samsun, Smyrna)	Blending for aroma in cigarettes	Primed leaves cured in sun on strings
Cigar filler	Cuba*; Philippines; Puerto Rico; United States (Pennsylvania, Ohio); Dominican Republic; Central and South America	Central bulk of cigars	Whole plant air-dried in ventilated sheds
Cigar binder†	United States (Connecticut River Valley, Wisconsin, Pennsylvania)	Binding the filler into shape	Whole plant air-dried in ventilated sheds
Cigar wrapper	Under cloth shade: United States (Connecticut River Valley, Georgia-Florida border); Cuba; Italy; new in Canada and Central and South America Unshaded: Sumatra	Outer leaf or wrapper of cigar	Primed leaves heat- and air-dried to a uniform golden brown or a bright green (Candella)
Dark air-cured and fire-cured	United States (Kentucky, Tennessee, Virginia)	Pipe, chewing, snuff	Whole plant air-dried in ventilated sheds

*Embargoed in the United States since 1962.
† Acreage in Connecticut River Valley dropped 90% after 1956 after replacement by processed sheet reconstituted from pulverized tobacco.

Curing (drying) is done in ventilated barns with natural or artificial heat. In some areas, machine-harvested leaves are packed in special frames for curing in heat-regulated, forced-air chambers. Drying time ($^1/_2$–6 weeks) and temperature (70–170°F or 21–77°C) influence the amount and kind of changes that occur in proteins, carbohydrates, organic acids, alkaloids, and enzymes in the leaf. Before use in cigars, cigarettes, pipes, chewing tobacco, or snuff, cured leaves are fermented by storing them 6 weeks to 2 years at about 15% moisture and 80–110°F (27–43°C). The methods used for harvesting, curing, and fermenting depend on the type of tobacco, intended use, and local custom. A flat product made from ground-up tobacco plus binding agents is often blended with tobacco leaf.

Economic importance. Tobacco is economically important in 66 countries and is grown to some extend in all but a few countries. Principal tobacco types, major world production areas, uses, and curing methods are shown in the **table**.

Gordon S. Taylor

Diseases. Diseases cause production problems in all countries where tobacco is grown. Annual losses of tobacco products worldwide average 10% or more; in addition, production costs are invariably increased. About 25 principal pathogens attack the crop from seed sowing through marketing.

Root knot, caused by the nematode *Meloidogyne* sp., is the most important disease. It occurs throughout the warmer countries, particularly in sandy soils. Diseased plants are stunted and have wilted leaves that yellow prematurely. Knots or galls form on the roots (**Fig. 2**). Infected plants become more susceptible to other diseases such as black shank and bacterial wilt.

Lesion nematodes (*Pratylenchus* sp.) are also important root pathogens that kill the feeder roots, consequently depleting the root systems and resulting in stunted, weakened plants.

Black shank, caused by the soil-inhabiting fungus *Phytophthora parasitica*, also is a warm-weather disease. Tiny swimming spores (zoospores) formed by the fungus penetrate and kill the roots, the bases of the stems turn black, and the plants quickly die (**Fig. 3**). Water may carry the zoospores long distances to begin new outbreaks.

Black root rot, caused by *Thielaviopsis basicola*, a fungus that attacks many species of plants, is widespread in cooler soils. The fungus forms black lesions on the roots, which then decay.

Fig. 2. Tobacco plant severely infected with root knot. (*USDA*)

Fig. 3. Black shank–infected tobacco plant with stalk split lengthwise to show typical decay and disking of the pith. Most of the roots are already dead. (*North Carolina State University*)

Fig. 4. Influence of (*a*) high (86°F or 30°C) and (*b*) low (68°F or 20°C) soil temperature on recovery of plants in black root rot–infested soil. The roots at 30°C were like those at 20°C at the beginning of the experiment. (*USDA*)

Fig. 5. Bacterial wilt: young plant showing the one-sided wilting of green leaves characteristic of the disease. (*USDA*)

Severely infected plants have only a few black stubby roots attached to the base of the stem (**Fig. 4**).

Blue mold, caused by the downy mildew *Peronospora hyoscyami*, attacks the leaves and stems in both the seed bed and field during cool, wet weather. Small plants are killed in a few days. Tiny airborne spores, produced in great numbers, may be blown hundreds of kilometers to start new epidemics.

Pseudomonas solanacearum, which causes bacterial wilt, attacks wounded roots of plants grown in sandy soils of warm climates. Great numbers of the rapidly producing bacteria produce toxins and also interfere with water uptake and conduction. Leaves wilt (**Fig. 5**) and the plant dies.

Wildfire (angular leaf spot), caused by *Pseudomonas tabaci*, is the most widespread leaf disease, caused by a bacterium that enters the leaf through the stomates. The pathogen thrives in water-congested tissue and causes circular, yellow-green spots on the leaves. Often the spots fuse to form irregular dead areas that may destroy the entire leaf (**Fig. 6**).

Mosaic occurs worldwide and is the principal viral disease of tobacco. The virus is spread by contact. Rubbing a diseased leaf and then a healthy one will produce infection. Leaves of diseased plants are mottled with intermingled light- and dark-green areas (**Fig. 7**).

Cucumber mosaic, veinbanding (caused by potato virus Y), and etch viruses infect many wild and cultivated plants. They are spread by leaf-feeding aphids (*Myzus* spp.) that carry the virus on their stylets from infected to healthy plants. Symptoms caused by different viruses are often similar and vary with the cultivar, virus strain, and age of the plant when infected. Moreover, plants may be infected with more than one virus. Symptoms range from a mild mottling or mosaic, to leaf distortion, scalding, necrosis, and stunting of the plants. The tomato spotted wilt virus, widespread in Europe, deforms and stunts plants. It is transmitted by thrips. Control is difficult.

Fig. 6. Mature tobacco leaf showing wildfire symptoms. The yellow halo around the younger lesions is a typical symptom. The larger coalescing lesions often show no halo. (*USDA*)

Fig. 7. Mottled leaves of a young plant infected with tobacco mosaic virus. The distortion seen on the leaf at the right commonly occurs on plants infected when they have four to six leaves. (*North Carolina State University*)

Weather fleck is caused by air pollution. Ozone is the principal damaging agent. Numerous small brown spots form on affected leaves. Such leaves are worthless as cigar wrappers.

Drought, hail, drowning, frost, heat, excessive sunlight, herbicide injury, nutrient deficiencies, and toxicities, alone or in combination with disease, frequently decrease both yield and quality by stunting the plants, causing premature yellowing or death of leaf tissue, or upsetting the proper carbon/nitrogen ratio necessary for desirable smoke aroma and flavor.

No single disease control measure will control any disease satisfactorily. Efficient control requires a continuous integrated management program including: disease-resistant cultivars, healthy transplants; rotation with other crops; effective pesticides; destruction of plant debris by disking or plowing immediately after harvest; and weed and insect control. *See* PLANT PATHOLOGY.

G. B. Lucas

Bibliography. G. B. Lucas, *Diseases of Tobacco*, 1975; H. D. Shew and G. B. Lucas (eds.), *Compendium of Tobacco Diseases*, 1991.

Toggle

Any of a wide variety of mechanisms, many used to open or close electrical contacts abruptly and all characterized by the control of a large force by a small one. The basic action of a toggle mechanism is shown in **illus.** *a*. When $\alpha = 90^\circ$, forces P and Q are independent of each other. Again, when $\alpha = 0^\circ$ the forces are isolated, force Q being sustained entirely by the frame, and force P serving only to hold the link in position. At $\alpha = 45^\circ$ from the symmetry $|P| = |Q|$, the mechanism serves to transfer the direction of forces to achieve equilibrium. If frame, connecting rod, and slider blocks with their pivots are sufficiently strong to support the maximum force encountered and if the coefficients of friction at the pin and slider joints are negligible, at intermediate positions the forces are related by the equation below.

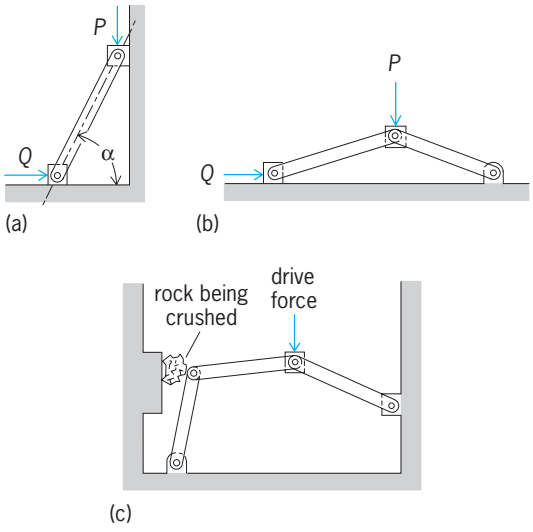
$$P = Q \tan \alpha$$

The essential characteristic of the mechanism arises from the tangent relation between forces. As $\alpha \to 0$, a small force at P can overcome a large force at Q. *See* COUPLING.

The mechanism finds application in such machines as printing presses, embossing machines, friction brakes and clutches, and rock crushers. In such devices, the frame sustains the output force, the drive supplying only a small force through a sufficient distance to bring the linkage into position. Arranged to drive slightly past dead center, the mechanism serves as a clamp that can be released by a small reverse force at P, yet will hold against a large force at Q. Such action finds application in workpiece clamps, tool holders, and electric switches and circuit breakers. A spring is generally added to assist the release once the mechanism has been returned through dead center.

Because the simple configuration of illus. *a* requires low-friction sliders, it is impractical. A more useful structure replaces the vertical slider with a second link pinned to the frame (illus. *b*), in which case input P sets up forces in both links. A further modification (illus. *c*) replaces the other slider with a link. *See* FOUR-BAR LINKAGE; LINKAGE (MECHANISM).

Frank H. Rockett



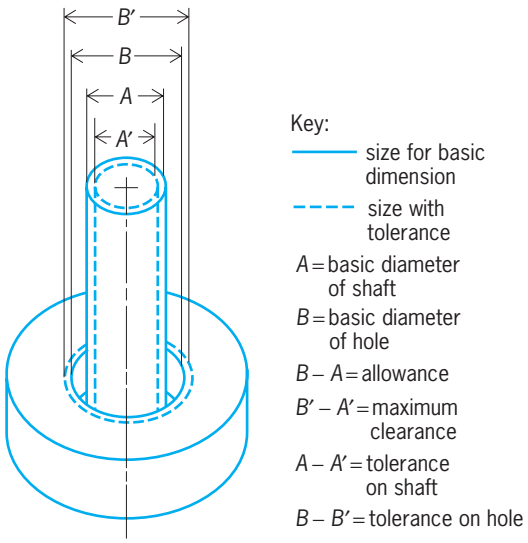
Toggle mechanism. (*a*) Simple structure. (*b*) Traditional configuration. (*c*) Typical application.

Tolerance

Amount of variation permitted or "tolerated" in the size of a machine part. Manufacturing variables make it impossible to produce a part of exact dimensions; hence the designer must be satisfied with manufactured parts that are between a maximum size and a minimum size. Tolerance is the difference between maximum and minimum limits of a basic dimension. For instance, in a shaft and hole fit, when the hole is a minimum size and the shaft is a maximum, the clearance will be the smallest, and when the hole is the maximum size and the shaft the minimum, the clearance will be the largest (see **illus.**).

If the initial dimension placed on the drawing represents the size of the part that would be used if it could be made exactly to size, then a consideration of the operating conditions of the pair of matting surfaces shows that a variation in one direction from the ideal would be more dangerous than a variation in the opposite direction. The dimensional tolerance should be in the less dangerous direction. This method of stating tolerances is called unilateral tolerance and has largely displaced bilateral tolerance, in which variations are given from a basic line in plus and minus values.

As an example, for a 1.5-in. shaft and hole for a free fit the standard allowance is 0.002 in. and the



Maximum and minimum limits of the basic dimensions of a machine part.

tolerance for hole and shaft is 0.001 in.:

$$\begin{aligned}\text{Max. shaft diameter} &= \text{nominal size} - \text{allowance}\\ &= 1.500 - 0.002\\ &= 1.498 \text{ in.}\end{aligned}$$

In the unilateral method for stating tolerance, the shaft diameter is $1.498^{+0.000}_{-0.001}$, or between 1.498 and 1.497 in. The diameter of the hole is $1.500^{+0.001}_{-0.000}$, or between 1.500 and 1.501 in. The maximum clearance is 1.501 minus 1.497, or 0.004 in., and the minimum clearance is 0.002 in. Paul H. Black

Bibliography. D. A. Madsen, *Geometric Dimensioning and Tolerancing*, rev. ed., 1988; E. Oberg and F. D. Jones, *Machinery's Handbook*, 23d ed., 1988.

Tomato

An important vegetable, belonging to the genus *Lycopersicon*, especially *L. esculentum*, that is grown for its edible fruit. *Lycopersicon* species are native to South America, especially Peru, and the Galápagos Islands. The tomato was first domesticated in Mexico, and received its name from the Aztec word *xitomate*.

The tomato was introduced to Europe in the mid-sixteenth century, and was prominently featured in the early herbals. It was grown for the beauty of its fruit but was not often eaten, except in Italy and Spain, because many considered it to be poisonous like its relative the deadly nightshade. The tomato has since become one of the most popular vegetables in the world.

Although native to the New World, the tomato was introduced to America from Europe. It has been grown in the United States since colonial days, but it became an important vegetable there only in the past century.

Classification and structure. The genus *Lycopersicon* is a member of the Solanaceae, the nightshade family. *Lycopersicon esculentum*, the familiar tomato, can be hybridized with each of the eight other species of *Lycopersicon*. It can be easily crossed with *L. pimpinellifolium*, the red currant tomato. Crosses with the more distantly related, green-fruited *Lycopersicon* species are more difficult; usually the tomato must be the maternal parent, embryo culture is sometimes required, and the hybrids have varying degrees of fertility. Despite these obstacles, tomato breeders have transferred many genes, particularly for disease resistance, from wild *Lycopersicon* species to the tomato. *See* BREEDING (PLANT).

Lycopersicon is allied to *Solanum*, a diverse genus that includes potato, eggplant, and over 2000 other species. Two *Solanum* species, *S. pennellii* and *S. lycopersicoides*, have been crossed with the tomato.

The tomato, and all other species with which it can be crossed, has 12 pairs of chromosomes. There is good homology between chromosomes of each species. Tomato chromosomes differ in length, centromere position, and heterochromatin distribution, and each of the chromosomes of the tomato can be distinguished cytologically.

The tomato is a herbaceous perennial, but is usually grown as an annual in temperate regions since it is killed by frost. Originally it had an indeterminate plant habit, continuously producing three nodes between each inflorescence. Many modern varieties, however, are determinate; they have the self-pruning (*sp*) gene, and therefore have fewer than three nodes between inflorescences with the stem terminating in an inflorescence.

The compound leaves are usually alternate and odd-pinnate. The branches are procumbent or partly erect, and the weak stem is sometimes supported by a stake in cultivation.

The inflorescence is a monochasial cyme of 4 to 12 perfect and hypogynous flowers (**Fig. 1**). Primitive tomatoes have the solanaceous trait of five flower parts, but modern tomato varieties often have more than five yellow petals and green sepals. The five anthers are joined around the pistil in *Lycopersicon*, one of the key distinctions from the closely related *Solanum* genus. Wild *Lycopersicon* species, which are self-incompatible and therefore are obligate cross pollinators, have their style exserted beyond the anther cone. Cultivated tomatoes are self-fertile, and their style length is more similar to the anther length, a characteristic that favors self-pollination.

The fruit is a berry with 2 to 12 locules containing many seeds. Most tomato varieties have red fruit, due to the red carotenoid lycopene. Different single genes are known to produce various shades of yellow, orange, or green fruit. There is no basis for the common belief that yellow-fruited tomatoes are low in acidity. Many greenhouse varieties have pink fruit, due to a single gene (*y*) that prevents formation of the yellow pigment in the epidermis of the fruit.

Distribution and economic importance. The tomato is the most important processed vegetable, constituting over 23 lb (10.4 kg) of the 54 lb (24.3 kg) of processed vegetables the average American consumes each year.

Over three-quarters of the acres of processing tomatoes grown in the United States are in California. Other leading states, in order of acreage of processing tomatoes, are Ohio, Indiana, and New Jersey. Florida is the leading state for fresh market tomatoes, followed by California, South Carolina, Alabama, and Texas. Sizable amounts of fresh market tomatoes are imported to the United States from Mexico in winter months. Tomatoes are grown in every state, including Alaska. Although most tomato varieties prefer a long, warm season, very early varieties have been bred for northern areas.

Much of the world's production of tomatoes is in the United States. Russia has the largest acreage of tomatoes, but its production is second to the United States. Other leading countries, in order of acreage, are Italy, Egypt, Mexico, Turkey, and Spain. Tomatoes are often grown in greenhouses in northern Europe, especially in the Netherlands and England. The greenhouse tomato production in the United States is centered in Cleveland, Ohio.

Cultural practices. Tomatoes prefer warm weather. Cool temperature, 10°C (50°F) and below, delays seed germination, inhibits vegetative development, reduces fruit set, and impairs fruit ripening. High temperature, above 35°C (95°F), reduces fruit set and inhibits development of normal fruit color. Tomato plants cannot tolerate frost, and home gardeners often cover plants to protect them.

The tomato plant is day-neutral, flowering when grown with either short or long days. This makes it possible to grow tomatoes outdoors during the short days of winter in frost-free areas, such as southern Florida and the Imperial Valley of California, as well as in more northern areas during the long days of summer.

In California, seeds for most commercial tomatoes are planted directly in the field. Transplants are used in shorter-season areas, with plants grown for about 6 weeks in a local greenhouse or shipped from a southern state where they are grown in the field. Home gardeners often grow their own transplants in a warm, sunny location, then harden them in a cold frame before transplanting.

A starter solution, soluble fertilizer applied when transplanting, is often beneficial. Tomatoes prefer fertile soil, but can be grown on a variety of soils with proper fertilization. Heavy nitrogen fertilization does not make plants vegetative, as is often believed, but rather delays maturity and promotes rank vine growth. Nitrogen fertilizer and irrigation are often withheld late in the season when tomatoes are to be mechanically harvested, since this increases the proportion of ripe fruit in a single harvest. Phosphorus is often applied in a band near the seed when direct-seeding.

Level, uniform fields with few stones are best for mechanical harvesting. Heavy soils should be avoided when direct-seeding because of the danger of the soil crusting before the seedlings emerge. A soil pH of 6.0 to 6.5 is recommended.

Tomatoes respond to irrigation in many areas. A uniform supply of soil moisture is needed to prevent excessive losses from cracking and blossom-end rot. Good drainage is essential for high yields.

Fig. 1. Inflorescence of tomato (*Lycopersicon esculentum*) with open flowers and flower buds in different stages of development.

Better fruit set is often obtained with greenhouse tomatoes if the flowers are vibrated to ensure pollination. This is usually not necessary in the field, since the wind suffices.

Harvesting. Tomatoes for processing are harvested when red ripe and are soon sent to a nearby cannery. Tomatoes for fresh market, however, are often harvested at an earlier stage of maturity when they are still firm and better able to tolerate shipment to distant markets. Mature green fruits will ripen normally, but more immature fruits should not be harvested since their quality will be inferior.

Ethylene, a gas that is the natural ripening hormone in tomatoes and other fruit, is sometimes applied to stimulate green fruit to ripen after harvest. Ethephon (2-chloroethylphosphonic acid), a growth regulator that liberates ethylene, has been extensively used commercially to promote ripening of tomatoes.

Most tomatoes for fresh market are harvested by hand, but almost all of the processing tomatoes in California are harvested mechanically. Mechanical harvesting is less common in other areas, but is increasing. The harvester cuts the vine off at ground level, elevates it on an inclined chain, then shakes the fruit off and discards the vine. Workers on the machine remove most of the immature and defective fruit, and the rest are conveyed to a bulk container. Since only a single harvest is possible with mechanical harvesting and bruising is possible, varieties and cultural practices are used that favor (1) a maximum proportion of simultaneously ripe fruits, (2) the ability of fruits to remain in sound condition for a prolonged period after ripening, (3) tight attachment of fruits in order to prevent them from falling off when the plant is lifted by the machine, and (4) resilient

fruits that are not easily bruised. *See* AGRICULTURAL MACHINERY.

Culinary and biological uses. The tomato is highly esteemed as a source of vitamin C. Tomato varieties differ in ascorbic acid content, and many tomato breeders are striving to increase the content of vitamin C. Most tomato varieties have about 20–25 mg ascorbic acid per 100 g, and one medium-sized tomato provides about half of the required daily allowance of vitamin C for adults. *See* ASCORBIC ACID.

Tomatoes are also a significant source of vitamin A. Red tomatoes have about 1000 international units (IU) of vitamin A per 100 g. Some orange-fruited varieties which have the *B* gene, such as Caro Red and Caro Rich, have 10 times as much β-carotene (provitamin A) as red tomatoes, and a single fruit will provide more than the required daily allowance of vitamin A. *See* VITAMIN A.

Tomatoes are a good source of protein, but most of it is in the seeds. Tomato juice contains 19 amino acids, principally glutamic acid.

Most tomato varieties have 4.5 to 7.0% soluble solids, much of it as fructose or glucose. Citric acid is the predominant acid in tomato juice. The pH of tomatoes for processing should be below 4.5 to prevent spoilage, but the belief that modern hybrid varieties are lower in acidity than older varieties and more susceptible to botulism is unfounded. The ratio of sugar to acid is important for tomato flavor. The biochemistry and genetics of several of the aromatic organic compounds that contribute to tomato flavor have been determined.

The tomato has been a favored organism for genetic studies. Over a thousand genes are known for the tomato, and several hundred of these have been located on their respective chromosomes. Features of the tomato plant that make it very desirable for genetic and physiological studies include its high rate of natural self-pollination, the ease of controlled cross pollination, the large number of seeds per fruit and per plant, its ease of propagation by seed and cuttings and of culture in the field and the greenhouse, the relatively short time for each generation to reach maturity, the economic importance of the crop, and the wide assortment of available mutants, some of them with known effects on vitamin formation, carotenoid biosynthesis, hormones, mineral nutrition, water relations, and other physiological processes. *See* AGRICULTURAL SCIENCE (PLANT); FOOD MANUFACTURING; PLANT PHYSIOLOGY.

R. W. Robinson

Germplasm. The ability of living organisms to pass on specific morphologic traits (form) and physiologic traits (function) to their offspring is due to the presence of germplasm, the self-reproducing deoxyribonucleic acid (DNA) found in the nuclei of their cells. There are large numbers of genes (discrete portions of the DNA) involved in such complex organisms, and after repeated cell divisions, significant numbers of minute changes occur in some of these genes. These changes may be expressed as variants, or mutants, of the parent plant. A small portion of these variants are better adapted than the parent to their environment; thus, they are "selected" in preference to the parent, or wild, type.

Fig. 2. Tomatoes, with genetic diversity, range from small-fruited wild types to large-fruited commercial types.

Like many other self-pollinated crop plants, the tomato is deficient in genetic diversity. As measured by molecular marker genes, the level of genetic variation prior to 1940 was vastly smaller in cultivated tomatoes than in related wild species. The same contrast was evident in genes for pest resistance, stress tolerance, fruit quality, and other economically desirable traits. Diversity among cultivated tomatoes was so depleted that little progress was made in breeding improved cultivars in the post-Mendelian period of 1900–1940. Recognition of this situation led tomato plant breeders to search for sources of increased diversity, such as spontaneous mutations (for example, determinate habit), induced mutations (for example, male sterility), and land races and related wild species (**Fig. 2**).

All wild species of *Lycopersicon* can be hybridized with *L. esculentum*, (with varying degrees of difficulty), permitting access to vast germplasm resources of *Lycopersicon*. Nearly all the major genes for disease resistance have been derived from land races and related wild species. Through experimental introgression, such enrichment of tomato germplasm has led to rapid improvement in tomato quality during the period from 1940 to the present, and has been partly responsible for up to a fivefold increase in yields per unit area compared with the preceding period. *See* AGRICULTURAL SCIENCE (PLANT); BREEDING (PLANT).

In the United States, tomato germplasm is maintained in two major public collections. The Tomato Genetics Resource Center at the University of California, Davis, has approximately 3400 accessions of wild species (9 species of *Lycopersicon* and 4 of related *Solanum*) and various genetic mutant and chromosomal variants. The U.S. Department of Agriculture/Agricultural Research Service (USDA/ARS) Plant Genetic Resources Unit in Geneva, New York, contains approximately 5000 items, mostly cultivars and other lines of *L. esculentum*, as well as representatives of the wild species. The main functions of each unit are to (1) acquire useful germplasm not already in the collection; (2) maintain adequate, viable seeds of each accession by procedures appropriate to

preserving their innate diversity; (3) evaluate each new accession in detail and repeat plantings of others for germination and phenotypic purity; (4) supply seed samples when requested by qualified investigators; (5) document all pertinent information in databases, which are also available on the World Wide Web. Both collections are backed up by samples in the National Seed Storage Laboratory at Fort Collins, Colorado, as insurance for long-term preservation.

Tomato breeders utilize these and other resources for developing improved cultivars. Progress to date has been accomplished largely through conventional breeding methods—hybridization (cross-mating of two genetically different plants) or backcrossing (first generation hybrid is crossed with parent generation) followed by pedigree selection of desired types. Recently, introgression of desired traits has been facilitated by the rapidly developing molecular technology of gene transposition (movement of DNA to new location in the chromosome); additional improvements are effected by transformation (movement and insertion of foreign DNA) and other manipulations. In the United States, F_1 (first-generation cross) hybrid cultivars constitute the majority of plantings for commercial production of fresh-market and processing tomatoes and for a large share of home gardens. Parent lines of such hybrids are cultivars or breeding lines derived as described above. *See* GENETIC ENGINEERING; SOMATIC CELL GENETICS.

This system of germplasm management and utilization facilitates the production of this nutritionally important crop for a steadily increasing worldwide market. Raymond L. Clark; Charles M. Rick

Diseases. Tomato plants are subject to well over 100 different diseases, most of which are caused by microorganisms (biotic agents), but some are caused by abiotic agents. The most frequent biotic agents are fungi, bacteria, viruses, and nematodes, while the most frequent abiotic agents are air pollutants, genetic disorders, and unbalanced mineral nutrition.

Some diseases of tomato occur wherever tomatoes are grown, whereas others may be restricted to certain regions. The absence of a disease in a region may be the result of the absence of the pathogen, resistance in the tomato cultivars, or unfavorable climate. A change in one or more of these factors may result in a sudden outbreak of disease. Some very common tomato diseases are discussed below.

Seedling damping-off. A wet rot of the roots, hypocotyl, or stem is characteristic of this disease. Death of seedlings usually occurs just before or just after seedling emergence from soils, or a few days after transplanting seedlings into soils. The three types of fungi that can cause this disease are *Rhizoctonia* spp., *Fusarium* spp., and *Pythium* spp. These fungi are all common inhabitants of most soils, but occur there in low populations. Prolonged wet weather or planting in soils that recently received green plant refuse increases the populations of the pathogens.

Root knot. Root knot is caused by a group of nematodes belonging to the genus *Meloidogyne*. The symptoms are conspicuous on roots because of abnormal growths called galls. The shoots of diseased plants appear pale green, are unvigorous, and sometimes wilt in hot weather. The nematodes occur naturally in soils in warm regions, and cause disease in many plants in addition to tomato. Some cultivars are resistant to root knot. *See* NEMATA (NEMATODA).

Wilts. Plants in the fruiting stage may wilt even though no obvious damage to roots has occurred and soil moisture is adequate for growth. Such wilting may be associated with a brown color in the stems. Three different microorganisms can cause this type of wilt, but each is different in its rapidity of wilt development. The fungus *Verticillium alboatrum*, most severe in cool weather, causes a slow-moving wilt with stunting, rather than killing, of the plant. The fungus *Fusarium oxysporum* f. sp. *lycopersici* causes a more rapidly moving wilt of the plant, which eventually dies. It is most severe in hot weather. The bacterium *Pseudomonas solanacearum* causes the most rapid wilt, and the plant usually dies soon after the first symptoms have been observed. All of these pathogens survive indefinitely in the soils. Resistant cultivars have been developed to all three wilt diseases.

Leaf spots and blights. Numerous necrotic spots and blights may develop on leaves and stems of tomato plants, some of which are very destructive and cause complete defoliation. The most damaging and common of these diseases are late blight and early blight. Late blight is caused by the fungus *Phytophthora infestans*. The first symptoms are greenish-black water-soaked spots. As the spots enlarge, a zone of white mold can be seen on the underside of leaves and at the margins of the spots when conditions are moist. Spread of the disease can be so rapid in wet cool weather that the entire plant will be killed. Early blight is caused by another fungus, *Alternaria solani*. As with late blight, spots may enlarge so that entire leaves are destroyed. The spots of early blight are different, however, in that conspicuous concentric rings occur in each lesion. A yellow color extends beyond the lesions as they age. Early blight is most severe on senescent leaves when plants bear enlarging fruit. High humidity and warm temperatures favor disease development.

Fruit rots. The two major rots of fruit are bacterial soft rot and anthracnose. Soft rot is caused by *Erwinia carotovora*. Breakdown of fruit, which can be very rapid at room temperatures, starts as a small water-soaked area, and continues until the fruit resembles a bag containing water. Insects with chewing larvae sometimes carry the bacterium from fruit to fruit in the field. As diseased fruit is harvested and washed, the bacterium is spread to healthy fruit. Many other fruits and vegetables are also susceptible to the soft-rot bacterium.

Anthracnose is a common disease of fruit that are allowed to grow close to the soil. The causal fungus, *Colletotrichum phomoides*, invades green fruit, but does not grow and cause rot until the fruit turn red.

In addition to the symptom of an expanding, sunken, circular rot, the signs of the fungus can be seen in the rotted portion, appearing as black areas, which may be arranged in concentric rings. In moist conditions, pink spore masses may appear on the surfaces of rotted areas. The fungus remains in soils on tomato debris from season to season, but since it only grows on tomatoes, rotation of tomatoes with other crops may be helpful to control the disease.

Mosaics. Several viruses cause mosaics of light green and dark green areas in leaves. These viruses also cause misshapen leaves, stunted growth, and sparse fruit set. Tobacco mosaic virus is probably the most common viral pathogen of tomato. The source of the virus is often tobacco products used by workers; their hands become contaminated with the virus and they transmit it from plant to plant during cultivation. Plants at all stages of growth are susceptible, though there are some resistant cultivars. Potato virus Y, another prevalent viral pathogen of tomato, produces symptoms very similar to those of tobacco mosaic. The virus differs, however, in that it is usually transmitted from plant to plant by aphids. Many solanaceous weeds are also susceptible to this virus, and they serve as reservoirs of the virus between crops. Two or more viruses may infect tomato plants at the same time; the multiple infections may cause more severe symptoms than those caused by either virus alone. *See* PLANT VIRUSES AND VIROIDS.

Abiotic diseases. Blossom-end rot of tomato fruit is caused by an imbalance in calcium in the fruit. Two to three weeks after fruit set, near the end of the fruit that is opposite the end attached to the plant, a water-soaked area may appear that rapidly enlarges and becomes sunken. Secondary fungi grow in the rotted area, which turns black. The calcium imbalances that cause this condition can themselves be caused by improper water content of soils or by the use of too much soluble fertilizer. Blotchy ripening of fruit, characterized by portions of the fruit turning yellow rather than red, is another abiotic problem of tomato fruit. The causes of this disorder are not clear, but certain cultivars are affected more than others. *See* PLANT PATHOLOGY.

Robert E. Stall

Bibliography. J. G. Atherton and J. Rudich, *The Tomato Crop: A Scientific Basis for Improvement*, 1987; W. A. Gould, *Tomato Production, Processing, and Technology*, 1992; J. B. Jones et al. (eds.), *Compendium of Plant Diseases*, 1991; C. M. Rick, Tomato, in J. Smartt and N. W. Simmonds (eds.), *Evolution of Crop Plants*, Longman, London, 1995; C. M. Rick and R. T. Chetelat, Utilization of related wild species for tomato improvement, *Acta Hort.*, 412:21–38, 1995; C. M. Rick and J. I. Yoder, Classical and molecular genetics of tomato: Highlights and perspectives, *Annu. Rev. Genet.*, 22:281–300, 1988; S. D. Tanksley and S. R. McCouch, Seed banks and molecular maps: Unlocking genetic potential from the wild, *Science*, 277:1063–1066, 1997; U. S. Department of Agriculture, *Agricultural Statistics—1998*, 1998; J. Yoder, *Molecular Biology of Tomato: Fundamental Advances and Crop Improvement*, 1993.

Tommotian fauna

The first diverse assemblages of unquestionable animal fossils at the Proterozoic-Phanerozoic transition, which marks the change from a predominantly microbial biosphere to a modern type of biosphere abundant with multicellular life. The name derives from the Early Cambrian Tommotian Stage in Siberia, where the significance of this fauna of early skeletal fossils (often referred to as "small shelly fossils") was first realized. However, the concept goes beyond the geographical and temporal boundaries of the Tommotian Stage. It can be traced back to the low-diversity assemblages of skeletal animal fossils appearing near the end of the Neoproterozoic, continuing into the Cambrian, in Siberia first in moderate diversity in the Manykaian (Nemakit-Daldynian) Stage, then burgeoning in the Tommotian Stage itself. *See* CAMBRIAN; FOSSIL.

Traditionally, the fauna was considered to predate the earliest trilobites and was therefore often referred to as "pre-trilobitic"; however, recent stratigraphic studies suggest that at least some of the Tommotian Stage correlates with trilobite-carrying beds elsewhere. A number of the characteristic Tommotian taxa are now also known to continue into post-Tommotian strata. Thus, freed of its tight stratigraphic constraints, the Tommotian fauna is known to have radiated from all continents, but it is particularly diverse and abundant on the Siberian Platform in Russia, in Australia, and in the belt of phosphorite-rich deposits that extends from the South China Platform through Mongolia, Kazakhstan, the Himalayas, and Iran. *See* TRILOBITA.

The Tommotian fauna, being mainly restricted to animals with hard parts, does not yield as complete a picture of animal anatomy and diversity as the more well-known Cambrian Burgess Shale and Chengjiang (South China) faunas or the Upper Cambrian orsten fauna, all of which are characterized by frequent preservation of soft tissues. Nonetheless, it has been instrumental in forming our understanding of the Cambrian explosion (dramatic evolutionary radiation of animals beginning about 545 million years ago), because its preservation is generally not dependent on extraordinary conditions and is therefore less spotty. Also, fossilized embryos of Tommotian animals make it possible to understand the complete life cycles of some of the most basal members of the metazoan evolutionary tree. *See* ANIMAL EVOLUTION; BURGESS SHALE.

Faunal composition. Characteristic morphologic features of the Tommotian fauna include mineralized tubes, spicules, sclerites, and shells, often belonging to animals of unknown affinities. The minerals involved are opal (a hydrated gel of silica), apatite (calcium phosphate), and aragonite/calcite (calcium carbonate)—the same minerals that are common in animal skeletons today.

Tubes. The tubes represent many different kinds of animals that built protective sheaths reinforced with minerals. One of the earliest known such forms, the late Neoproterozoic *Cloudina*, had a multilayered

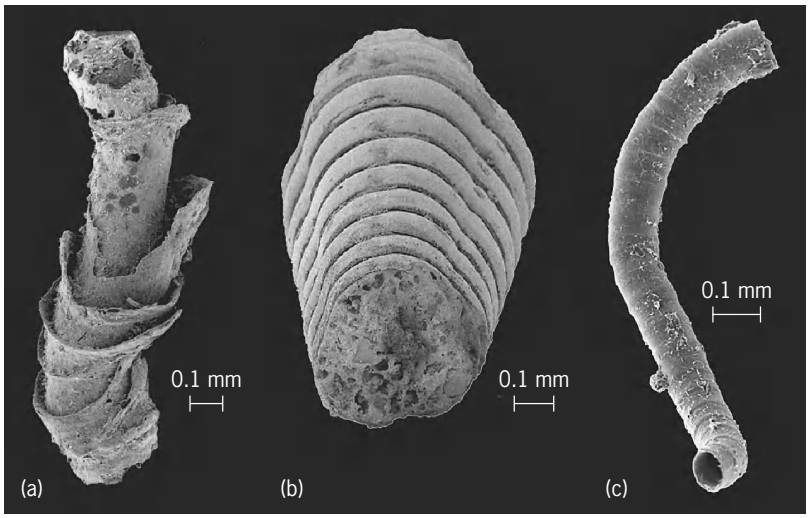


Fig. 1. Tube-dwelling animals of the Tommotian fauna. (*a*) *Cloudina*, one of the earliest animals with a mineralized skeleton. (*b*) *Aculeochrea*, an anabaritid showing the three-rayed symmetry typical of the group. (*c*) *Hyolithellus*, an animal reinforcing its tube with calcium phosphate.

cone-in-cone structure consisting of thin calcitic laminae (**Fig. 1***a*). Already in these early mineralized animal skeletons there is evidence of attacks from shell-boring predators, suggesting that exoskeletons initially arose as a response to predation. *Cloudina* is now known to be associated with several other forms of skeletal fossils. The triradially symmetrical tubes of anabaritids (Fig. 1*b*) partly overlap *Cloudina* in stratigraphic range and are a particularly characteristic component of the Proterozoic-Cambrian boundary beds. Anabaritids may be related to cnidarians, although triradial symmetry is not a characteristic feature of later cnidarians. Associated embryos show definitive cnidarian features, although the connection between these embryos and adult anabaritids is only implied from their joint occurrence. *See* CNIDARIA.

Other possible cnidarians are represented by forms such as *Olivooides*, which shows similarities with modern scyphozoans (jellyfish). A rich embryonic material of *Olivooides* shows that it developed directly within the egg, without going through a free planula stage (**Fig. 2**).

Other tube-dwelling forms are even more difficult to put in their proper phylogenetic place because of the lack of anatomically significant characters. These include the hyolithelminths (Fig. 1*c*) and other taxa that similarly reinforced their tubes with calcium phosphate, as well as forms that used calcium carbonate for the same purpose. There were also a number of animals that constructed their tubular sheaths of sclerotized organic matter or strengthened them with agglutinated sedimentary particles.

Spicules. Spicules (spikelike supporting structures) of various composition are widespread in the Animal Kingdom today. Although a number of mineralized spicules have been reported from the late Neoproterozoic, all of these reports are contested. Sponge spicules, however, are already considered a conspicuous component of the Tommotian fauna, and there are a number of spicules probably belonging to animals other than sponges.

The cup-shaped archaeocyathans (**Fig. 3**) long were of uncertain affinities and regarded by some as representing an extinct phylum. The presence of calcareous basal skeletons in several groups of recent sponges, however, has led to the now generally accepted idea that archaeocyathans are a group of

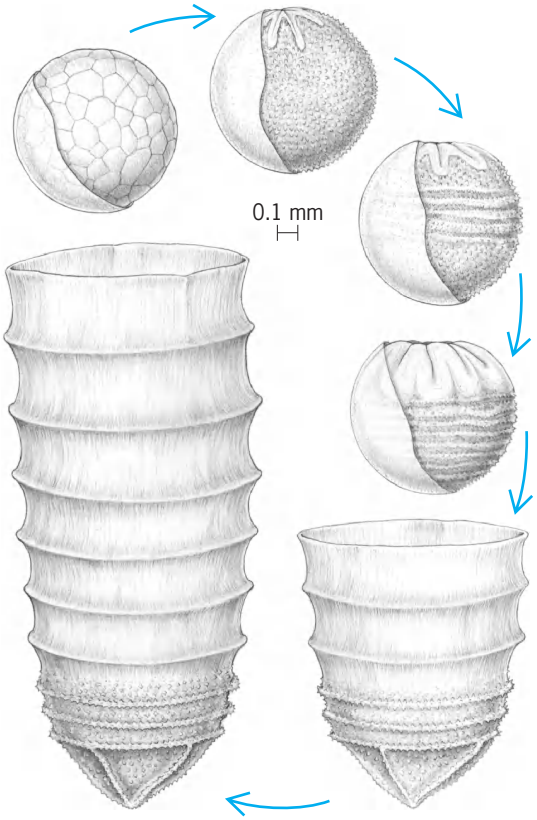


Fig. 2. Reconstruction of life cycle of the probably scyphozoan *Olivooides*, based on fossilized embryos and polyps.

Fig. 3. An archaeocyathan, interpreted to be a calcareous sponge.

calcified sponges that proliferated in the Early Cambrian and died out before the Ordovician. They are common in reefs but do not appear to have had the capacity of constructing reef frameworks. *See* ARCHAEOCYATHA.

Sclerites. A conspicuous element of Tommotian fossil assemblages is sclerites (hardened plates) belonging to composite exoskeletons, scleritomes. In many cases, the body shape of the bearer and the distribution of sclerites on the body are not known, but finds of complete scleritomes or even bodies in shale deposits give occasional and crucial insights. Thus the star-shaped composite sclerites of the chancelloriids are known to belong to a cactuslike animal that in its organization seems closest to sponges: sedentary, sac-shaped bodies with an apical orifice and no evidence of internal organs. These sclerites belong to a type called coelosclerites, consisting of a mineralized envelope around a space originally filled with soft tissue and showing no evidence of accretionary growth. The halkieriids are characterized by scale-shaped coelosclerites, and finds of complete specimens show the animal to have been slug-shaped, with two large anterior and posterior shell plates in addition to the sclerites (**Fig. 4***a*). Embryos of a segmented animal co-occurring with such sclerites have been proposed to belong to halkieriids. It is not clear whether coelosclerites are a convergent feature, independently evolved in several groups, or whether they were inherited from a common ancestor of the various groups of coeloscleritophorans.

Other sclerites were growing by stepwise accretion and often used calcium phosphate rather than calcium carbonate as shell mineral. The most widespread group of such phosphatic sclerite bearers are the tommotiids (Fig. 4*c*, *d*), of which so far no complete skeletons have been found. Tommotiids show a variety of sclerite shapes and ultrastructures and may in fact represent a polyphyletic assemblage of lineages that independently acquired a phosphatic scleritome. Other phosphatic sclerites include tooth- or hook-shaped objects as well as a variety of plate-like types, most of which are of unknown nature. The earliest known brachiopod is represented in the Tommotian fauna by a phosphatic shelled form.

Shells. Mollusk-like forms are a common component of the Tommotian fauna, although their phylogenetic placement continues to be contentious. A number of gastropodlike shells (Fig. 4*e*) occur, but it is not clear whether the soft parts did in fact possess the torsion characteristic of crown-group gastropods. Laterally flattened shells showing what appears to be an incipient dorsal hinge (Fig. 4*f*) have been interpreted as ancestral to the later rostroconchs and bivalves. A number of Tommotian taxa have been assigned to polyplacophorans, but all of these suggestions are controversial. Examples are the halkieriids (Fig. 4*a*) and siphogonuchitids (Fig. 4*b*), which have been interpreted alternatively as stem-group polyplacophorans, stem-group mollusks, and stem-group lophotrochozoans (a phylum grouping including mollusks, annelids, and brachiopods). *See* ANNELIDA; BIVALVIA; BRACHIOPODA; GASTROPODA; MOLLUSCA; POLYPLACOPHORA.

Evolutionary significance. The phylogenetic relationships of the Tommotian animals are very incompletely understood. Although some forms have with reasonable confidence been interpreted to belong to the crown group or stem group of a living phylum, others are more problematic. To some extent this is because many of them are incompletely known, perhaps only from disarticulated sclerites and no soft parts. More pertinent, however, may be the observation that the Tommotian fauna represents the first major adaptive radiation of animals and so is likely to include a number of short-lived lineages that do not share anatomical features to any greater extent than do living phyla.

The appearance in the fossil record of the Tommotian fauna near the base of the Cambrian has often been interpreted as being primarily a biomineralization event rather than a major radiation. This has then been attributed to extrinsic (for example, seawater chemistry) as well as intrinsic (for example, evolution of biomineralization pathways) factors. However, biomineralization not involving skeleton formation can be shown to be much older than the Cambrian. In addition, independent lines of evidence (trace fossils, soft-body preservation, organic microfossils) show the Cambrian explosion to have

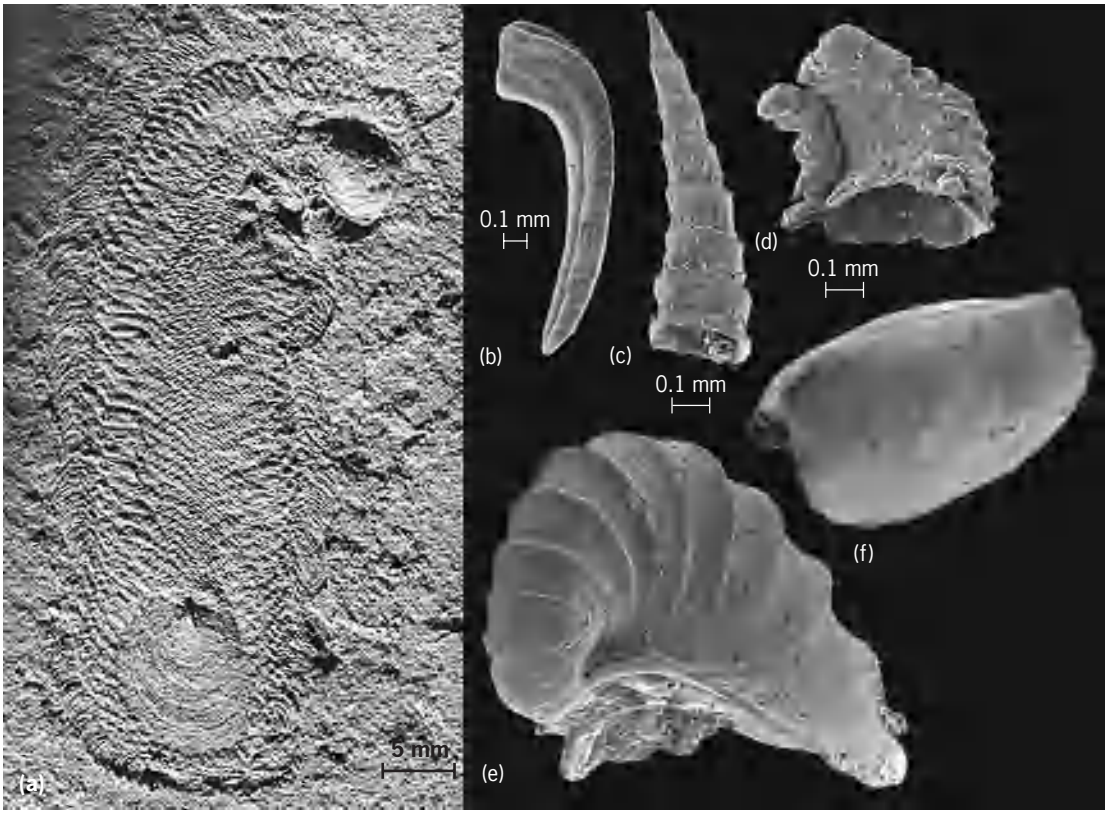


Fig. 4. Sclerite-bearing animals and probably mollusks. (*a*) *Halkieria*, a scale-and-plate-clad coeloscleritophoran (*reprinted with permission from M. S. Conway and J. S. Peel, Articulated halkieriids from the Lower Cambrian of north Greenland, Nature, 345:802–805, 1990, © Nature*). (*b*) *Siphogonuchites*, coeloscleritophoran, represented by a loose sclerite. (*c*) Sclerite of *Lapworthella*, a tommotiid. (*d*) Sclerite of *Eccentrotheca*, tommotiid. (*e*) *Archaeospira*, a possible early gastropod. (*f*) *Watsonella*, an early mollusk possibly representing an ancestral lineage to rostroconchs and bivalves (*reprinted with permission from A. P. Gubanov, A. V. Kouchinsky, and J. S. Peel, The first evolutionary-adaptive lineage within fossil molluscs, Lethaia, 32:155–157, 1999 © Taylor & Francis Group*).

involved nonskeletalized animals and nonanimals as well. Thus, the evolution of biomineralized skeletons assumes the same general significance as the evolution of various other kinds of animal tissues during this radiation event. Stefan Bengtson

Bibliography. S. Bengtson et al., Early Cambrian fossils from South Australia, *Mem. Ass. Australasian Palaeontol.*, 9:1–364, 1990; S. Bengtson and S. Conway Morris, Early radiation of biomineralizing phyla, in J. H. Lipps and P. W. Signor (eds.), *Origin and Early Evolution of the Metazoa*, Plenum, New York, 1992; E. Landing and K. E. Bartowski, Oldest shelly fossils from the Taconic allochthon and late Early Cambrian sea-levels in eastern Laurentia, *J. Paleontol.*, 70:741–761, 1996; V. V. Missarzhevskij, Drevnejshie skeletnye okamenelosti i stratigrafiya pogranichnykh tolshch dokembriya i kembriya [The oldest skeletal fossils and stratigraphy of the Precambrian-Cambrian boundary beds], *Trudy Geologicheskogo Instituta AN SSSR*, 443:1–237, 1989; Yi Qian and S. Bengtson, Palaeontology and biostratigraphy of the Early Cambrian Meishucunian Stage in Yunnan Province, South China, *Fossils and Strata*, 24:1–156, 1989; A. Yu. Rozanov, Problematica of the Early Cambrian, in A. Hoffman and M.H. Nitecki (eds.), *Problematic Fossil Taxa*, vol. 5, Oxford University Press, New York, 1986; A. Yu. Rozanov et al., Tommotskij yarus i problema nizhnej granitsy kembriya [The Tommotian Stage and the problem of the lower boundary of the Cambrian], *Trudy Geologicheskogo Instituta AN SSSR*, 206:1–380, 1969 [English transl. Amerind Publishing Co., New Delhi, 1981]; A. Zhuravlev and R. Riding (eds.), *Ecology of the Cambrian Explosion*, Columbia University Press, New York, 2000.

Ton of refrigeration

A rate of cooling that is equivalent to the removal of heat at 200 Btu/min (200 kilojoules/min), 12,000 Btu/h (13 megajoules/h), or 288,000 Btu/day (300 MJ/day). This unit of measure stems from the original use of ice for refrigeration. One pound of ice, in melting at 32°F (0°C), absorbs as latent heat approximately 144 Btu/lb (335 J/kg), and 1 ton (0.9 metric ton) of ice, in melting in 24 h, absorbs 288,000 Btu/day (300 MJ/day). In Europe, where the metric system is used, the equivalent cooling unit is the frigorie, which is a kilogram calorie, or 3.96 Btu. Thus 3000 frigories/h is approximately 1 ton of refrigeration. A standard ton of refrigeration is one developed at standard rating conditions of 5°F (−15°C) evaporator and 86°F (30°C) condenser temperatures, with 9°F (−13°C) liquid subcooling and 9°F (−13°C) suction superheat. *See* REFRIGERATION. Carl F. Kayan

Tone (music and acoustics)

Physically, a sound that is composed of discrete frequency (or sine-wave) components; psychologically, an auditory sensation that is characterized foremost by its pitch or pitches.

Physical meaning. The physical definition distinguishes a tone from a noise, wherein the components form a continuum of frequencies. Tones may be pure, consisting of a single frequency, or they may be complex. Complex tones, in turn, may be periodic or not periodic. Periodic complex tones repeat themselves at rapid regular intervals. They have frequency components that are harmonics—discrete frequencies that are integer multiples of a fundamental frequency. For example, the orchestra tunes to the tone of an oboe, a periodic tone consisting of a fundamental frequency of 440 hertz, a second harmonic component with a frequency of 880 Hz, a third harmonic at 1320 Hz, and so on. In general, musical instruments that generate continuous sounds—the bowed strings, the brasses, and the woodwinds—create such periodic tones. Tones that are not periodic (aperiodic) have frequency components that do not fit a harmonic series. Percussive instruments such as kettledrums and bells make such aperiodic tones. *See* HARMONIC (PERIODIC PHENOMENA); MUSICAL ACOUSTICS; MUSICAL INSTRUMENTS; PERIODIC MOTION.

Psychological meaning. Pitch is a sensation of highness or lowness that is the basic element of melody. Periodic complex tones tend to have a single pitch, which listeners will match by a pure tone having a frequency equal to the fundamental frequency of the periodic complex tone. Aperiodic complex tones tend to have multiple pitches. A second psychological attribute of complex tones is tone color or timbre. Tone color is often represented by descriptive adjectives. The adjectives may be linked to the physical spectrum. Thus, a tone with strong harmonics above 1000 Hz may be called "bright." A tone with no harmonics at all above 1000 Hz may be called "dull" or "stuffy." *See* PITCH; PSYCHOACOUSTICS; SOUND.

William M. Hartmann

Bibliography. D. Deutsch (ed.), *The Psychology of Music*, Academic Press, New York, 1998; W. M. Hartmann, *Signals, Sound, and Sensation*, Springer-Verlag, New York, 1997; B. C. J. Moore, *Introduction to the Psychology of Hearing*, Academic Press, London, 1997; T. D. Rossing, *The Science of Sound*, Addison-Wesley, Reading, MA, 1990.

Tongue

An organ located at the base of the oral cavity and found in all vertebrate animals. It is best developed in terrestrial vertebrates, where it takes on the functions of food procurement, food transport, and acquisition of chemosensory signals. The tongue generally is not a significant independent organ in fish, and it is secondarily reduced in organisms that feed aquatically, such as crocodilians and some turtles.

Within terrestrial vertebrates, there is considerable variability in the specific structure of the tongue, the degree of participation of the hyoid skeleton (that is, a complex of bones at the base of the tongue which supports the tongue and its muscles), and the mechanisms of movement. In birds the tongue is merely a thickened epithelium that overlies the hyoid apparatus. Movement is produced by moving various hyoid elements. In most amphibians, including both frogs and salamanders, the hyoid provides extensive support, but considerable intrinsic tongue musculature exists. In squamate reptiles (lizards and snakes) and mammals the tongue is largely independent of the hyoid apparatus and is composed entirely of muscle. The musculature is tightly packed in the tongue and is generally arranged in three mutually perpendicular planes. In the mammalian tongue the musculature is arranged into longitudinal, transverse, and vertical bundles. Organs composed entirely of muscle and lacking independent skeletal systems are widespread, and include elephant trunks and the tentacles and arms of cephalopod mollusks. Such organs have been termed muscular hydrostats.

In muscular hydrostats, muscle acts as both the effector of movement and as the support for that movement. The most important biomechanical feature of a muscular hydrostat is that its volume is constant, so that any decrease in one dimension will cause a compensatory increase in at least one other dimension. Tongue protrusion is primarily produced by elongation of the tongue, as well as movement of the tongue in space. Elongation occurs when muscle that decreases cross-sectional area contracts, a phenomenon found in the circular muscles of snakes and most lizards, or in transverse muscles in most mammals. The geometry of a muscular hydrostat may provide for amplification of muscle effect, or leverage, so that in muscular hydrostats that are protruded over long distances, the resting length of the organ is long in relation to its width. Bending requires simultaneous contraction of longitudinal muscle with another muscle, usually circular, transverse, or radial. In bending, one muscle acts to elongate the organ, while another resists this elongation at a specific spot, producing a bend. One of the primary advantages of a muscular hydrostat is that bending is not restricted to movement at joints, and the highly subdivided muscular and neural systems seen in mammalian tongues in particular produce movements that are remarkably specific, complex, and diverse.

While muscular-hydrostatic movements characterize the tongue of most mammals and many lizards and snakes, many of the most spectacular tongue projectors, such as chameleon lizards and plethodontid salamanders, do not use this mechanism in protrusion. In these animals the tongue is used as a prey capture mechanism and may be protruded with force for very long distances. These organisms have developed separate mechanisms in which the muscular tongue is projected ballistically from the body. In both, a muscle squeezes a process of the hyoid apparatus to generate the projectile force. In the

chameleon the tongue is projected off a hyoid base that remains in the mouth; in plethodontid salamanders the entire hyoid apparatus, which carries the tongue, is projected. *See* TASTE. Kathleen K. Smith

Tonsil

Localized aggregation of diffuse and nodular lymphoid tissue found in the region where the nasal and oral cavities open into the pharynx. The lymphoid tissue consists of small, closely packed round cells called lymphocytes supported in a specialized connective tissue framework called reticular tissue. When lymphocyte production is active, rounded and more densely packed clusters or nodules of these cells appear in the diffuse lymphoid tissue. The most active nodules possess lighter staining centers composed of somewhat larger, less densely packed lymphocytes showing evidence of cell division. Such areas are called germinal centers. In the tonsillar regions, the lymphoid tissue lies just beneath the lining epithelium. The tonsils are important sources of blood lymphocytes. They often become inflamed and enlarged, necessitating surgical removal.

Palatine tonsil. The two palatine (faucial) tonsils are almond-shaped bodies measuring about 1 × 0.5 in. (2 × 1 cm) and are embedded between folds of tissue connecting the pharynx and posterior part of the tongue with the soft palate (**Fig. 1**). These are the structures commonly known as the tonsils. The openings of 10 to 20 pits (crypts) which extend deep into the organ may be seen on the surface. The stratified squamous epithelium of this region covers the surface of the tonsil and lines the crypts (**Fig. 2**). A fibrous capsule separates the tonsil from the underlying muscle. Extensions from the capsule form supporting septa within the tonsil. Lymphoid tissue occupies all interstices between the capsule, septa, and epithelium. Crypts frequently become filled with detached epithelial cells, living and dead lymphocytes, and exuding fluids. Such sequestered masses form an excellent culture medium for the growth of certain bacteria and fungi. The protective quality of the crypt epithelium may be weakened by the passage of large numbers of lymphocytes through it.

Lingual tonsil. The lingual tonsil occupies the posterior part of the tongue surface. It is really a collection of 35–100 separate tonsillar units, each having a single crypt surrounded by lymphoid tissue. Each tonsil forms a smooth swelling about 0.08–0.1 in. (2–4 mm) in diameter. Since gland ducts open into these crypts, the contents are flushed out, and lingual tonsils rarely cause trouble. The epithelium is again stratified squamous, and a thin capsule is present around each unit.

Pharyngeal tonsil. The pharyngeal tonsil (called adenoids when enlarged) occupies the roof of the nasal part of the pharynx and is covered with pseudostratified ciliated columnar epithelium. The organ consists of a series of radiating folds leading forward from the region where the roof of the nasal pharynx

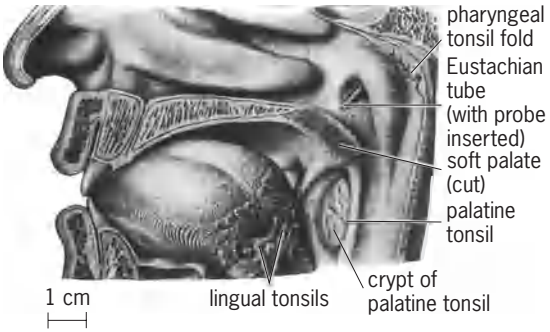


Fig. 1. Dissection showing the tonsil in relation to other structures in the pharyngeal region.

joins the posterior pharyngeal wall (**Fig. 3**). When cut at right angles to the folds, the intervening spaces resemble crypts. Septa are found in the folds, but a distinct capsule is lacking. Lymphoid tissue and lymphocytic invasion of the epithelium are similar to those of the other tonsils (**Fig. 4**). This tonsil may enlarge to block the nasal passage, forcing mouth breathing.

Development. Lymphocytic infiltration for the palatine tonsil begins at the site of the disappearing second pharyngeal pouch during the third fetal month. The pharyngeal and lingual tonsils appear during the fourth and fifth fetal months, respectively. Tonsils reach their maximum size during childhood and subsequently regress.

Function. All tonsils produce lymphocytes which are added to the circulating blood via the plexus

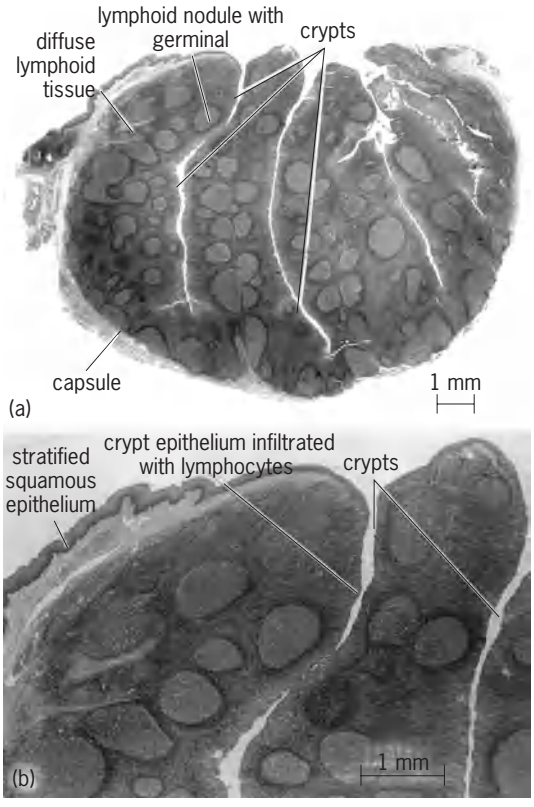


Fig. 2. Photomicrographs of a child's tonsil. (*a*) Vertical section. (*b*) Higher magnification of *a*.

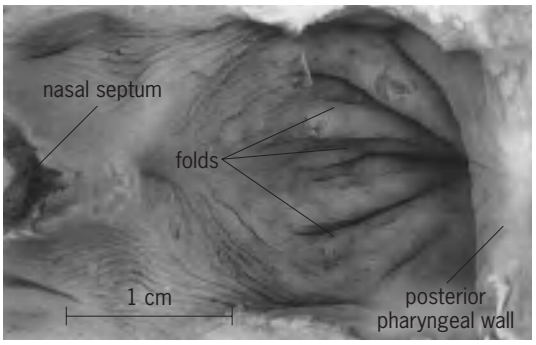


Fig. 3. Photograph of the roof of the nasal pharynx of a newborn baby showing the pharyngeal tonsillar folds.

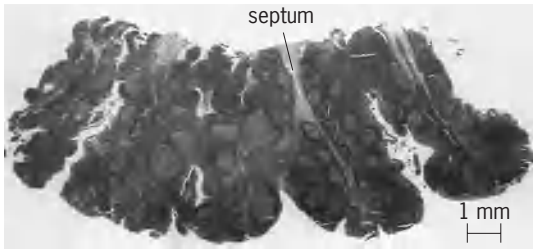


Fig. 4. Photomicrograph showing a section across the pharyngeal tonsil folds of a child.

of lymph capillaries which surrounds the lymphoid tissue. The flow of lymph is always away from the tonsillar sites. No other function has been firmly established. The three sets of tonsils along with lesser amounts of intervening lymphoid tissue form a complete ring around the upper reaches of the digestive and respiratory systems. Because of this strategic location, a protective function has been suggested. It is thought by some that protection, in response to entering bacteria, may be afforded through the production of antibodies. Evidence has shown that plasma cells (many of which are present in the tonsils), lymphocytes, or both are implicated in this process. *See* HEMATOPOIESIS; LYMPHATIC SYSTEM.

Theodore Snook

Tonsillitis

An inflammation of the tonsil. Tonsillitis is a nonspecific term usually referring to bacterial or viral infection involving all or part of Waldeyer's ring, a collection of lymphatic tissue encircling the pharynx. It consists primarily of the tonsils (palatine tonsils), adenoids (pharyngeal tonsils), and lingual tonsils. The palatine tonsils are readily seen at the edges of the soft palate between the anterior tonsillar pillar (glossopalatinus muscle) and the posterior tonsillar pillar (pharyngopalatinus muscle). The pharyngeal tonsils (adenoids) cover the vault and posterior wall of the nasopharynx and cannot be seen unless the nasopharynx is examined with a mirror or the soft palate is retracted. The lingual tonsils are located at the base of the tongue.

The purpose of Waldeyer's ring, situated at the entrance to the respiratory tract and the alimentary tract, is presumed to be related to immunization. For example, it has been shown in the 7-day-old calf and piglet that there were only small numbers of bacteria in the lymphatic drainage of the head, intestines, and lungs but that there were large numbers in the tonsils. With increased age to adulthood, the number of bacteria decreased. This suggests that the tonsils may aid in the development of immunity to the bacteria of the animal's environment.

Virus infections. The viral flora of enlarged tonsils and adenoids have been studied. Fifty-eight percent of adenoids and 26% of enlarged tonsils contained adenovirus types I, II, or V. These latent viruses were most common in the first decade, and their incidence decreased sharply at puberty. The presence of these adenoviruses may be related to tonsillar hyperplasia, wherein there is an increase in the number of secondary lymph follicles and an increase in the size of the individual follicles.

Bacterial infections. Bacterial tonsillitis is caused by beta-hemolytic *Streptococcus* and possibly by *Staphylococcus aureus*. It involves the entire Waldeyer's ring, as well as the pharyngeal mucous membrane, and more correctly should be termed pharyngotonsillitis. Attacks of bacterial tonsillitis occur with varying frequency and severity in different individuals. They are most frequent in children younger than 6 years. Unfortunately, the appearance of tonsils between attacks does not correlate well with the number or severity of infections and gives no indication as to whether past attacks have been bacterial or viral. There is likewise little correlation between clinical history and microscopic findings. *See* STAPHYLOCOCCUS; STREPTOCOCCUS.

Complications. The complications of tonsillitis depend on which tonsil is involved. Recurrent adenoiditis with adenoid hypertrophy is frequently associated with recurrent otitis media, middle-ear fluid, and at times nasal obstruction with mouth breathing and snoring. Acute palatine tonsillitis may be complicated by peritonsillar abscess which may develop lateral to the tonsillar capsule.

Treatment. Streptococcal pharyngotonsillitis is treated with penicillin for at least 10 days in persons not allergic to the drug. Removal of the adenoids is considered when there is residual middle-ear fluid. Palatine tonsils must be removed after peritonsillar abscess, but otherwise their removal depends upon the frequency of recurrent attacks of bacterial pharyngotonsillitis in relation to the patient's age. *See* TONSIL.

James A. Donaldson

Bibliography. D. D. DeWeese and W. H. Saunders, *Textbook of Otolaryngology*, 7th ed., 1987; P. M. Stell (ed.), *Scott Brown's Otolaryngology: Laryngology*, vol. 5, 5th ed., 1988.

Tooth

Any one of the structures found in the mouth of most vertebrates which, in their most primitive form, were conical and were usually used for seizing, cutting up, or chewing food, or for all three of these

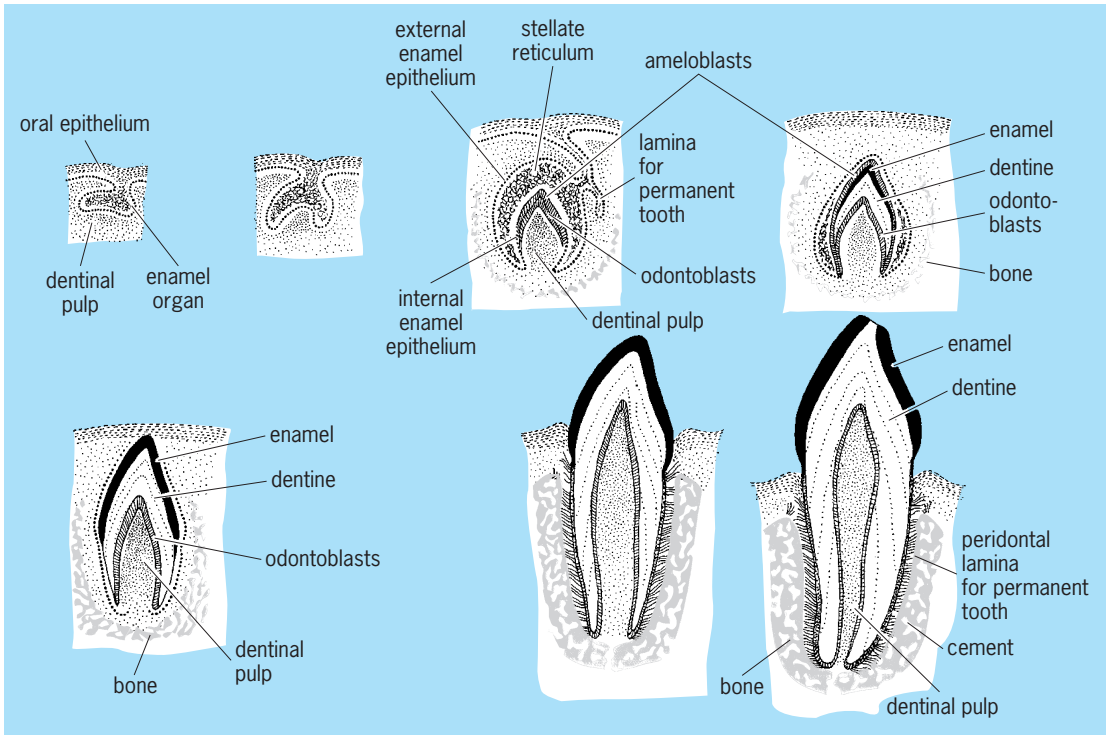


Fig. 1. Stages in the development of a tooth.

purposes. Although the true enamel of more advanced vertebrates is ectodermal, the remaining components of the teeth are mesodermal in their embryological origins (**Fig. 1**). The basic tissues that make up the vertebrate tooth are enamel, dentin, cementum, and pulp. Tooth replacement in vertebrates other than mammals may be explained by a rhythmical wave of impulses inducing the tooth germs, proceeding from the front to the back of the jaw. In their evolutionary origins teeth are derivatives of bony tubercles which developed on the outer part of the body of primitive agnathans, fishlike stem vertebrates, forming a protective shell around them. An easy confirmation of how the tubercles are thought to have developed into teeth can be seen in the identical embryonic development of the scales and teeth of young sharks.

Each tooth develops as a result of interaction between the thickening ectoderm (oral epithelium), which forms an enamel organ, and a mesodermal dental papilla. The ameloblast cells of the enamel organ secrete enamel, and the odontoblast cells of the dental papilla secrete the dentin. As the cells of the tooth bud proliferate rapidly, the tooth extends into the mesoderm, the outer layer becoming the external enamel epithelium, the inner one forming the internal enamel epithelium. The external and internal layers of enamel epithelium are separated by a mass of cells called the stellate reticulum. As the advanced tooth bud invaginates to form a cup-shaped structure around a mesodermal center, the inner epithelium cells form the ameloblast, which in turn first induces the formation of odontoblasts from the underlying mesodermal cells. As the odontoblasts begin to lay down dentin and the ameloblasts begin to secrete enamel, the two tissues are laid down in direct contact with each other.

Tissue components. Enamel is the hardest tissue in the body because of the very high concentration, about 96%, of mineral salts. The remaining 4% is water and organic matter. The enamel has no nerve supply, although it is nourished to a very slight degree from the dentin it surrounds. The fine, microscopic hexagonal rods (prisms) of apatite which make up the enamel form patterns that minimize breakage during mastication. The tiny rods are held together by a cementing substance.

Dentin, a very bonelike tissue, makes up the bulk of a tooth, consisting of 70% of such inorganic material as calcium and phosphorus, and 30% of water and organic matter, principally collagen. The rich nerve supply makes dentin a highly sensitive tissue. Although this sensitivity serves no obvious physiological function, it may be partly explained by the fact that in jawless, armored stem vertebrates the bony tubercles were not covered by enamel, and thus could function as a skinlike sensory system. As the tubercles became functional teeth, their sensory function was lost but the sensitivity remained.

Cement is a calcified tissue, a type of modified bone less hard than dentin, which fastens the roots of teeth to the alveolus, the bony socket into which the tooth is implanted. In mammalian herbivores, such as the rhinoceros, horse, and rabbit, which have evolved tall teeth, cementum is deposited between the enamel folds to strengthen the delicately folded enamel, as well as to form part of the occluding surface.

A miscellaneous tissue, consisting of nerves, fibrous tissue, lymph, and blood vessels, known as the pulp, occupies the cavity of the tooth surrounded by dentin. Lining the wall of the pulp cavity are odontoblasts which originally produced dentin tubules. If stimulated by decay or other external contact to the overlying dentin, these cells may produce secondary dentin. This is not as well organized in its structure as the first dentin, and it resembles bone.

Replacement. In most groups of vertebrates other than mammals, the individual teeth of the marginal dentition on the edge of the jaws are replaced throughout the life of the animals. Deep within the tissues of the tooth germs, teeth are constantly formed, grow, erupt, function, and are shed. Experiments on the tooth replacement of reptiles and amphibians resulted in an interesting hypothesis which seems to explain the previously unintelligible system of rhythmic replacement of teeth in nonmammalian vertebrates. Some contend that the spacing of waves of stimulation (which are genetically determined) in jaws, proceeding from front to back, results in tooth rudiments (or anlages) which begin development at rhythmically spaced time intervals. Spacing the beginning of development of tooth germs, assuming that the rate of development of teeth is even, results in the alternating appearance of erupting teeth. In **Fig. 2***a*, by the time the tooth-initiating impulse has traveled from position 1 to the position past 3, another anlage has been initiated at position 1. This means that a new tooth is started on its developmental path at a frequency slightly greater than two intervals between the induced tooth germs.

If the spacing of the tooth germs was exactly two intervals, the odd and even teeth would alternate exactly in the lower vertebrate jaw. The apparent irregular arrangement of the young, mature, and old teeth, both odds and evens, is explained by the assumption of the hypothesis that the spacing between the impulses is slightly greater than two intervals.

Evolution and phylogeny. The jawless fossil ostracoderms had no teeth as such, and presumably they were filter feeders. Teeth, along with jaws, first made their appearance among the placoderms. Although many sharks have simple conelike teeth, they usually have accessory cusps on the teeth, making their edges sharp and serrate. The teeth of various sharks and rays have been modified by being flattened into plates to crush mollusks, crustaceans, and other hard-shelled sea animals.

Both living and fossil lungfishes developed fan-shaped plates studded with ridges, or blunt, conical teeth. In the actinopterygian branch of bony fishes, relatively simple, conical teeth predominate, although a great variety of teleosts develop peculiar teeth of their own. Many teleosts have curved teeth which bend backward on a fibrous hinge, allowing the prey to enter the mouth but preventing its escape.

The probable direct ancestors of tetrapods, the crossopterygian bony fishes, as well as the earliest tetrapods, had conical teeth with groovelike folds of enamel cutting radially into the tooth all around.

In most of the reptiles simple conical teeth are predominant, although tooth plates were developed in the shellfish-eating placodonts and in several groups of herbivorous dinosaurs. Some lizards also evolved rather elaborate dentitions. Such reptiles are the ornithischian hadrosaur, the triceratopsian dinosaurs, and the recent iguanids. The mammallike reptiles evolved a great variety of complicated cheek teeth and a dentition which was differentiated into precanine and postcanine parts.

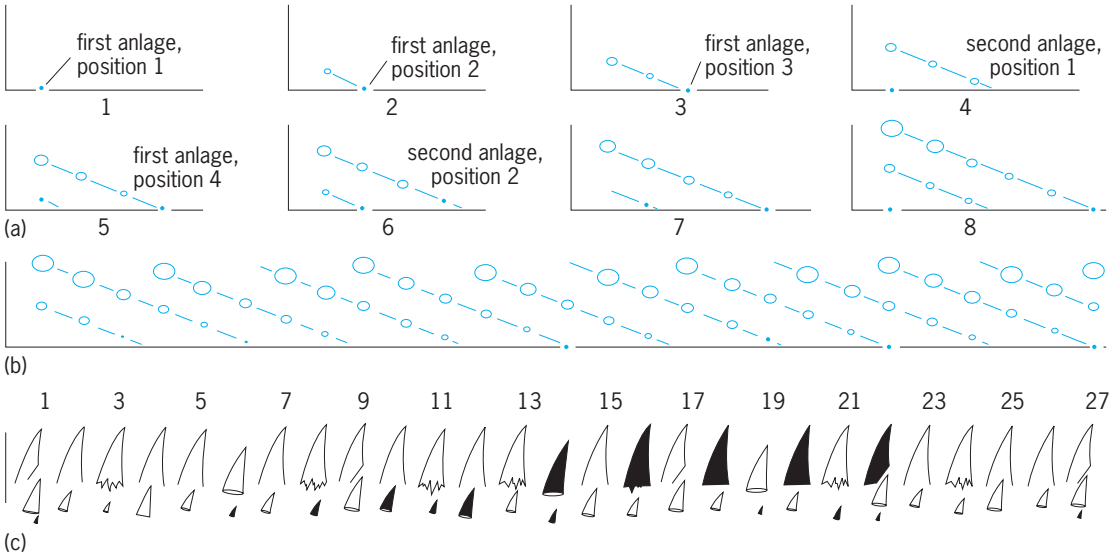


Fig. 2. Schema showing tooth replacement in lower vertebrates. (*a*, *b*) Horizontal lines represent the strand of tissue which connects the tooth germs in most vertebrates; vertical lines indicate the distance from the tooth germs to the surface of the jaw. In positions 1 to 8 of *a*, the dots on the horizontal lines represent the formation of tooth rudiments (anlages). As a result of this type of development, mature developing teeth in the jaw will be arranged as seen in *b*, and (c) they may appear from the side. In *c*, the white teeth are odd; the black are even.

Taxonomic studies. Teeth play a very important role in the study of mammalian evolution and the classification of living and fossil species. Because enamel is the hardest substance in the body, it is also most often preserved in rocks. Mammalian teeth often reflect adaptations to a particular kind of food and are therefore usually the most characteristic hard part of mammals. The crown of a tooth is genetically predetermined, and it is (at least in low-crowned teeth) fully formed before the teeth erupt. The only changes of the mammalian tooth crown, therefore, result from wear, breakage, or decay. Because it is very important that the complicated crown surfaces of the lower and upper molars occlude correctly, stringent genetic requirements determine occlusion.

Patterns of dentition. The dentition of therian mammals, at least primitively, consists of four different kinds of teeth. The incisors (I) are usually used for nipping and grasping; the canines (C) serve for stabbing or piercing; the premolars (Pm) grasp, slice, or function as additional molars; and the molars (M) do the chewing, cutting, and grinding of the food. Primitively the placentals have 40 teeth and the marsupials 50; the formulas below indicate their distribution in each side of the skull and jaw:

Primitive placental formula

$$\mathrm{I}\frac{1,2,3}{1,2,3} \quad \mathrm{C}\frac{1}{1} \quad \mathrm{Pm}\frac{1,2,3,4}{1,2,3,4} \quad \mathrm{M}\frac{1,2,3}{1,2,3}$$

Primitive marsupial formula

$$\mathrm{I}\frac{1,2,3,4,5}{1,2,3,4,5} \quad \mathrm{C}\frac{1}{1} \quad \mathrm{Pm}\frac{1,2,3}{1,2,3} \quad \mathrm{M}\frac{1,2,3,4}{1,2,3,4}$$

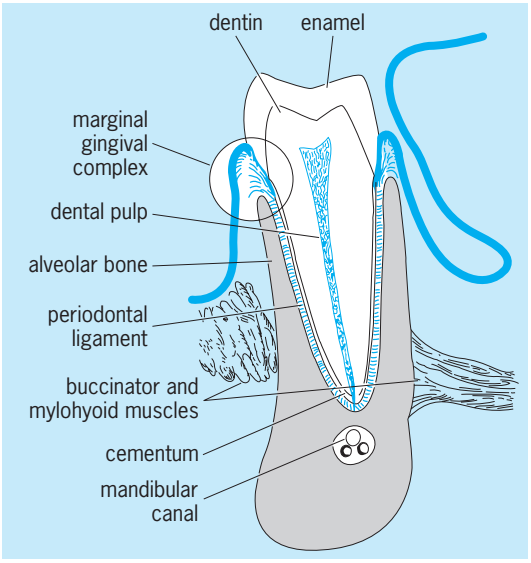
The numbers in the formulas depict the teeth in the jaw by their specific position in the series. For example, Pm_2 refers to a second lower premolar, whereas M^2 stands for a second upper molar. *See* MAMMALIA.

In therian mammals, probably because of the intricacies and vital importance of tooth occlusion, only part of the first (or "milk") dentition is replaced. This second, or permanent, dentition is made up of incisors, canines, and premolars; as a rule only one premolar is replaced in marsupials. Although the molars erupt late in development and are permanent, that is, not replaced, they are part of the first, or deciduous, dentition. *See* DENTITION. Frederick S. Szalay

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1987; R. Renner, *An Introduction to Dental Anatomy*, 1985; A. S. Romer, *Vertebrate Paleontology*, 3d ed., 1966; A. S. Romer and T. S. Parsons, *The Vertebrate Body*, 6th ed., 1986; J. Z. Young, *The Life of Vertebrates*, 1991.

Tooth disorders

Diseases and disturbances of the teeth and associated structures, including abnormal formation and growth of the teeth and jaws, tooth decay, inflammation of the tissues housing the roots of the teeth, and various diseases of the jaw bones. Tooth disorders are important because of their integral relationship to general body health, the essential role of the teeth in mastication and speech, and the contribution of normal dental function to the individual's psychological sense of well-being.



Structural features of the normal periodontal tissues.

Dentin, a calcified tissue containing viable cell processes, is the principal constituent of teeth. The dentin of the crown is covered by enamel, a nonviable tissue, and that of the root by cementum, a tissue closely resembling bone. The dental pulp contains nerves, blood vessels, and lymphatics which provide sensation and nutrition to the dentin. The root of the tooth is united to its body housing by the collagen fibers of the periodontal ligament (see **illus.**).

Abnormal formation. Defective formation of dentin and enamel are referred to respectively as dentinogenesis and amelogenesis imperfecta, and may be caused by febrile illness during the period of tooth formation, in hypophosphatasia, or by faulty calcium or phosphorus metabolism, such as occurs in rickets. In other cases, the causes remain unknown. Some or all of the teeth may fail to form completely, or extra or supernumerary teeth may be present. Abnormalities in the growth of the bones of the jaws and face with resulting faulty tooth position and malocclusion affect about 70% of all humans; in about 5%, these abnormalities are sufficiently severe to be disfiguring and handicapping.

Decay. In most societies, dental decay, or caries, is one of the most important and common tooth disorders. Decay rates are highest in individuals in the age range of 11–18 years who belong to low socioeconomic groups. Cavities arising in these individuals occur mostly in the developmental pits and fissures of the tooth crown and on the smooth surfaces between the teeth near the gum margin. Decay of the root surfaces becomes prevalent in individuals in the age range of 55–65 years. Decay occuring during the adolescent years is caused by a class of

microorganisms referred to as the cariogenic streptococci, of which *Streptococcus mutans* is the predominant member. However, a susceptible host and a cariogenic diet containing sucrose are also essential factors. Although the mechanism by which bacteria cause decay is not completely understood, most experts believe that cariogenic organisms, by using sucrose, and to a lesser extent other sugars, produce polymers which bind the organisms to the tooth surface and acids which cause demineralization resulting in cavity formation. Once the carious lesion penetrates the enamel and enters the dentin, the viable cell processes are affected and the tooth may become painful. At this stage, the organisms have relatively ready access to the pulp, where they may cause the formation of microabscesses. The associated inflammation and swelling cause severe pain and toothache. The pulp may become necrotic, causing formation of a periapical abscess, granuloma, or cyst in the alveolar bone at the root tip where the nerves and vessels enter the pulp. At this stage of disease, the tooth can be treated successfully by resolving the infection, removing the necrotic pulp, and obliterating the pulp chamber with inert filling material.

Fluoride administration, either in the drinking water at a concentration of 1 part per million (ppm), or by other routes at comparable levels, effectively reduces dental decay about 50–70%. This level of fluoride consumption appears to be without toxic or undesirable side effects, although concentrations in excess of 3–5 ppm may lead to browning or mottling of the enamel. The mechanism by which fluoride causes decreased decay rates is not understood, although some experts believe that the ion combines with hydroxyapatite, leading to the formation of more perfect crystals which are more resistant to acid attack. Fluoride may also enhance remineralization of beginning lesions.

Periodontal disease. Whereas dental decay is the principal cause of tooth loss in young individuals, inflammatory disease of the tissues surrounding the teeth, referred to as periodontal diseases, causes most of the tooth loss in adults. *See* PERIODONTAL DISEASE.

Jaw diseases. Diseases of the jaws may affect the teeth. The most common jaw disorders, other than abnormalities of growth and development discussed above, fall into four categories: (1) inflammation of the jawbone caused by infections such as osteomyelitis; (2) cysts associated with the teeth as well as those located in bone sutures of the jaws; (3) benign and malignant tumors of the jaws including osteoma, fibroma, sarcoma, tumors derived from tooth tissues, multiple myeloma, leukemia, and carcinoma; and (4) systemic diseases such as generalized skeletal abnormalities produced by endocrine dysfunction in which the jaws are affected. *See* DENTISTRY; TOOTH.

Roy C. Page

Bibliography. E. Newbrun, *Cariology*, 3d ed., 1989; E. Newbrun (ed.), *Fluorides and Dental Caries*, 3d ed., 1986; S. Schluger et al., *Periodontal Disease*, 2d ed., 1989.

Topaz

A mineral best known for its use as a gemstone. Crystals are usually colorless but may be red, yellow, green, blue, or brown. The wine-yellow variety is the one usually cut and most highly prized as a gem. Corundum of similar color sometimes goes under the name of Oriental topaz. Citrine, a yellow variety of quartz, is the most common substitute and may be sold as quartz topaz.

Topaz crystal habits. (*After C. Klein and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., Wiley, 1993*)

Topaz is a nesosilicate with chemical composition $Al_2SiO_4(F,OH)_2$. The mineral crystallizes in the orthorhombic system and is commonly found in well-developed prismatic crystals with pyramidal terminations (see **illus.**). It has a perfect basal cleavage which enables it to be distinguished from minerals otherwise similar in appearance. Hardness is 8 on Mohs scale; specific gravity is 3.4–3.6.

Topaz is found in pegmatite dikes, particularly those carrying tin. It is also formed during the late stages of the solidification of rhyolite lavas. The minerals characteristically associated are tourmaline, cassiterite, fluorite, beryl, and apatite. It is also found as rolled pebbles in stream gravels. Fine yellow and blue crystals have come from Siberia and much of the wine-yellow gem material from Minas Gerais, Brazil. In the United States topaz has been found near Florissant, Colorado; in Thomas Range, Utah; in San Diego County, California; and near Topsham, Maine. *See* GEM; SILICATE MINERALS.

Cornelius S. Hurlbut, Jr.

Topographic surveying and mapping

The measurement of surface features and configuration of an area or a region, and the graphic expression of those features. Surveying is the art and science of measurement of points on, above, or under the surface of the Earth. Topographic maps show the natural and cultural features of a piece of land. The natural features include configuration (relief), hydrography, and vegetation. The cultural features include roads, buildings, bridges, political boundaries, and the sectional breakdown of the land. Topographic maps are used by a wide variety of people, such as engineers designing a new road; backpackers finding their way into remote areas; scientists describing soil or vegetation types, wildlife habitat, or hydrology; and military personnel planning field operations.

Presentation of information. Topographic map features can be expressed in a variety of ways. Maps that show natural and cultural features only in plan view are called planimetric maps, while maps that

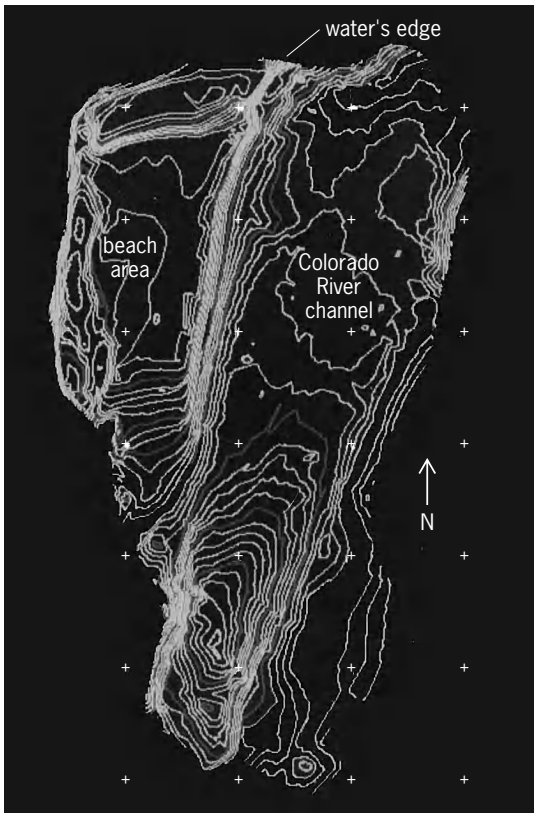


Fig. 1. Topographic and hydrographic map of the beach and river channel at River Mile 194 of the Colorado River in the Grand Canyon.

show relief are called hypsometric maps. Contour lines join points along a line of the same elevation across the ground. Contours show not only the elevation of the ground but also the geomorphic shape of features. Steepness of slope is determined by how close together the contour lines fall: the closer together the lines are, the steeper the slope. This is the most common way to show the physical characteristics of a piece of land (**Fig. 1**). *See* CONTOUR.

Hachures, also called vertical shading, are short lines drawn in the direction of slope: closely spaced and heavy for steep areas, farther apart and finer for more gentle slopes. Hachures give an impression of the shape of the land, but no indication of the actual elevation being represented. Shading and tinting are used in much the same way.

On maps that are used for navigation purposes, form lines are sometimes used. Form lines are like contour lines in that they follow lines of the same elevation, but they are not drawn with as high a degree of accuracy. This gives an impression of the shape of the features but does not show the actual elevation of the line itself.

A digital terrain model (DTM) is a computer-generated grid laid over the topographic information, which can then be rotated, tilted, and vertically exaggerated to give a three-dimensional view of the ground from different perspectives, including oblique representations. This technology is an excellent presentation tool: it utilizes the advances that have been made in computer mapping and drafting software.

Survey control. Prior to starting collection of data, a network of known horizontal and vertical control points must be established. This gives a reference location to the map, whether it be a local coordinate system or a state plane coordinate system, or geographically referenced to the rest of the world with latitude and longitude. The network also allows measurements made from several different locations in the same coordinate system to fit together into the same reference datum (the basis for the coordinate system).

Field methods. These include ground surveys, geographic positioning systems, and hydrographic surveys.

Ground surveys. Alidade and transit stadia surveys are performed, on occasion, but because of the increased speed and accuracy (centimeter-level measurements) most ground-based survey data are collected with a total station survey instrument and an electronic data collector. This information is transferred to a computer for conversion to coordinate locations and elevations. These data can then be imported into a modeling software package that generates contours at a specified interval by using description codes from the collected data to show breaks in the topography, and to draw in features such as roads, railway lines, streambeds, or structures. The information can then be plotted to create a hard copy of the maps at the desired scale to adequately display the information for the map's intended purpose. *See* SURVEYING INSTRUMENTS.

Geographic positioning system (GPS). This type of survey can be made to define topography. The Geographic Positioning System utilizes orbiting satellites operated by the U.S. Department of Defense. The satellites send out radio signals giving the precise location of the satellite and the time to a receiver on the ground. Several satellites are observed at the same time, and the position of the receiver on the ground is determined by correcting the signals sent by the satellites to a known point on the ground. The technology of the Geographic Positioning System provides virtually the same accuracy as a total station survey, but the receivers are too expensive to be in wide use. As the radio signals cannot pass through solid obstructions, this type of surveying is sometimes difficult, if not impossible, in cases of heavy tree canopy, tall buildings blocking the sky, or other features interfering with the signals.

Hydrographic surveys. The survey of a body of water, whether the ocean, a harbor, bay, lake, reservoir, or river, is a hydrographic survey. This type of survey is used to define the shoreline, depth of the water, shoals, rock outcrops, or other navigation hazards, as well as tidal changes and currents. Hydrographic surveys are used for dredging and maintenance of harbors, determining siltation of estuaries and bays, locating navigable channels, determining the sediment transport in rivers, and evaluating habitat for aquatic wildlife.

To produce a map of the bottom of a body of

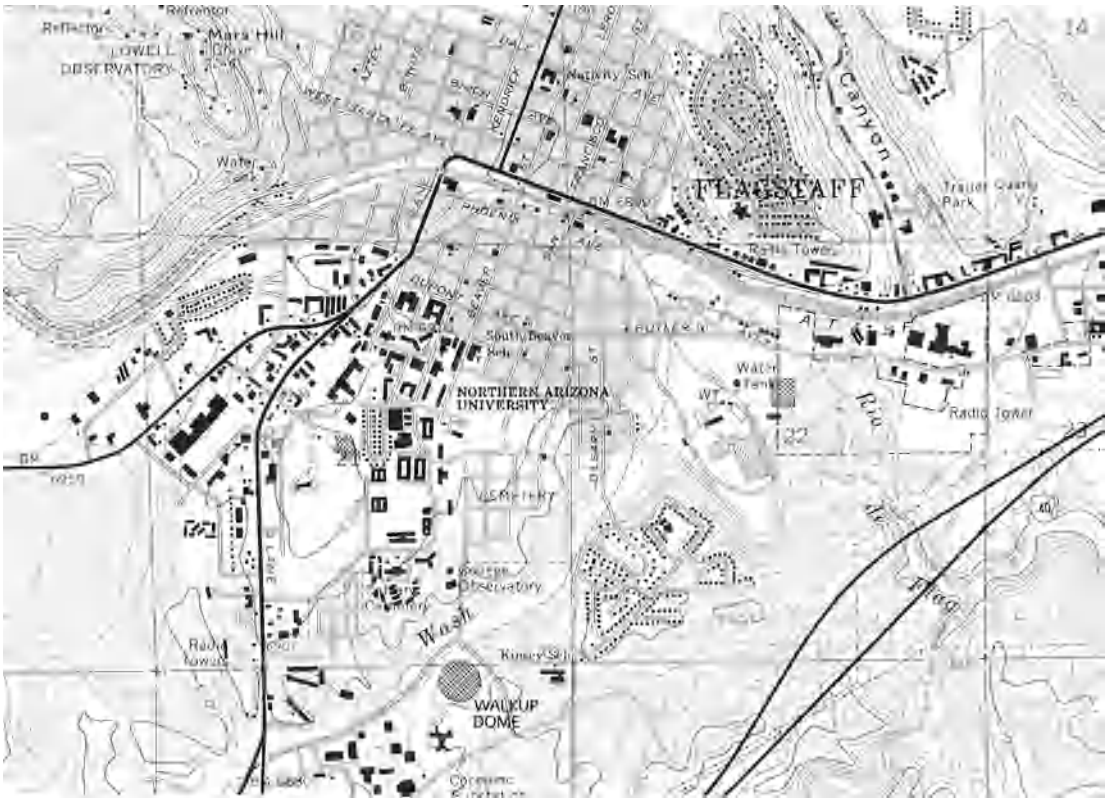


Fig. 2. Portion of USGS quadrangle sheet, showing topography and cultural information.

water requires combining the location of a point with the depth of the water at that point. Determining depth can be done in a variety of ways, from simply using a weight on the end of a graduated line dropped from a boat, to using a boat equipped with sonar to measure depth. After a control network has been established, the position of the boat can be recorded. A survey instrument set up on shore can be used to track the boat, with distance and angle measurements recorded and time-integrated with the recorded depths. This can be done either by hand recording or by using electronic devices that send the measurements from the shore to the boat by radio, where a computer program combines and stores the data. This process creates depth and position data for a great many points in a short amount of time. The position of the boat can also be determined by inertial guidance systems, distance measurements made from two or more radio beacons, or Global Positioning System locations. *See* HYDROGRAPHY; SONAR.

Photogrammetry and remote sensing. These techniques involve the art, science, and technology of using photography to obtain reliable measurements. The first aerial photography took place in 1858 to map the countryside. Emperor Napoleon III ordered aerial reconnaissance photos taken to prepare for the battle of Solferino in northern Italy in 1859. English, Russian and Germans used kites and rockets to bring single, multiple, and even gyro-stabilized cameras over their mapping areas. Photographs can be taken from airplanes, helicopters, and even satellites; thus the term remote sensing is applied to this technology.

By using pairs of photographs, a stereo image can be produced form which very accurate locations and elevations can be determined. This can then be transferred to a hard-copy map of a specific scale. The U.S. Geological Survey (USGS) has produced topographic maps of most of the United States showing topography, hydrology, vegetation, cultural features, and political boundaries (**Fig. 2**). The USGS has produced topographic maps of the Moon and Mars, and is mapping other planets and moons as well, by using radar and digital images from satellites near these bodies.

Although photogrammetry is most commonly used to map the ground, it can also be used to map objects that are difficult to study in other ways. The objects can be as different as the face of a dam, an astronomic radio reflector that deforms because of environmental conditions, or even the shapes of living organisms. Different types of applications use different types of collection techniques. Microwave, infrared, radar, ultraviolet, and multispectral data collection can be focused on different needs, such as thematic mapping, resource surveys, and topographic maps for foresters, engineers, geographers, geologists, surveyors, and others with special mapping needs. *See* AERIAL PHOTOGRAPH; PHOTOGRAMMETRY; REMOTE SENSING.

Plotting. Survey data are plotted primarily by mechanical devices, rather than being drafted by hand. This allows for very quick and accurate generation of maps as well as ease of making changes in scale, color, orientation, or location on the map page. These devices use ink pens, ink that is sprayed onto the medium, wax that is heated and applied to paper,

or coloring that is applied electrostatically. Some of the devices that use ink or wax are capable of producing several hundred colors, enabling the mapmaker to generate a highly customized product for a special purpose.

Maps need to show scale, contour interval, and a north arrow to illustrate orientation (magnetic, geodetic, polar, or grid north). A graphic scale or grid ticks with the coordinate location represented allows distance on the ground to be determined. A location map is also useful to provide the map reader with information on the general area that the large-scale map is detailing. *See* CARTOGRAPHY; MAP PROJECTIONS; MAP REPRODUCTION. Chris Brod

Bibliography. J. M. Anderson and E. M. Mikhail, *Surveying: Theory and Practice*, 7th ed., 1997; R. Graham, *Digital Aerial Survey: Theory and Practice*, 2002; E. M. Mikhail, J. S. Bethel, and J. C. McGlone, *Introduction to Modern Photogrammetry*, 2001; P. R. Wolf and R. C. Brinker, *Elementary Surveying*, 9th ed., 1997.

Topological dynamics

The study and application of topological transformation groups. Topological dynamics originated in the late-nineteenth-century investigations of the qualitative behavior of the solutions to the differential equations of classical mechanics by Henri Poincaré. The emergence of topological dynamics as a mathematical discipline occurred in the twentieth century with an abstract formulation of certain qualitative features of the classical systems by G. D. Birkhoff.

Topological transformation groups. A topological transformation group, or simply transformation group, is a triple (G, X, π), where G is a topological group called the phase group, X a topological space called the phase space, and π a continuous mapping of $G \times X$ onto X satisfying the homomorphism rule $\pi(gh,x) = \pi(g,\pi(h,x))$, $g,h \in G$, $x \in X$. The set $\{\pi(g,x)|g \in G\}$ is called the orbit through x. When G is the group $\mathbf{R}$ of real numbers, the transformation group is called a flow; when G is the group $\mathbf{Z}$ of integers, the transformation group is called a cascade. *See* ABSTRACT ALGEBRA; GROUP THEORY; TOPOLOGY.

Homomorphism and isomorphism. Let (G, X, π) and (G, X', π') be transformation groups. A continuous mapping φ of X onto X' is called a homomorphism if it is equivariant, that is, $\varphi(\pi(g,x)) = \pi'(g,\varphi(x))$. If φ is one-to-one with a continuous inverse, it is called an isomorphism.

Invariant measure. The Borel subsets of a topological space are the elements of the smallest σ-algebra of subsets of the space containing the open sets; a Borel measure is a measure on the Borel sets. A borel measure μ is invariant for a transformation group (G, X, π) if for all Borel sets A, $\mu(A_g) = \mu(A)$, where $A_g = \{\pi(g,x)|x \in A\}$. The use of invariant measures in the study of dynamical systems is original with Poincaré.

Classical dynamical systems. A classical (autonomous) dynamical system is a first-order system of ordinary differential equations (1). The functions F_i,

$$\frac{dx_i}{dt} = F_i(x_1, \ldots, x_n) \qquad (1 \le i \le n) \tag{1}$$

defined on an open set X in n-dimensional space ($\mathbf{R}^n$), are assumed to be such that (a) for each $y \in X$ there is a unique solution to Eq. (1), $x(t)$, $-\infty < t < \infty$, with $x(0) = y$; and (b) the solution in (a) varies continuously in y,t. A flow $(\mathbf{R}, X, \pi)$ may be associated to Eq. (1) by defining $\pi(t, y)$ to be the value at time t of the solution in (a). The homomorphism rule $\pi(s + t, y) = \pi(s, \pi(t, y))$ is a consequence of the uniqueness statement in Eq. (1). *See* DIFFERENTIAL EQUATION.

Hamiltonian systems. Let $X = U \times \mathbf{R}^m$, where U is an open subset of $\mathbf{R}^m$. Denote coordinates in X by $(q,p) = (q_1, \ldots, q_m, p_1, \ldots, p_m)$. A continuously differentiable (C^1) function H on X determines a hamiltonian system, given by Eqs. (2). In terms of the

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \qquad \frac{dp_i}{dt} = \frac{\partial H}{\partial q_i} \qquad (1 \le i \le m) \tag{2}$$

Lagrange equations for a conservative system with m degrees of freedom, $q_1, \ldots, q_m$ represent position coordinates, $p_1, \ldots, p_m$ represent impulses or momenta, and $H = T + V$ is total energy (kinetic plus potential). H is called the hamiltonian of Eq. (2). *See* HAMILTON'S EQUATIONS OF MOTION; LAGRANGE'S EQUATIONS.

Liouville's theorem. Assume the functions F_i in Eq. (1) are C^1 on X. Lebesgue measure on X is an invariant measure for the associated flow if and only if the vector $F = (F_1, \ldots, F_n)$ has divergence 0, Eq. (3).

$$\operatorname{div} F = \sum_{j=1}^{n} \frac{\partial F_j}{\partial x_j} = 0 \tag{3}$$

An immediate consequence of this theorem is that the flow associated to a hamiltonian system with a C^2 hamiltonian function preserves Lebesgue measure.

Poincaré recurrence theorem. This theorem, probably the first theorem in "ergodic theory," asserts that if a flow has a finite invariant measure μ, then with respect to μ almost every point $x \in X$ is Poisson-stable; for any neighborhood U of x there exist arbitrarily large (positive and negative) values of t such that $\pi(t,x) \in U$. Every Poisson-stable point belongs to the nonwandering set, the closed flow-invariant set of points x such that for any neighborhood U, $U \cap U_t \neq \phi$ for arbitrarily large values of t. In the case of Eq. (1), if X has finite volume and div $F = 0$, it follows that every point belongs to the nonwandering set. *See* STATISTICAL MECHANICS.

Flows on manifolds. Let X be a differentiable manifold. A vector field F on X generates the flow $(\mathbf{R}, X, \pi)$ if $(d/dt)\pi(t, x) = F(\pi(t, x))$. That is, for every C_1 function f on X, $(d/dt)f(\pi(t, x))|_{t=0} = F(x)f(x)$. In the case of a classical system of Eq. (1), the vector field is given by Eq. (4). If X is a compact

$$F(x) = \sum_{j=1}^{n} F_j(x) \frac{\partial}{\partial x_j} \tag{4}$$

manifold, every C_1 vector field on X generates a C_1 flow. *See* MANIFOLD (MATHEMATICS).

Integrals. Flows on manifolds occur naturally in the study of classical systems of Eq. (1). A C^1 function H is an integral of the system defined by Eq. (1) if its gradient is everywhere perpendicular to the vector field defined by Eq. (4); that is, Eq. (5) is satisfied.

$$\sum_{j=1}^{n} F_j(x)\,\frac{\partial H}{\partial x_j} = 0 \tag{5}$$

For example, the hamiltonian of a hamiltonian system defined by Eq. (2) is automatically an integral of that system. By the chain rule, if H is an integral, the level surface $H = c$ is flow-invariant. For almost all real numbers c, this level surface is empty or else an $(n - 1)$-dimensional differentiable manifold (Sard's theorem), and in the latter case $(\mathbf{R},\{H = c\},\pi)$ is a flow on a manifold.

Geodesic flow. Let M be a compact riemannian manifold with tangent bundle TM, and let T_1M be the bundle of tangent vectors of unit length. If v_p is a unit tangent vector at p, there exists a unique geodesic $x(t)$, $-\infty < t < \infty$, parametrized by arc length "in the direction v_p" (so that negative arc length makes sense) with $x(0) = p$, $(dx/ds)|_{s=0} = v_p$. Define $\pi(t,v_p) = (dx/ds)|_{s=t}$. Since arc length is the parameter, $\pi(t,v_p)$ is a unit tangent vector [at $x(t)$], and $(\mathbf{R},T_1M,\pi)$ is a flow, called geodesic flow. *See* FIBER BUNDLE; RIEMANNIAN GEOMETRY.

Principle of least action. Let U be an open set in $\mathbf{R}^m$. Given a positive definite kinetic energy, defined by Eq. (6), and a potential energy $V(q)$, the Lagrange equations of motion are given by Eq. (7) where $L =$

$$T(q,\dot{q}) = {}^1\!/_2 \sum_{i,j=1}^{m} a_{ij}(q)\dot{q}_i\dot{q}_j \tag{6}$$

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i} = 0 \qquad (1 \le i \le m) \tag{7}$$

$T - V$. [The impulses p_i in Eq. (2) are defined by Eq. (8).] The principle of least action of Maupertius-

$$p_i = \sum_{j=1}^{m} a_{ij}(q)\dot{q}_j \tag{8}$$

Euler-Lagrange-Jacobi states that among all the curves C on which the total energy $H = T + V$ is constant (h), the one joining a pair of points x_0, x_1 in U and representing a true motion of the system is the one which minimizes the "action integral"

$$\int_C 2T$$

Using Eq. (9), one can define a metric by Eq. (10).

$$T = h - V = (h - V)^{1/2}\left(\sum_{i,j=1}^{m} a_{ij}(q)\dot{q}_i\dot{q}_j\right)^{1/2} \tag{9}$$

$$ds^2 = 2(h - V(q))\sum_{i,j=1}^{m} a_{ij}(q)\,dq_i dq_j \tag{10}$$

The principle of least action now says that geodesics for ds^2 are true motions of the system. The speed of travel along a geodesic is $ds/dt = h - V$, but by a "change of clock" this speed can be made to be 1, and the geodesic flow results. *See* LEAST-ACTION PRINCIPLE.

Whitney sums. Let X be a differentiable manifold with tangent bundle TX. TX is the Whitney sum of continuous bundles A and B over X, $TX = A \oplus B$, if for each x the fibers of A and B at x are complementary subspaces of the tangent space at x. One can speak similarly of a Whitney sum of three or more bundles.

Hyperbolic structure. Let X be a compact riemannian manifold on which the length of a tangent vector is denoted $\|v\|$. A hyperbolic structure for a flow $(\mathbf{R}, X, \pi)$ is a decomposition of TX into a Whitney sum of three bundles, E, E^u, and E^s, each of which is flow-invariant (for the derivative of the flow) and such that (i) E is tangent to the flow (that is, to the orbits) and (ii) there exist constants $c < \infty$ and $\lambda > 0$ such that inequalities (11) and (12) are satisfied. E^s is

$$\|d\pi(t,x)v_x\| \;\le\; ce^{-\lambda t}\|v_x\| \tag{11}$$

$$(v_x \in E^s_x,\ t \ge 0)$$

$$\|d\pi(t,x)v_x\| \;\le\; ce^{\lambda t}\|v_x\| \tag{12}$$

$$(v_x \in E^u_x,\ t \le 0)$$

called the stable bundle and E^u the unstable bundle. The concept of a hyperbolic structure is also used for a diffeomorphism (or for the cascade generated by its repeated application); it differs from the hyperbolic structure of a flow only in the requirement that $TX = E^s \oplus E^u$, necessary because one cannot speak of a tangent to an orbit of a cascade.

Anosov flow. A differentiable flow on a manifold of dimension greater than 1 is said to be an Anosov flow if there exists a hyperbolic structure for the flow. The most important example of an Anosov flow is the geodesic flow on a compact riemannian manifold of negative sectional curvature.

Anosov diffeomorphism. A diffeomorphism is an Anosov diffeomorphism if there exists a hyperbolic structure for the diffeomorphism. The known examples are algebraic in nature. For example, let A be an $n \times n$ matrix with integer entries and determinant 1. As a linear transformation of $\mathbf{R}^n$, A sends the integer lattice $\mathbf{Z}^n$ to itself and therefore defines an automorphism of the quotient $\mathbf{R}^n/\mathbf{Z}^n = T^n$, the n-dimensional torus. If A has no eigenvalues of absolute value 1, the automorphism ("hyperbolic automorphism") is an Anosov diffeomorphism. J. Franks and A. Manning have shown that any Anosov diffeomorphism of the torus is isomorphic to some hyperbolic automorphism of the same torus; that is, the corresponding cascades are isomorphic.

Structural stability. A differentiable flow $(\mathbf{R}, X, \pi)$ on a compact manifold X is "structurally stable" if the phase portrait (or orbit structure) is insensitive to small perturbations in the equations (vector field) governing the flow. It is possible to give the set of all C^1 vector fields, which has a natural vector space structure, a norm (the C^1 norm) with respect to which it is a Banach space. A C^1 vector field F is structurally stable if there exists $\epsilon > 0$ such that whenever

G is a C^1 vector field with $\|F - G\| < \epsilon$, G and F generate flows with isomorphic orbit structure; that is, there is a homeomorphism of X which takes G orbits into F orbits, although "time" may not be preserved. The equations governing any physical system can in practice only be approximated, but when the "true" system is structurally stable, a sufficiently good approximation to it will have the same orbit structure. One can also define structural stability for diffeomorphisms (which also have a natural C^1 topology).

Anosov's theorem. Every Anosov flow or diffeomorphism is structurally stable. This implies the geodesic flow on a compact, negatively curved riemannian manifold is structurally stable; similarly, a hyperbolic automorphism of the torus is structurally stable.

Symbolic dynamics. Let Λ_r be a set with $r > 1$ elements, say $\Lambda_r = \{1,2,\ldots,r\}$. Provided with a natural metric, the set X_r of all bisequences $x = \{x_n\}^{\infty}_{n=-\infty}$ of elements from Λ_r is a compact metric space, and the left shift, σ, $(\sigma x)_n = x_{n+1}$, a homeomorphism of this space. This defines a cascade called the shift (on r symbols), written (σ, X_r) [rather than $(\mathbf{Z}, X_r, \pi)$]. Symbolic dynamics is the study of the shift, for its own sake and for application to other systems.

J. Hadamard (1898) set up a correspondence between certain symbolic sequences and geodesics on a negatively curved surface. Utilizing this correspondence, M. Morse (1922) proved the corresponding geodesic flow has minimal sets not consisting of a single closed orbit. (A minimal set is a nonempty closed invariant set containing no proper subset with the same property.) Generalizing this approach, Y. G. Sinai and R. Bowen have constructed an elaborate symbolic dynamics for flows and diffeomorphisms which are suitably "hyperbolic" (for example, Anosov). The theory has numerous applications, among them (a) a generalization of Morse's result to Anosov flows with smooth invariant measure, in particular the geodesic flow on a compact negatively curved manifold; (b) the statement that every minimal set for an Anosov flow is one-dimensional (that is, "small"); and (c) remarkable asymptotic formulas for the distribution of periodic orbits in Anosov flows (with application to the distribution of closed geodesics on a negatively curved manifold).

Ergodic theory. Ergodic theory is the study of measure-preserving transformations. Let X be a set, $\mathscr{B}$ a σ-field of subsets of X, and μ a measure on $\mathscr{B}$. A transformation T of X to itself is measure-preserving if $T^{-1}B \in \mathscr{B}$ and $\mu(T^{-1}B) = \mu(B)$ for all $B \in \mathscr{B}$. If T is one-to-one onto, and if both T and T^{-1} are measure-preserving, T is said to be an automorphism of $(X, \mathscr{B}, \mu)$. If G is a topological group, a representation of G on (X,B,μ) is a homomorphism $(g \to T^g)$ from G to the automorphism group of $(X, \mathscr{B}, \mu)$ with the property that $T^g x$ is measurable on the product space $G \times X$. Representations T^g and $T^g{}_0$ on $(X, \mathscr{B}, \mu)$ and $X_0, \mathscr{B}_0, \mu_0)$ are isomorphic if there exists a one-to-one equivariant map φ of X onto X_0 such that $\mu_0(\varphi B) = \mu(B)$ for all $B \in \mathscr{B}$, $\mu(\varphi^{-1}B_0) = \mu_0(B_0)$ for all $B_0 \in \mathscr{B}_0$.

Birkhoff ergodic theorem. Let T^t be a representation of $\mathbf{R}$ on a probability space $(X, \mathscr{B}, \mu)$. Birkhoff's ergodic theorem says that if f is integrable on X, then for almost all $x \in X$ the limit given by Eq. (13) exists;

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f(T^t x)\, dt = \bar{f}(x) \tag{13}$$

furthermore, $\bar{f}$ is integrable, and Eq. (14) holds. A

$$\int_X f(x)\mu(dx) = \int_X \bar{f}(x)\mu(dx) \tag{14}$$

similar statement holds for a representation of $\mathbf{Z}$, the integral in Eq. (13) being replaced by a sum

$$\frac{1}{T} \sum_{t=0}^{T-1} f(T^t x)$$

Ergodicity and mixing. A group of automorphisms of a probability space is ergodic (or metrically transitive) if every measurable invariant set has measure 0 or 1. In the presence of ergodicity the function $\bar{f}$ in Eq. (13) is almost everywhere equal to

$$\int_X f(x)\mu(dx)$$

and hence "time averages equal space averages." Because of this interpretation, ergodicity is important for flows with invariant measure which are unstable, that is, sensitive to perturbations of the initial conditions (for example, Anosov flows).

An equivalent formulation of ergodicity (for $\mathbf{R}$) is the condition that Eq. (15) is satisfied for all A,B

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mu(A \cap T^t B)\, dt = \mu(A)\, \mu(B) \tag{15}$$

$\in \mathscr{B}$. The condition for mixing is an "unaveraged" analog of Eq. (15) given by Eq. (16). Between the

$$\lim_{t\to\infty} \mu(A \cap T^t B) = \mu(A)\, \mu(B) \tag{16}$$

two notions is the notion of weak mixing, defined by Eq. (17). If A and $T^t B$ are independent sets, then by

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \left|\mu(A \cap T^t B) - \mu(A)\mu(B)\right| dt = 0 \tag{17}$$

definition $\mu(A \cap T^t B) = \mu(A)\mu(T^t B) = \mu(A)\mu(B)$. Ergodicity and the two kinds of mixing thus express, to varying degrees, the notion of asymptotic independence. D. V. Anosov and Sinai have proved an Anosov diffeomorphism with smooth invariant measure is mixing; an Anosov flow with smooth invariant measure is ergodic and, if it is weak-mixing, mixing; the geodesic flow on a compact, negatively curved riemannian manifold is mixing (in dimension two, a classical theorem of E. Hopf and G. A. Hedlund). Stronger statements have since been made.

Entropy. Let T be a measure-preserving transformation of a probability space (X, B, μ). Define $\eta(t)$, $0 \le t \le 1$, to be 0 if $t = 0$ and $-t \log t$ otherwise. If $P = \{A_1, \ldots, A_r\}$ is a partition of X into pairwise disjoint measurable sets, let P_N, $N \ge 1$, be the partition whose elements are sets of the form

$$A_{j_0} \cap T^{-1}A_{j_1} \cap \cdots \cap T^{-(N-1)}A_{j_{N-1}}$$

$$1 \le j_0, \ldots, j_{N-1} \le r$$

Define $$H(P_N) = \sum_{A \in P_N} \eta(\mu(A))$$

The limit $$\lim_{N\to\infty} \frac{H(P_N)}{N} = h(T, P)$$

exists and is called the entropy of T with respect to P. One has $0 \le h(T, P) \le \log r$; $h(T, P) = 0$ if and only if P is measurable with respect to the σ-algebra generated by $T^{-1}P, T^{-2}P, \ldots$. Thus entropy is a measure of "determinism." Finally $h(T)$, the entropy of T, is the supremum of $h(T, P)$ over all P. In some cases $h(T) = \infty$; but, for example, if T is a diffeomorphism of a compact differentiable manifold X, and if μ is smooth, then Kushnirenko's theorem states that $h(T) < \infty$. *See* ENTROPY.

Ornstein's theorem. If $p_1, \ldots, p_r$ are probabilities, it is possible to define a Borel probability measure μ on the space X_r of symbol sequences which governs the process of selecting elements from Λ_r independently and according to the given probabilities. The shift σ preserves μ, and is called a Bernoulli shift. A. N. Kolmogerov and Sinai showed that

$$h(\sigma) = \sum_{j=1}^{r} \eta(p_j)$$

since entropy is an isomorphism invariant, this settled the long-standing question of whether the two shift with probabilities ($^1/_2$, $^1/_2$) is isomorphic to the three shift with probabilities ($^1/_3$, $^1/_3$, $^1/_3$). (They are not isomorphic because $\log 2 \neq \log 3$.) D. Ornstein (1969) established the deep result that entropy is a complete invariant among the Bernoulli shifts; two Bernoulli shifts with the same entropy are isomorphic. *See* PROBABILITY.

Bernoulli shifts as models. Ornstein's results have led to the measure theoretic classification of a number of cascades and flows. N. Friedman, Ornstein, Sinai, and R. Azencott showed that a C^2 Anosov diffeomorphism with smooth invariant measure is Bernoulli; Y. Katznelson showed that an ergodic automorphism of the n-dimensional torus is Bernoulli; Ornstein and B. Weiss showed that the geodesic flow on a compact riemannian manifold of negative curvature is Bernoulli (where a flow is defined to be Bernoulli if each of its time t maps is Bernoulli); and M. Ratner showed that an Anosov flow with smooth invariant measure is Bernoulli.

Minimal transformation groups. A transformation group is minimal if every orbit is dense in the phase space. (The concept is due to Birkhoff.) It is of interest to classify minimal transformation groups, one reason being that in an arbitrary compact phase space there exist nonempty closed invariant sets which are minimal. In what follows, X is a compact metric space with metric $d(x, y)$, (G, X, π) and $\pi(g, x)$ will be written (G, X) and gx, and transformation groups (arbitrary G) will be called flows.

Equicontinuous flows. A flow is equicontinuous if for each $\epsilon > 0$ there exists $\delta > 0$ such that $d(gx, gy) < \epsilon$, for all $g \in G$, whenever $d(x, y) < \delta$.

$$D(x, y) = \sup_g d(gx, gy)$$

is a compatible G invariant metric for X when (G, X) is equicontinuous. Since the group of isometries of (X, D) is, with a natural topology, compact, G maps homomorphically into a compact metric group acting on X. If (G, X) is minimal, the closure of the image of G in this group acts transitively on X. The study of minimal equicontinuous flows is therefore essentially the study of transitive compact point actions.

Distal flows. A point $x \in X$ is distal for (G, X) if

$$\inf_g d(gx,gy) = 0$$

implies $y = x$. This flow is distal if every point is distal. Every equicontinuous flow is distal, but there exist minimal distal flows which are not equicontinuous. (However, R. Ellis has shown that if G is finitely generated and X totally disconnected, distality implies equicontinuity. For example, a minimal distal subset of the shift is finite.) Minimal distal flows have been characterized by H. Furstenberg in terms of "isometric extensions." A homomorphism $\varphi:(G, X) \to (G, X')$ defines an isometric extension if there is a continuous function R on $\Delta = \{(x, y) \in X \times X | \varphi(x) = \varphi(y)\}$ such that (a) $R(gx, gy) = R(x, y)$ for all $g \in G$, and (b) for each $x' \in X'$, R defines a metric on $\varphi^{-1}x'$. Furstenberg's structure theorem states that a minimal distal flow can be "built up" by a transfinite sequence of isometric extensions and "inverse limits" beginning with the trivial flow (one-point phase space).

Point distal flows. If there exists a distal point with dense orbit, the flow (G, X) is point-distal. Point-distal flows are minimal and not necessarily distal. For example, there exist nonequicontinuous, hence nondistal, cascades with totally disconnected phase space. Furstenberg's structure theorem is generalized to point-distal flows using the notion of almost one-to-one extension, a homomorphism $\varphi:(G, X) \to (G, X')$ such that for some $x' \in X'$, $\varphi^{-1}x'$ is a single point. W. Veech and Ellis have shown that every point-distal flow has an almost one-to-one extension which can be built up from the one-point flow by a transfinite sequence of isometric extensions, almost one-to-one extensions, and inverse limits. A general structure theory for minimal flows has been established.

William A. Veech

Bibliography. E. Akin, *The General Topology of Dynamical Systems*, 1993; J. Alexander (ed.), *Dynamical Systems*, 1988; V. I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics*, 1968, reprint 1989; J. DeVries, *Elements of Topological Dynamics*, 1993; R. Mane, *Ergodic Theory and Differentiable Dynamics*, 1987; D. Ruelle, *Elements of Differentiable Dynamics and Bifurcation Theory*, 1989; C. L. Siegel and J. K. Moser, *Lectures on Celestial Mechanics*, 2d ed., 1994.

Topology

The branch of mathematics that studies the qualitative properties of spaces, as opposed to the more delicate and refined geometric or analytic properties. While there are earlier results that belong to the field, the beginning of the subject as a separate branch of mathematics dates to the work of H. Poincaré during 1895–1904. The ideas and results of topology have a central place in mathematics, with connections to almost all the other areas of the subject.

The difference between topological and geometric properties is illustrated by the example of a space with three separate pieces. The exact shapes of the pieces constitute a geometric property of the space, and the study of these shapes is in the domain of differential geometry, but the fact that the space has three separate pieces is a qualitative or topological property. As another example, if a round sphere is deformed to be pear-shaped (or even more irregularly shaped, like the surface of the Earth), then the geometric notions of distance, straight line, and angle are changed, but the topological properties of the surface are left unchanged. However, if a handle is added by cutting two holes in the sphere and connecting them by a curved pipe, then the topology of the surface is changed (**Fig. 1**).

Development. Four major areas of topology are algebraic topology, homotopy theory, general topology, and manifold theory. Algebraic topology, the first area of modern topology to be developed, is concerned with associating algebraic invariants to geometric spaces in order to measure higher-dimensional analogs of the number of pieces of a space or the number of handles of a surface. Algebraic topology has tremendous influence on other branches of mathematics, both direct (application of the invariants of algebraic topology to problems from other areas of mathematics and physics) and indirect (application in other contexts of ideas arising from algebraic topology).

As algebraic topology developed, it became clear that if one function could be continuously deformed to another (that is, if they were homotopic), then these two functions behaved in the same way as far as the invariants of algebraic topology were concerned.

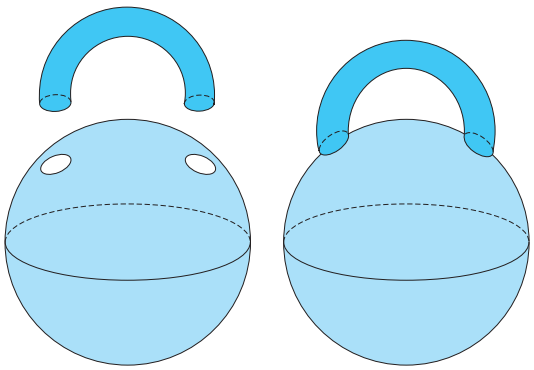


Fig. 1. Process of adding a handle to a 2-sphere. (*a*) Cutting of holes in sphere. (*b*) Connecting of holes by a curved pipe.

This led naturally to the study of invariants that remain unchanged as the maps are deformed by homotopies, that is, homotopy invariants. This study, which is an offshoot of algebraic topology, is called homotopy theory. Some of the most interesting homotopy invariants are the higher homotopy groups. These proved extremely difficult to compute, even for spaces as simple as the sphere, and are the subject of much investigation.

Early in the development of topology it was realized that the foundations of the subject needed attention. General or point-set topology studies the relationships between the basic topological properties that spaces may possess.

Before abstract topological spaces were defined, there were numerous examples of spaces arising from geometric and analytic problems. The most important of these is a class of spaces known as manifolds. Both because of their ubiquitous appearance throughout mathematics and because they possess extraordinarily rich topological properties, manifolds became one of the central objects of study in topology. The basic theme in manifold theory is to find sufficient algebraic invariants to classify, that is, to list comprehensively, all manifolds, and to give methods for evaluating these invariants in geometric cases. *See* MANIFOLD (MATHEMATICS).

General topology. The coordinate space of calculus, $\mathbf{R}^n$, is the space of all ordered n-tuples $(x_1, \ldots, x_n)$ of real numbers. On this space there is the pythagorean (or euclidean) distance function. If $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are points of $\mathbf{R}^n$, then the distance from x to y is given by Eq. (1).

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} \qquad (1)$$

This distance function indicates how the points of $\mathbf{R}^n$ are related to each other, and as discussed below, it gives $\mathbf{R}^n$ the structure of a topological space. An open subset of $\mathbf{R}^n$ is a subset $U \subset \mathbf{R}^n$ with the property that for each $x \in U$ there is a real number $\delta > 0$ such that all points y for which $d(x, y) < \delta$ are also contained in U. (The notation $x \in U$ means that x is a member of U.) The intersection of any finite number of open subsets of $\mathbf{R}^n$ is itself an open subset, and the union of any collection of open subsets of $\mathbf{R}^n$ is an open subset. *See* SET THEORY.

Topological spaces. This situation is vastly generalized by the notion of an abstract topological space. A topological space, or a space for short, is a pair consisting of a set S, called the underlying set of the space, and a collection of subsets $\{U_\alpha\}_{\alpha \in A}$ called the open subsets of the space. The collection of open subsets is required to satisfy four axioms:

1. The empty subset is an open subset of the space.
2. The entire set S is an open subset of the space.
3. If $U_1, \ldots, U_n$ is a finite collection of open subsets of the space, then the intersection $U_1 \cap \cdots \cap U_n$ is an open subset of the space.
4. If $\{U_\beta\}_{\beta \in B}$ is any collection of open subsets of the space, then the union $\cup_{\beta \in B} U_\beta$ is an open subset of the space.

The open subsets of a space X that contain a given point x are thought of as containing all the points of X close to x. Thus, a topology also describes how the points of a space cohere. The collection of open sets in $\mathbf{R}^n$ as defined above form a topology on $\mathbf{R}^n$.

A subspace A of X consists of a subset T of the set S underlying X with the induced topology; that is, a subset of A is open in A if and only if it is the intersection of an open set of X with T. The unit interval I is the subspace of $\mathbf{R}^1$ given by Eq. (2).

$$I = \{t \in \mathbf{R}^1 \mid 0 \leq t \leq 1\} \tag{2}$$

If X and Y are spaces, then a new space, denoted $X \times Y$, can be formed. The underlying set is simply the set of ordered pairs $\{(x,y) | x \in X, y \in Y\}$. The topology, called the product topology, is as follows: Any subset of $X \times Y$ of the form $U \times V$, where U is an open subset of X, and V is an open subset of Y, is open in $X \times Y$. More generally, any union of such sets is open in $X \times Y$.

Continuous maps and homeomorphisms. A set function f: $X \to Y$ is said to be continuous if, for every open subset U of Y, the subset of points of X that map by f to U, denoted $f^{-1}(U)$, is an open subset of X. Continuous functions are called maps. They preserve the notions of coherence in the spaces X and Y in the sense that close points in X are sent by f to close points in Y. This notion generalizes the usual notion of continuity from calculus for maps between subspaces of the real line. *See* CALCULUS.

A homeomorphism from X to Y is a continuous map f from X to Y, which is a bijection on the underlying sets and whose inverse function is also continuous. General topology is the study of topological spaces and continuous maps between them. In general topology, two spaces are considered equivalent if there is a homeomorphism between them.

An example of homeomorphic spaces is the unit circle $S^1 = \{(x, y) \in \mathbf{R}^2 | x^2 + y^2 = 1\}$ and the ellipse $\{(x, y) \in \mathbf{R}^2 | x^2 + (y^2/4) = 1\}$. The map $(x, y) \to (x, 2y)$ sends the circle to the ellipse. The map $(x, y) \to (x, y/2)$ sends the ellipse to the circle. Since these maps are continuous and inverses of each other, they are homeomorphisms. Hence, the circle and the ellipse are homeomorphic. The boundary of the unit square in the plane is made up of four straight lines meeting at right angles. It is homeomorphic to the unit circle. On the other hand, there is a continuous map from the half-open unit interval $J = \{t \in \mathbf{R}^1 | 0 \leq t < 1\}$ to S^1 given by $t \to (\cos(2\pi t), \sin(2\pi t))$. Though this map is a bijection, its inverse is not continuous since the inverse function sends (1,0) to 0 and sends points close to (1,0) with negative y coordinates to points near 1 instead of near 0. In fact, the circle and the half-open interval are not homeomorphic.

These examples indicate why topology is often described as the study of properties of spaces that remain constant under "deformation and bending but without tearing"; these words roughly describe a homeomorphism. Geometric properties such as size, curvature, straightness, and angles are completely lost when only properties invariant under homeomorphism are considered.

The circle and the 2-sphere $S^2 = \{(x, y, z) \in \mathbf{R}^3 | x^2 + y^2 + z^2 = 1\}$ are not homeomorphic. This result follows from one of the most celebrated theorems in general topology, called the invariance of domain theorem. It states that if U and V are nonempty open subsets of $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and if U and V are homeomorphic, then $n = m$.

Spaces with special properties. Most topological spaces that arise from other parts of mathematics have many special properties. For example, a topological space is said to be Hausdorff if for every pair of distinct points x and y in the space there are open subsets U and V such that $x \in U$, $y \in V$, and $U \cap V =$ (that is, the intersection of U and V is empty). The space $\mathbf{R}^n$ and all its subspaces are Hausdorff. Another important property is compactness. A topological space is compact if for any collection of open subsets whose union is the entire space there is a finite subcollection whose union is also the entire space. An important theorem says that a subspace X of $\mathbf{R}^n$ is compact if and only if it satisfies the following two conditions:

1. X is bounded in the sense that there is a real number $\mathbf{R} > 0$ such that all points of X have distance at most $\mathbf{R}$ from the origin.
2. X is closed in the sense that $\mathbf{R}^n - X$ is an open subspace of $\mathbf{R}^n$.

Much of general topology focuses on compact Hausdorff spaces.

Another important class of spaces is the metric spaces. A metric space is a set S equipped with a distance function (or metric) d:$S \times S \to \mathbf{R}$ satisfying the following three axioms:

1. For all $x, y \in S$, $d(x, y) \geq 0$.
2. $d(x, y) = 0$ if and only if $x = y$.
3. $d(x, y) + d(y, z) \geq d(x, z)$.

The pythagorean distance in $\mathbf{R}^n$ satisfies these axioms. If (S, d) is a metric space, then it has an induced topology defined analogously to the topology on $\mathbf{R}^n$. Many theorems of general topology concern what sorts of topological properties are sufficient to ensure that a space admits a metric inducing the given topology.

Algebraic topology. Algebraic topology is the study of algebraic invariants associated to topological spaces. An algebraic invariant of a space is an algebraic object associated to the space that remains unchanged if the space is replaced by a homeomorphic space. By an algebraic object is meant either an algebraic structure, such as a group, ring, or field, or an element of an algebraic structure.

Homology. This is the first and most important invariant of algebraic topology. Only a vague description of the main idea behind homology will be given. The homology groups $H_i(X)$, $i = 1, 2, \ldots$, of a space X are abelian groups associated to the space that measure the higher-dimensional analog of the number of pieces, that is, the number of higher-dimensional holes in X. The nth-homology group $H_n(X)$ is constructed out of the n-dimensional cycles (a cycle is a geometric compact object with no boundary) in X that go around the n-dimensional holes. In order to form the homology group, it is necessary to divide out by (that is, to set equal to zero) those cycles that

do not surround any hole of X and hence are the boundary of a compact $(n+1)$-dimensional geometric object in X. *See* GROUP THEORY.

While the homology groups are difficult to define and understand, once defined they can be computed for any space that is presented in a reasonable geometric fashion. For example, the unit n-sphere for $n \geq 1$, which is the subset of $\mathbf{R}^{n+1}$ given by Eq. (3),

$$S^n = \{x \in \mathbf{R}^{n+1} \mid d(x,0) = 1\} \tag{3}$$

has a single n-dimensional hole, which is surrounded by the entire sphere, and no other holes of any dimension. Thus, the homology groups of $H_i(S^n)$ are zero for all $i > 0$ except for $i = n$, and $H_n(S^n)$ has a single generator corresponding to the sphere itself. In particular, the n-sphere and the m-sphere are not homeomorphic if $n \neq m$. This homological computation can be used to give another proof of the invariance of domain theorem discussed above. This was one of the first applications of homology theory and indicated its power.

Euler characteristic. This is one of the most frequently used invariants of algebraic topology. A space X has an Euler characteristic when it has only finitely many nonzero homology groups and each of these is finitely generated. The Euler characteristic $\chi(X)$ is given by Eq. (4), where $\beta_i(X)$ is the ith-Betti number of X, which is by definition the rank of the ith-homology group of X.

$$\chi(X) = \sum_i (-1)\beta_i(X) \tag{4}$$

Fundamental group. This invariant of algebraic topology, also called the Poincaré group, formed from the set of loops in X that begin and end at a preassigned point $x_0 \in X$, or are based at x_0. Two such loops γ_1 and γ_2 can be composed by simply tracing γ_1 followed by γ_2 to give a new loop based at x_0. Two loops are said to be equivalent if the first can be continuously deformed through a family of loops based at x_0 to the second. The composition yields a group structure on the set of equivalence classes of loops based at x_0. This group is the fundamental group of X based at x_0 and is denoted $\pi_1(X, x_0)$. A space is said to be simply connected if its fundamental group is the trivial group.

An important result computes the fundamental groups of the spheres: The sphere S^n is simply connected if $n > 1$; and the fundamental group of the circle is isomorphic to the group of integers and is generated by the loop going once around the circle.

Applications. In applying algebraic topology to problems from other areas of mathematics, it is necessary to compute the values of the invariants. To a large degree this is feasible for the homology groups and the fundamental group. Therefore, these invariants play very important roles in differential and algebraic geometry. *See* ALGEBRAIC GEOMETRY; DIFFERENTIAL GEOMETRY.

An important theorem for computing an invariant of algebraic topology by geometry is the Gauss-Bonnet theorem. It states that, if S is a compact riemannian surface, then the Euler characteristic $\chi(S)$ is given by Eq. (5), where $K(x)$ is the gaussian curvature of the surface and dvol is the volume element. *See* RIEMANNIAN GEOMETRY.

$$\chi(X) = \frac{1}{2\pi} \int_S K(x)\text{dvol} \tag{5}$$

The ideas of algebraic topology have had profound effect even in situations that are not directly topological. The homology of groups is an important tool in Galois theory and representation theory. Much progress in number theory has revolved around finding reasonable analogs to the invariants of algebraic topology. *See* NUMBER THEORY.

Homotopy theory. In algebraic topology the deformation of continuous maps is extremely important. The defnition of homotopy formalizes this notion. A homotopy of maps from X to Y is a continuous map given by expression (6), where I is the unit interval.

$$H: X \times I \to Y \tag{6}$$

Setting h_t equal to the restriction of H to the slice $X \times \{t\}$ gives a family of continuous maps from X to Y. The homotopy H is said to be a homotopy from h_0 to h_1, and h_0 and h_1 are said to be homotopic. The set of homotopy classes of maps from X to Y can then be introduced. The task of homotopy theory is to compute in some effective way the set of homotopy classes of maps from one space to another.

Not only do the invariants of algebraic topology associate an algebraic structure to each topological space, but they also associate to a continuous map from X to Y an algebraic map (a homeomorphism) from the structure associated to X to that associated to Y. For example, for any map $f: X \to Y$ there is an induced homeomorphism $f_*: H_n(X) \to H_n(Y)$ called the map induced on homology. Such invariants are called homotopy invariants if any time f is homotopic to g, the algebraic maps induced by f and g on the algebraic structures are equal. Both homology and the fundamental group are homotopy invariants.

One application of these ideas is the Lefschetz fixed-point formula. For any space X that belongs to a large class of compact spaces (including all manifolds, as defined below), any map $f: X \to X$ has a fixed point if the Lefschetz number $\lambda(f)$ defined by Eq. (7) is nonzero. Here, $T_i(f)$ is the trace of the map $f_*: H_i(X,\mathbf{Q}) \to H_i(X,\mathbf{Q})$ on homology with rational coefficients $\mathbf{Q}$. [The trace is defined as follows. A basis is chosen for the vector space $H_i(X,\mathbf{Q})$. In this basis the linear map f^* is given by a square matrix, and its trace is the sum of the diagonal entries. The trace does not depend on the basis chosen.] The Lefschetz fixed-point formula implies that any map of the disk to itself has a fixed point and that any map of the n-sphere to itself that is homotopic to the identity has a fixed point if n is even.

$$\lambda(f) = \sum (-1)^i T_i(f) \tag{7}$$

Other basic homotopy invariants from algebraic topology are the higher homotopy groups. The set of homotopy classes of maps from S^n to a space X, where the maps are required to send the point

$(1, 0, \ldots, 0) \in S^n$ to a base point $x_0 \in X$, is denoted $\pi_n(X, x_0)$. The fact that collapsing the equator of S^n to a point yields a space homeomorphic to two n-spheres glued together at a single point leads to a law of composition on this set that makes it a group. For $n = 1$, this group is simply another presentation of the fundamental group. For $n > 1$, these groups are called the higher homotopy groups. These are homotopy invariants. One of the first and most important theorems about these groups states that the homotopy group $\pi_n(X, x_0)$ is an abelian group for any $n \geq 2$ and any space X.

It is natural to try to compute the homotopy groups of spheres. The following results hold for a given point x_0 in S^n:

1. $\pi_k(S^n, x_0) = 0$ if $k < n$.
2. $\pi_n(S^n, x_0)$ is isomorphic to the group of integers.
3. $\pi_k(S^1, x_0) = 0$ for all $k \neq 1$.

These results show a close relationship between homology and homotopy. Surprisingly, it was shown that $\pi_3(S^2, x_0)$ is nontrivial; in fact it is isomorphic to the group of integers. This result established the computation of the homotopy groups of spheres as a major area of homotopy theory. Another important result, obtained in 1950, states that $\pi_k(S^n, x_0)$ is a finite group unless $k = n$ or n is even and $k = 2n - 1$; in the exceptional cases, $\pi_k(S^n, x_0)$ is isomorphic to the group of integers plus a finite group. It is also known that for every $n \geq 2$ the sphere S^n has infinitely many nonzero homotopy groups. While there are many other partial results in determining the homotopy groups of spheres, a solution of this problem still seems very distant.

Manifold theory. The spaces that arise most often outside of topology are manifolds. They have always been a focus of much of the work in topology. A topological n-dimensional manifold (or topological n-manifold) is a Hausdorff topological space that is a union of countably many open sets $\{U_i\}$, each of which is homeomorphic to an open subset of $\mathbf{R}^n$. The homeomorphisms $\varphi_i: U_i \to V_i \subset \mathbf{R}^n$ give local coordinates $(x_1, \ldots, x_n)$ on U_i. Thus, a manifold is covered by open sets that have local coordinates like the coordinates of $\mathbf{R}^n$. These pairs (U_i, φ_i) are called local coordinate charts. An equivalence between topological manifolds is homeomorphism.

Examples of manifolds. Euclidean space $\mathbf{R}^n$ is of course an n-manifold. The 2-sphere $S^2 \subset \mathbf{R}^3$ can be shown to be a 2-manifold, or surface, as follows: A stereographic projection from the north pole x of the sphere to the plane tangent to the south pole gives a homeomorphism from $S^2 - \{x\}$ to $\mathbf{R}^2$ (**Fig. 2**). Similarly, a stereographic projection from the south pole y gives a homeomorphism from $S^2 - \{y\}$ to $\mathbf{R}^2$. Thus, S^2 can be covered with two open sets, each of which is homeomorphic to $\mathbf{R}^2$. This proves that S^2 is a 2-manifold. Similarly, stereographic projections in higher dimensions can be used to show that S^n is an n-manifold.

The surface of a doughnut is an example of a 2-manifold; it is called the torus. More generally, the k-holed torus, also called the surface of genus k, is a 2-manifold (**Fig. 3**).

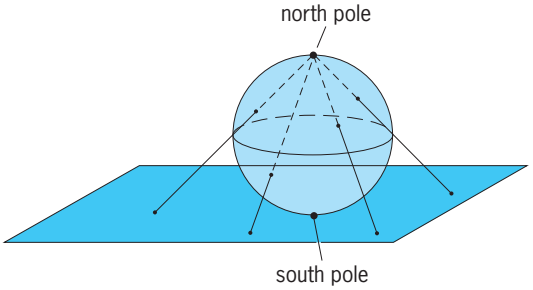


Fig. 2. Stereographic projection of a sphere from the north pole to the plane tangent to the south pole.

Smooth manifolds. The manifolds that arise from geometry and analysis carry an extra structure called a smooth (or differential) structure. A smooth structure on a topological manifold M is a collection of local coordinate charts, $\varphi_i: U_i \to V_i$, such that for every i and j the overlap function, which is the composition (8), is a smooth map (that is, an infinitely

$$\varphi_i(U_i \cap U_j) \xrightarrow{\varphi_i^{-1}} U_i \cap U_j \xrightarrow{\varphi_j} \varphi_j(U_i \cap U_j) \quad (8)$$

differentiable map). Technically speaking, two such systems of smoothly overlapping coordinate systems define the same smooth structure if their union is a smooth structure. A smooth manifold is a topological manifold with a smooth structure. The coordinates that were given above on S^2 have smooth overlap, so that S^2 has a smooth structure.

If M is a smooth manifold, then there is a well-defined notion of a smooth real-valued function on M, and of differentiation of such smooth functions. More generally, there is a natural notion of a smooth (or infinitely differentiable) map between smooth manifolds. A diffeomorphism or isomorphism between smooth manifolds is a smooth homeomorphism whose inverse is also a smooth map. Differential topology is the study of smooth structures on manifolds and smooth maps between smooth manifolds. In differential topology two manifolds are considered equivalent if there is a diffeomorphism between them.

Homology groups of manifolds. Some of the first results in algebraic topology concern the homology groups of compact manifolds. Poincaré proved a fundamental result, called Poincaré duality, which says, roughly, that the homology groups of a compact n-manifold in dimensions greater than $n/2$ can be computed from the groups in dimensions less than or equal to $n/2$.

The homology groups of compact surfaces have been completely determined: The first homology of the surface of genus g is a free abelian group of rank $2g$. In particular, the information about the genus

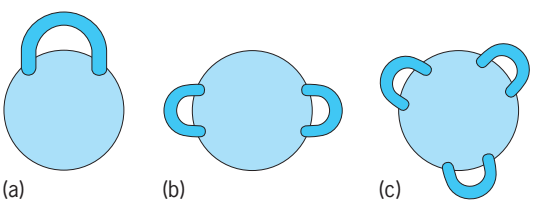


Fig. 3. Examples of 2-manifolds. (*a*) Surface of genus 1. (*b*) Surface of genus 2. (*c*) Surface of genus 3.

(or number of handles) can be recovered from the homology.

De Rham's theorem computes the homology groups of a smooth manifold from the calculus of the manifold. Many consequences flow from this result, implying that certain analytic or geometric quantities can be viewed as homology classes.

Classification of manifolds. One of the main problems in manifold theory is the classification of all manifolds (or smooth manifolds) up to homeomorphism (or diffeomorphism). There is also the auxiliary question of the difference (if any) between the two classification schemes; that is, whether every topological manifold admits a smooth structure, and, if so, how many nondiffeomorphic structures.

Results for surfaces were obtained during the nineteenth century. They can be formulated as follows: (1) Every topological surface admits a smooth structure and only one up to diffeomorphism; and (2) if S and S' are compact surfaces, then S and S' are homeomorphic if and only if their homology groups are isomorphic.

Based on this result, Poincaré asked whether or not the homology of a compact 3-manifold determined it up to homeomorphism, in particular, in the case of the 3-sphere. He answered this question in the negative by defining a new invariant, the fundamental group, and finding a 3-manifold Σ with the homology of S^3 but with a nontrivial fundamental group. Since the fundamental group of S^3 is the trivial group, Σ is not homeomorphic to S^3.

Then Poincaré refined his question to what is known as the Poincaré conjecture: if a 3-manifold has the homology of the 3-sphere and is simply connected, then is it homeomorphic to S^3? Despite much effort, this problem is unresolved.

S. Smale realized that higher-dimensional manifolds could be easier to study than 3-manifolds, and in 1961 he answered in the affirmative the generalization of Poincaré conjecture in all dimensions greater than or equal to 5. This was followed by considerable activity, starting with the classification of smooth structures on S^n for $n \geq 5$. While it has long been known that a complete classification is logically impossible in dimensions at least 4, by 1975 the classification for manifolds of dimension at least 5 had been reduced to problems in algebra and homotopy theory.

There are strong results about when a topological manifold admits a smooth structure. The answer is "not always," but the obstructions are well understood and are homotopy-theoretic in nature. A compact manifold of dimension at least 5 has only finitely many distinct smooth structures, and an upper bound for the number is given in terms of the algebraic topological invariants of the manifold.

By 1975 it appeared that the techniques applied in the higher dimensions could not be made to work in dimensions 3 and 4 because there was not enough room to maneuver. W. Thurston then introduced a vast program to study 3-manifolds. His idea was to cut up any 3-manifold into pieces that have especially nice metrics, with eight possible types. The most prevalent and interesting is the hyperbolic metric. Thurston's conjecture about the existence of these nice metrics vastly generalizes the Poincaré conjecture, which concerns the case of spherical metrics. Thurston has established his conjecture in many cases (but not, unfortunately, in the case of simply connected manifolds), lending much evidence that it is true.

In 1980 there were two major advances concerning 4-dimensional manifolds. M. Freedman showed that the high-dimensional techniques can be made to work for simply connected topological 4-manifolds. As a result, he obtained a complete classification of these manifolds, including an affirmative answer to the 4-dimensional version of Poincaré's conjecture. This completes the answer to Poincaré's conjecture in all dimensions except the one in which he originally made it.

S. Donaldson defined new algebraic invariants for smooth 4-manifolds and used them to show that some of the topological manifolds that Freedman had constructed did not admit smooth structures. He also used the invariants to show that certain well-known smooth manifolds, which were homeomorphic by Freedman's result, were not diffeomorphic. It is now known that many topological 4-manifolds admit infinitely many distinct smooth structures, whereas in all other dimensions the corresponding number is finite. In particular, $\mathbf{R}^4$ admits uncountably many distinct smooth structures, where in all other dimensions $\mathbf{R}^n$ has a unique smooth structure.

Donaldson's invariants are constructed using the moduli space of classical solutions (solutions to the Euler-Lagrange equations) of an action functional for a gauge theory. This action functional plays an important role in high-energy theoretical (quantum) physics. Other action functions and the properties of the quantum physical systems they describe can also be used to define manifold invariants, especially for three-manifolds and four-manifolds. Examples are the Seiberg-Witten invariants of four-manifolds (which are closely related to Donaldson's invariants), Chern-Simons invariants for three-manifolds, and the so-called quantum cohomology of a symplectic manifold. The deep interplay between mathematics and theoretical physics (particularly quantum field theory and string theory) is one of the primary sources of new insights in geometry and manifold theory. *See* ACTION; GAUGE THEORY; QUANTUM FIELD THEORY; SUPERSTRING THEORY; VARIATIONAL METHODS (PHYSICS). John W. Morgan

Bibliography. G. L. Cain, *An Introduction to General Topology*, 1994; S. Eilenberg and N. Steenrod, *Foundations of Algebraic Topology*, 1952; D. S. Freed and K. K. Uhlenbeck, *Instantons and Four-Manifolds*, 2d ed., 1991; M. J. Greenberg and J. R. Harper, *Algebraic Topology: A First Course*, 1981; M. W. Hirsch, *Differential Topology*, 1976, reprint 1988; J. G. Hocking and G. S. Young, *Topology*, 1961, reprint 1988; R. C. Kirby and L. C. Siebenmann, *Foundational Essays on Topological Manifolds, Smoothings and Triangulations*, 1977; S. Lefschetz, *Topology*, 2d ed., 1990; E. Spanier, *Algebraic Topology*, 1966, reprint 1995.

Torbanite

A variety of coal that resembles a carbonaceous shale in outward appearance. It is fine-grained, brown to black, and tough, and breaks with a conchoidal or subconchoidal fracture. The name torbanite is derived from the initial discovery site of the material in 1850 at Torbane Hill, Linlithgowshire, Scotland. Torbanite is synonymous with boghead coal and is related to cannel coal. It is derived from colonial algae identified with the modern species of *Botryococcus braunii* Kütz and antecedent forms.

Major deposits of torbanite occur in Australia, Tasmania, New Zealand, Scotland, and South Africa. The South African deposit, which is in the Ermelo district of the Transvaal, yields 20–100 gal of oil per ton (80–400 liters per metric ton) on retorting. High-assay torbanite yields paraffinic oil, whereas low-assay material yields asphaltic oil. *See* COAL; SAPROPEL.

Irving A. Breger

Torch

A gas-mixing and burning tool that produces a hot flame for the welding or cutting of metal. The torch usually delivers acetylene and commercially pure oxygen producing a flame temperature of 5000–6000°F (2750–3300°C), sufficient to melt the metal locally. The torch thoroughly mixes the two gases and permits adjustment and regulation of the flame. Acetylene requires 2.5 times its volume of oxygen for complete combustion and, being an endothermic compound of carbon and hydrogen, can produce a higher flame temperature than other fuel gases. *See* ACETYLENE; WELDING AND CUTTING OF MATERIALS.

Torches are of two types: low-pressure and high-pressure. In a low-pressure, or injector, torch, acetylene enters a mixing chamber, where it meets a jet of high-pressure oxygen (see **illus.**). The amount of acetylene drawn into the flame is controlled by the velocity of this oxygen jet. In a high-pressure torch both gases are delivered under pressure. Heat developed at the work is controlled principally by the size of the nozzle or tip fitted to the torch. The larger the tip the greater the required gas pressure. Small flames are used with thin-gage metals; large flames are necessary for thick metal parts.

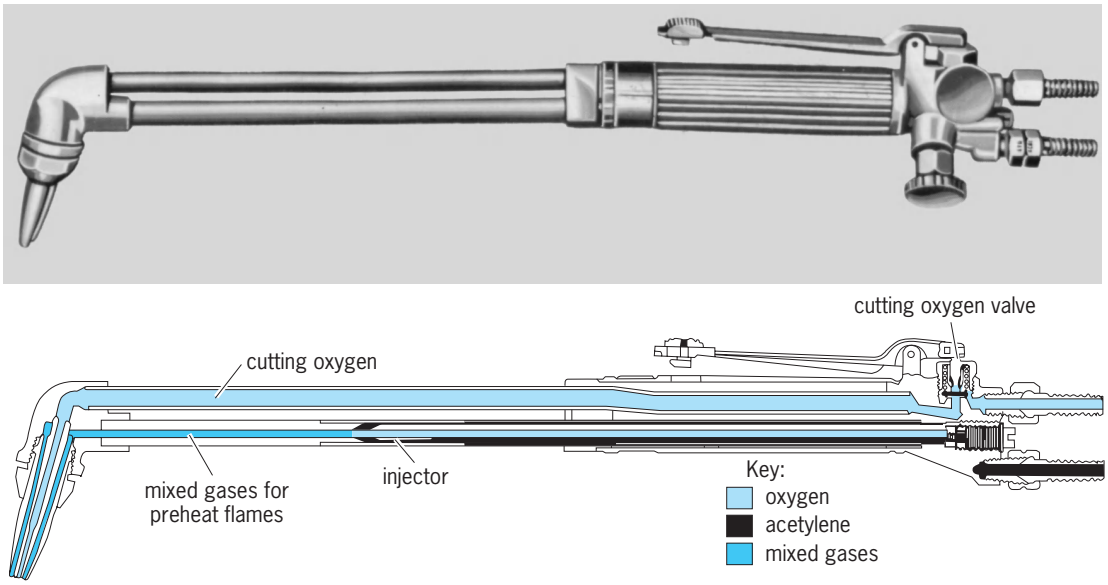
A welding torch mixes the fuel and gas internally and well ahead of the flame. For cutting, the torch delivers an additional jet of pure oxygen to the center of the flame. The oxyacetylene flame produced by the internally mixed gases raises the metal to its ignition temperature. The central oxygen jet oxidizes the metal, the oxide being blown away by the velocity of the gas jet to leave a narrow slit or kerf. In the case of iron, the oxides fuse at a lower temperature than the iron or steel so that the oxides form, melt, and blow away before the adjacent metal fuses. The temperature for the cutting action, once initiated, is maintained by the oxidization of the iron. Intricate shapes are accurately cut in low-carbon steel by torches automatically guided, such precision cutting being called flame machining.

Frank H. Rockett

Tornado

A violently rotating, tall, narrow column of air (that is, a vortex), typically about 300 ft (100 m) in diameter, that extends to the ground from a cumulonimbus cloud. The vast majority of tornadoes rotate cyclonically (counterclockwise in the Northern Hemisphere). Of all atmospheric storms, tornadoes are the most violent. *See* CLOUD; CYCLONE.

Visual appearance. Tornadoes are made visible by a generally sharp-edged, funnel-shaped cloud pendant from the cloud base, and a swirling cloud of dust and debris rising from the ground (**Fig. 1**). The funnel consists of small waterdroplets that form as moist



Low-pressure injector cutting torch. (*Linde Co.*)

Fig. 1. The Cordell, Oklahoma, tornado of May 22, 1981, in its decay stage. (*National Severe Storms Laboratory/ University of Mississippi Tornado Intercept Project*)

air entering the tornado's partial vacuum expands and cools. The condensation funnel may not extend all the way to the ground and may be obscured by dust. Many condensation funnels exist aloft without tangible signs that the vortex is in contact with the ground; these are known as funnel clouds. Tornado funnels assume various forms: a slender smooth rope, a cone (often truncated by the ground), a thick turbulent black cloud on the ground, or multiple funnels (vortices) that revolve around the axis of the overall tornado.

Many tornadoes evolve as follows: The tornado begins outside the precipitation region as a dust whirl on the ground and a short funnel pendant from a wall cloud on the southwest side of the thunderstorm; it intensifies as the funnel lengthens downward, and attains its greatest power as the funnel reaches its greatest width and is almost vertical; then it shrinks and becomes more tilted, and finally becomes contorted and ropelike as it decays. A downdraft and curtain of rain and large hail gradually spiral from the northeast cyclonically around the tornado, which often ends its life in rain. *See* HAIL; PRECIPITATION (METEOROLOGY); THUNDERSTORM.

Parent storm. Most tornadoes and practically all violent ones develop from a larger-scale circulation, the mesocyclone, which is 2–6 mi (3–9 km) in diameter and forms in a particularly virulent variety of thunderstorm, the supercell. The mesocyclone forms first at midaltitudes of the storm and in time develops at low levels and may extend to high altitudes as well. The tornado forms on the southwest side (Northern Hemisphere) of the storm's main updraft, close to the downdraft, after the development of the mesocyclone at low levels. Some supercells develop up to six mesocyclones and tornadoes repeatedly over great distances at roughly 45-min intervals. Tornadoes associated with supercells are generally of the stronger variety and have larger parent cyclones. Hurricanes during and after landfall may spawn numerous tornadoes from small supercells located in their rainbands. *See* HURRICANE.

Damage. Tornadoes are classified as weak, strong or violent, or from F0 to F5 on the Fujita (F) scale of damage intensity. Sixty-two percent of tornadoes are weak (F0 to F1). These tornadoes have maximum windspeeds less than about 110 mi/h (50 m/s) and inflict only minor damage, such as peeling back roofs, overturning mobile homes, and pushing cars into ditches. Thirty-six percent of tornadoes are strong (F2 to F3) with maximum windspeeds estimated to be 110–200 mi/h (50–90 m/s). Strong tornadoes extensively damage the roofs and walls of houses but leave some walls partially standing. They demolish mobile homes, and lift and throw cars. The remaining 2% are violent (F4 to F5), with windspeeds in excess of about 200 mi/h (90 m/s). They level houses to their foundations, strew heavy debris over hundreds of yards, and make missiles out of heavy objects such as roof sections, vehicles, utility poles, and large, nearly empty storage tanks. Engineers have concluded that structural damage is almost always due to wind-associated forces and missiles, not to the sudden reduction in atmospheric pressure. Typically, most debris is ejected out of a tornado along its direction of motion, so that the damage, apart from the characteristic well-defined long narrow path, appears to be from straight-line winds. However, ample signs of the circulation are usually evident in the overall debris distribution of large tornadoes. Multiple-vortex tornadoes that cross fields accumulate soil and vegetation in cycloidal rows, typically 6 in. (15 cm) high and 5 ft (1.5 m) wide, that are visible from the air. The marks are cycloidal because of the circular motion of an individual vortex about the tornado axis combined with the forward motion of the overall tornado. The rows are created by small debris being drawn into the bases of the vortices and left behind in lines of litter in the wakes of the vortices. The individual vortices occur at the locations of the strongest winds, lowest pressure, and most rapid pressure falls. Therefore, multivortex tornadoes passing through cities produce cycloidal swaths of more intense damage. *See* WIND.

Statistics. The typical tornado (described here for the Northern Hemisphere; north/south direction is the opposite for the Southern Hemisphere) is weak and short-lived (1–2 min), moves from the southwest at 30 mi/h (15 m/s), and inflicts damage to an area of about 1 mi × 150 ft (2 km × 50 m). In extreme cases, the length and width of the damage path may exceed 100 mi (150 km) and 2.5 mi (4 km), respectively; the lifetime may be over 1 h; and the translation speed may reach 68 mi/h (30 m/s). The majority of tornadoes (60%) approach from the southwest, and only 4% have a westward component of motion.

Climatology. Tornadoes occur most often at latitudes between 20° and 60°, and they are relatively

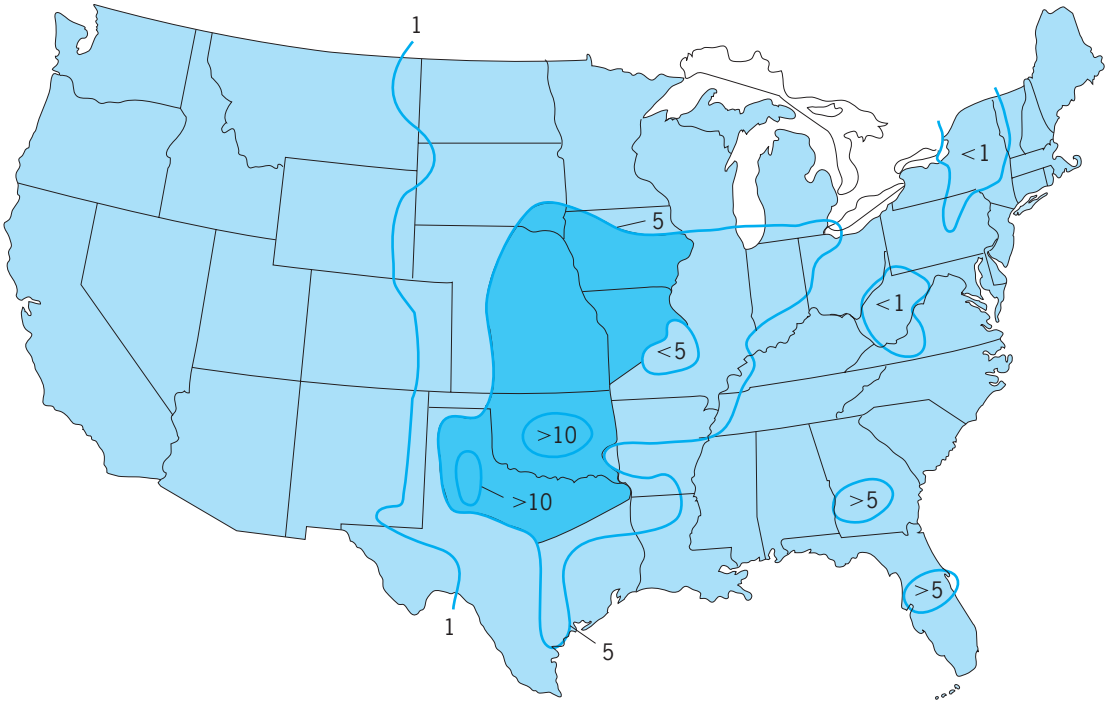


Fig. 2. Map showing average annual number of tornado occurrences per 10,000 mi^2 (26,000 km^2) from National Severe Storms Forecast Center data, 1953–1975. Shaded area shows the approximate location of so-called Tornado Alley. Contour lines indicate the number of tornadoes per 10,000 mi^2/year. West of the 1 contour line there is less than 1 per 10,000 mi^2 per year. (*After R. E. Peterson, ed., Proceedings of the Symposium on Tornadoes: Assessment of Knowledge and Implications for Man, Texas Technical University, 1976*)

frequent in the United States, Russia, Europe, Japan, India, South Africa, Argentina, New Zealand, and parts of Australia. Violent tornadoes are confined mainly to the United States, east of the Rocky Mountains, and (with less frequency) to the Bangladesh-Assam area. The world's highest frequency, 5–10 tornadoes annually per 10,000 mi^2 (26,000 km^2), occurs in the area known as Tornado Alley, which extends from Texas through Oklahoma and Kansas into Iowa (**Fig. 2**).

In winter in the United States, tornadoes are confined to the southern states, and generally have longer tracks because of stronger steering currents. The tornado belt generally shifts northward with the jet stream and the advance of tropical air through the central states and into the northern states during spring and summer. May is the most active month with 20% of the annual number of tornadoes; January is the least active with only 3%. Tornadoes occur at all hours but are most frequent between 1500 and 1900 local standard time because solar heating of the Earth's surface influences thunderstorm initiation.

Atmospheric conditions. Essentially, there are five atmospheric conditions that set the stage for widespread tornado development: (1) a surface-based layer, at least 3000 ft (1 km) deep, of warm, moist air, overlain by dry air at midlevels; (2) an inversion separating the two layers, preventing deep convection until the potential for explosive overturning is established; (3) rapid decrease of temperature with height above the inversion; (4) a combination of mechanisms, such as surface heating and lifting of the air mass by a front or upper-level disturbance, to eliminate the inversion locally; (5) pronounced vertical wind shear (variation of the horizontal wind with height). Specifically, storm-relative winds in the lowest 6000 ft (2 km) should exceed 20 knots (10 m/s) and veer (turn anticyclonically) with height at a rate of more than 10°/1000 ft (30°/km). (Storm-relative wind is the wind measured with respect to a coordinate system that moves with the storm.) Such conditions are prevalent in the vicinity of the jet stream and the low-level jet. Incidentally, the ground-relative winds often are quite similar to the storm-relative winds in strength and directional turning.

The first three conditions above indicate that the atmosphere is in a highly metastable state. There is a strong potential for thunderstorms with intense updrafts and downdrafts. The fourth condition is the existence of a trigger to release the instability and initiate the thunderstorms. The fifth is the ingredient for updraft rotation. When conditions warrant, the Storm Prediction Center issues tornado watches about 1 h in advance for rectangular areas of roughly 25,000 mi^2 (65,000 km^2) that indicate that the potential for tornadoes exists in the area for a 6-h period. *See* AIR MASS; FRONT; JET STREAM; TEMPERATURE INVERSION; THUNDERSTORM.

The reason that the world's worst tornadoes generally occur in the central United States and in Bangladesh-Assam is that both regions are located in the midlatitude belt of westerlies, with warm oceans equatorward and mountains to their west. During the spring, both regions are often in the transition zone between warm and cold air masses, so that they lie beneath the jet stream. Cyclones and upper-level disturbances generally develop in the vicinity of the jet stream and migrate along it. Ahead of cyclones

approaching from the west, warm, very moist air flows off the ocean into the region at low levels. At midlevels the air has been dried out by passage over the mountains to the west. Furthermore, the geography of both regions favors the development of low-level jets and associated strong low-level vertical wind shear.

Detection. Microbarographs, radars, satellites, and sferic detectors have all been used to identify potentially tornadic storms. Tornado warnings of about 1 h duration are issued by local weather service offices for specific countries on the basis of radar observations and reports from trained spotters; such warnings have contributed to a 50% decline in the annual numbers of deaths since 1950. At night, flashes from downed power lines may reveal the existence of a tornado. Conventional warning radars measure only the spatial distribution of the reflectivity of precipitation. Rotation is inferred by a hook-shaped appendage on the southwest side of the radar echo, observed sometimes at close range. The best detection device is a Doppler radar which measures the velocity components along the radar beam of raindrops and hailstones in addition to the reflectivity. Since 1988, a warning network of approximately 150 Doppler radars has been installed across the United States. Two types of vortex signature appear in Doppler velocity fields, the mesocyclone signature and the tornadic vortex signature (TVS). The TVS is the Doppler velocity field of the flow near the tornado, poorly resolved because the radar beam is broader than the tornado. It is generally seen only at close range because the width of the radar beam and the elevation of the radar horizon above the ground increase with range. At long range, "radar-indicated" tornadoes are based on detection of the much wider mesocyclone. For large tornadoes with signatures that first appear in the parent cloud, both signatures may be detected aloft 20 min or longer prior to touchdown, allowing warnings to be issued with considerable lead time. The TVS extends upward as well as downward, reaching heights of 40,000 ft (13 km) in rare cases. However, 50% of TVSs form as a column from the ground to more than 1 mi (2 km) and extend upward in time. Tornadoes associated with this type of signature form rapidly in 5–10 min and may occur with little advance warning. Untested detection instruments include an infrasonics sensor and seismic tornado detector. Home barometric devices based on detecting rapid pressure falls are inadequate because the warning may precede the tornado strike by only seconds. *See* DOPPLER RADAR; RADAR METEOROLOGY; SATELLITE METEOROLOGY; SFERICS; STORM DETECTION.

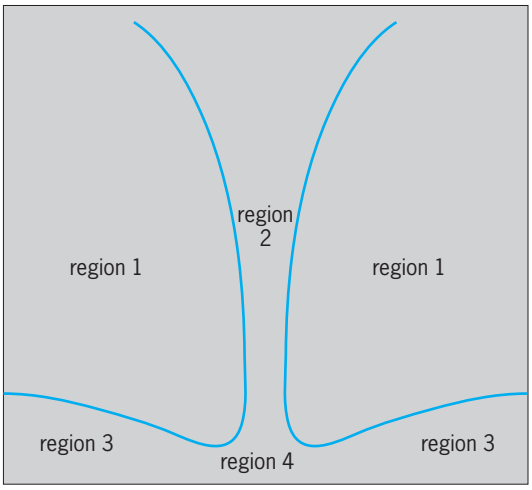


Fig. 3. Different flow regions of a tornado. (*After E. Kessler, ed., Thunderstorm Morphology and Dynamics, 2d ed., University of Oklahoma Press, 1986*)

Measurements. Winds and pressure deficits in tornadoes are not well known because tornadoes are difficult to intercept and extremely dangerous. Various estimates have been obtained from photogrammetric analyses of tornado movies, engineering analyses of structural failures and missiles generated by the vortex, analyses of the cycloidal marks left by multivortex tornadoes, chance readings from anemometers and microbarographs (when not destroyed) and attempts to introduce probes directly, Doppler radar measurements, and analysis of the size and shape of the funnel cloud. Portable Doppler radars driven to within a mile or two of large, violent tornadoes have recorded wind speeds of up to 280 mi/h (125 m/s). As yet, there is no evidence of wind speeds in excess of 300 mi/h (130 m/s). Substantial vertical velocities, up to 180 mi/h (80 m/s), have been documented. Radial inflow velocities may reach 110 mi/h (50 m/s) in narrow bands close to the ground. A surface pressure drop of 55 millibars (5.5 kilopascals) was recorded on a sensor placed in the path of a large, violent, multivortex tornado in 1995. This is probably not the maximum pressure deficit of the tornado because the measurement was neither in the center of the tornado nor in a satellite vortex. The temperature (also recorded) dropped a few degrees Celsius during the tornado owing to the low pressure. Temperature above the ground inside a tornado is unknown. *See* AIR PRESSURE.

Origins of rotation. The basic measure of local spin in a fluid is the vorticity vector, which is the curl or rotation of the wind vector and is aligned parallel to the spin axis in the direction of advance of a right-hand screw. The vorticity at a specified point and time is equal to twice the angular velocity of a minute spherical parcel of air that is centered on the point at that time. The strong vertical wind shear generally present prior to development of severe storms is associated with horizontal vorticity that is up to 100 times greater than the background vertical vorticity associated with the Earth's rotation and large-scale cyclones. Theory suggests that a thunderstorm updraft could develop a tornado directly from the background vertical vorticity in a few hours. However, tornadoes can be generated more rapidly from horizontal vorticity. Furthermore, three-dimensional computer models generate storms with rotating updrafts in suitably sheared environments with the Earth's rotation "switched off." Thus, it is widely believed that the mesocyclone forms as a result of the updraft tipping preexisting horizontal

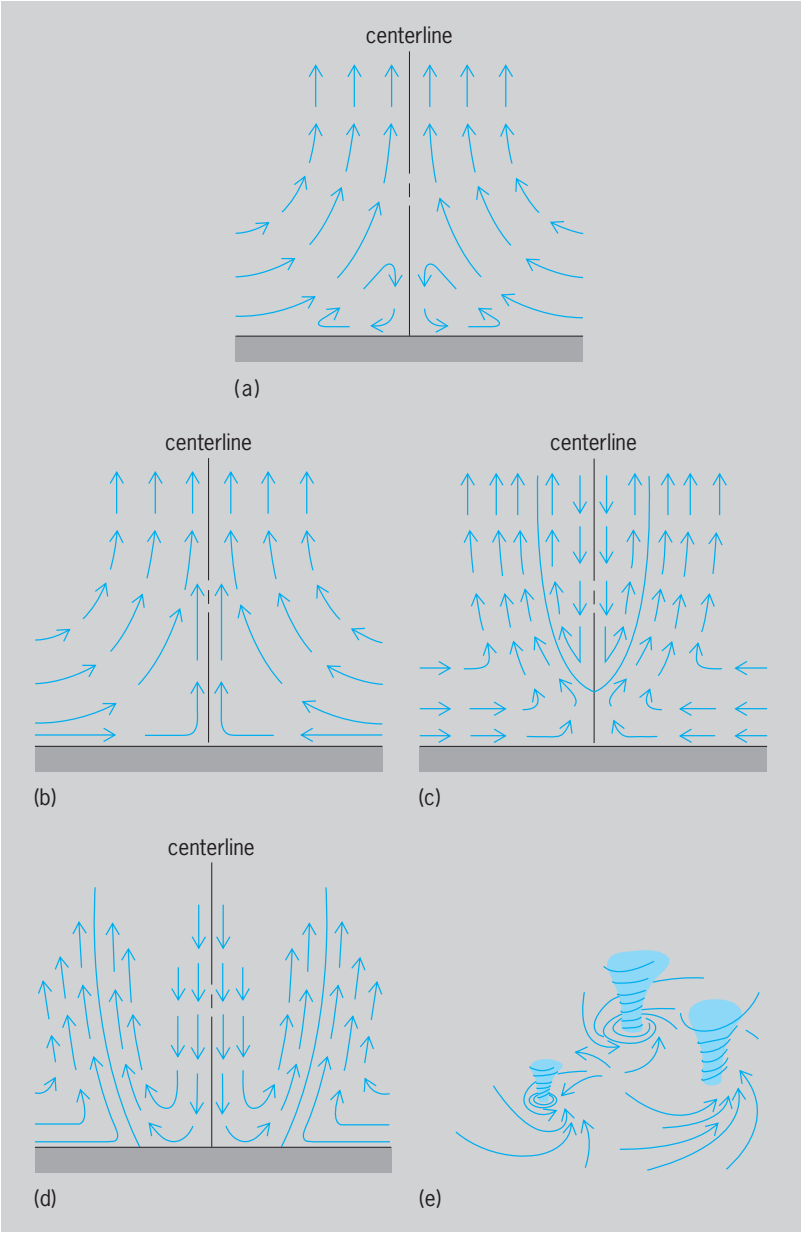


Fig. 4. Effect of increasing swirl ratio on the tornado flow. (*a*) Weak swirl—no tornado as flow in boundary layer separates and passes around corner region. (*b*) One-cell tornado (axial updraft). (*c*) Vortex breakdown (axial downdraft aloft does not reach ground). (*d*) Two-cell tornado with downdraft impinging on ground (core radius increases rapidly with increasing swirl ratio). (*e*) Multiple vortices (two to six with increasing swirl ratio). Only the flow in a radial-height cross section is depicted in *a–d*. (*After E. Kessler, ed., Thunderstorm Morphology and Dynamics, 2d ed., University of Oklahoma Press, 1986*)

vorticity toward the vertical. The updraft rotates as a whole when the vorticity is streamwise (that is, the spin axis is along the flow direction) relative to the storm. At this stage there is little rotation at very low levels because the vertical vorticity is being produced in rising air. *See* THUNDERSTORM; VORTEX.

Generation of rotation near the ground is a more complicated process. It does not happen in storm simulations if the evaporation of precipitation is switched off. In simulated storms, it originates in an initially dry airstream that flows through a precipitation region at the rear of the rotating updraft, is chilled by evaporating rain, subsides, and flows along the ground. The left side of this airstream is drawn up into the storm's main updraft, and a wall cloud (markedly lowered cloud base) forms in the process because the cool air is nearly saturated. The rotation is generated around the periphery of this downdraft in the form of a toroidal circulation (the coldest air in the center of the downdraft sinks faster than surrounding air). Each vortex ring is elongated in the downstream direction owing to the transport of vorticity by the horizontal wind, and tilted by the gradient of vertical wind since its back edge remains in the generation region in the downdraft and its leading edge moves downstream out of the downdraft. The vertical vorticity produced by the tilting is cyclonic on the left side of the airstream, and anticyclonic on the right side. The cyclonic side is drawn into the storm's rotating updraft, thus connecting the mesocyclonic vortex already present in the updraft to the ground. Horizontal contraction and vertical stretching of the lower part of this vortex by the updraft causes it to spin faster. Near the ground this spin-up is enhanced by radial inflow induced by friction, the same effect that causes accumulation of the tea leaves at the center of a stirred cup of tea. A cyclonic tornado may form simply as the result of continued uninterrupted contraction of this column for around 15 min. The process may be aborted prematurely by the lower part of the vortex being overtaken by cold sinking air, the same process that seems to cause tornadoes to decay. The anticyclonic side of the airstream is further away from the storm's updrafts and seldom is spun up into an anticyclonic tornado. Detailed analysis of data collected in field experiments such as VORTEX (Verification of the Origins of Rotation in Tornadoes Experiment), conducted in 1994–1995, which utilize exciting new mobile instrumentation, should provide information on tornado genesis.

A different mechanism applies to small tornadoes and waterspouts that develop in environments with insignificant vertical shear. These tend to develop along wind-shift lines, which become dynamically unstable and roll up into rows of vertical vortices, about 0.6 mi (1 km) in diameter. If one of the vortices happens to be located under the updraft of a cumulus congestus or cumulonimbus cloud, then a tornado may form as a result of the vertical stretching effect. These tornadoes develop when the temperature at low levels decreases strongly with height, and, unlike supercell tornadoes, occur early in the life of the parent cloud.

Energy source. The rate at which kinetic energy is produced in a tornado is estimated to be 10^3 MW in order of magnitude. To explain the high energy density of tornadoes, various exotic energy sources have been postulated in the past, including electrical heating from "rapid-fire" lightning discharges in a confined area. However, it is now believed that the tornado's energy is derived mostly from the buoyant potential energy that is released in the parent cloud by condensing water vapor.

Flow structure. Theory, laboratory simulations, computer models, and analysis of films reveal four basic flow regions in tornadoes (**Fig. 3**). The first (region 1) is the outer flow away from the ground

and the axis of rotation. This region consists of horizontally converging, rising air that approximately conserves its angular momentum. Consequently, the air spins faster as it approaches the axis. The core is region 2. It surrounds the axis, extending outward roughly to the radius of maximum winds. The core is approximately in solid-body rotation. The core flow is generally stable against radial displacements; and so it supports centrifugal waves, which are often seen moving up or down the funnel. This stability also prevents air from region 1 from being entrained into the core. The flow along the axis may be either upward or downward, and there may be stagnation points aloft separating axial upflow and downflow. In large tornadoes, the axial flow is probably downward all the way to the ground. Interaction with the ground creates a turbulent boundary layer, perhaps a few hundred meters deep, of reduced tangential velocities. In this third region, the outward centrifugal forces no longer can balance the inward pressure gradient forces (as they do in region 1). The resulting net inward force drives a strong radial inflow into region 4, the corner region at the foot of the vortex, where the strongest winds are located. Here, conservation of mass dictates a strong updraft into the core, either along the axis or surrounding a central downdraft. In the corner region, inflowing parcels overshoot the radius that they would reach in the absence of friction. Closer penetration to the axis would cause parcels to rotate faster if angular momentum losses were small enough. Conversely, loss of angular momentum would lead to slower rotation if the overshoot were negligible. The former happens to be the larger effect for all parcels except for those in the lowest few meters. Paradoxically friction causes more intense tornadoes!

Laboratory simulations reveal that the structure of turbulent vortices depends largely on a single flow parameter, the swirl ratio, which is roughly the ratio of the tangential and vertical velocities in the parent rotating updraft. As the swirl ratio increases, the core widens, and the flow changes (**Fig. 4**). At a moderate swirl ratio (Fig. 4*c*), the vortex breaks down from a high-speed rotating jet (with the wind more upward than tangential) to a slower broader vortex with downflow along the axis. The high wind speeds in the jet are associated dynamically with large pressure deficit on the axis. Below the breakdown point the flow is faster than the propagation speed of centrifugal waves so the waves are unable to propagate downward and modify the flow to a slower state. Tornadoes with this structure are narrow but strong, and have the capability of lofting heavy objects such as pickup trucks 100 ft (30 m) in the air. At large swirl ratio (Fig. 4*d*), the downflow along the axis reaches the surface, and parcels flowing into the tornado along the ground do not come close to the axis before rising. This type of tornado is often violent owing to large ambient angular momentum and low central pressure associated with a warm, light core, which arises from air descending along the axis and warming adiabatically. At still larger swirl ratios, transitions occur, first from a single vortex to two, then to three (Fig. 4*e*), up to six vortices. Extreme winds occur in these secondary vortices. *See* VORTEX.

Robert Davies-Jones

Bibliography. C. Church et al. (eds.), *The Tornado: Its Structure, Dynamics, Prediction, and Hazards*, 1993; R. Davies-Jones, Tornado formation and structure, *Sci. Amer.*, 1995; R. Davies-Jones, Tornadoes, *Sci. Amer.*, 273:34–41, 1995; E. Kessler (ed.), *Thunderstorm Morphology and Dynamics*, 2d ed., 1986; R. E. Peterson (ed.), *Proceedings of the Symposium on Tornadoes: Assessment of Knowledge and Implications for Man*, Texas Technical University, 1976; J. T. Snow, The tornado, *Sci. Amer.*, 250:86–97, 1984.

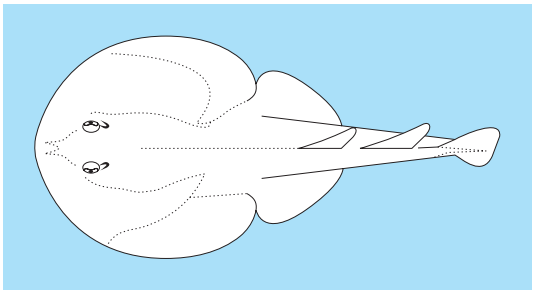
Torpediniformes

An order of batoid fishes occurring in the subclass Elasmobranchii and known as the torpedo electric rays and numbfishes. Typical members of Torpediniformes are identifiable by their flat, pancakelike body sector (disc); relatively robust tail sector; smooth skin; small or obsolete eyes; distinct caudal fin; and ovoviviparous development (that is, producing eggs that develop internally and hatch before or soon after extrusion). Of special interest is a pair of enlarged electric organs located on the disc lateral to the gill slits. These kidney-shaped organs can best be seen from the ventral side, although the columnar structures that compose them occupy the thickness of the disc. These electric organs, which may constitute 17% of the total body weight, deliver shocks up to 220 volts. The voltage depends on the species, its size, and physical condition. The electric organs are thought to be used primarily for feeding and defense. Electric rays are poor swimmers, depending primarily on the tail because the disc is rather inflexible and of little use in locomotion, although it is flexible enough for capturing prey. They spend most of their time partially buried in sand or mud and feed on a variety of invertebrates, including crustaceans, mollusks, and worms, as well as small fishes. Members of the genus *Torpedo* are reported to reach a length of 1.8 m (6 ft) and a weight of 44 kg (100 lb). Torpediniforms occur in intertidal waters to deep waters [1070 m or 3510 ft in the case of blind species (*Benthobatis*)] in temperate to tropical zones of all oceans.

The order comprises two families: Torpedinidae and Narcinidae.

Torpedinidae (torpedo electric rays). The disc is truncate anteriorly or emarginate; the rostrum is reduced; the jaws are extremely slender and lack labial cartilages. There are two genera, *Torpedo* (21 species) and *Hypnos* (1 species). *Torpedo* differs from *Hypnos* in having well-developed dorsal and caudal fins vs. very small dorsal and caudal fins, and in having worldwide distribution vs. distribution limited to off Australia.

Narcinidae (electric rays or numbfishes). The disc is rounded anteriorly; a rostrum is present; the jaws are stout and the labial cartilages strong. There are nine genera and 37 species in two subfamilies, Narcininae (numbfishes) with four genera and 26 species and Narkinae (sleeper rays) with five genera and



Example of a Narcinidae species. (*Courtesy of J. S. Nelson, Fishes of the World, 4th ed., Wiley, 2006*)

11 species. Narcininae (see **illustration**) differs from Narkinae in having long, strongly protractile jaws vs. short, weakly protractile jaws; a deep groove around the mouth and lips vs. shallow grooves; a broad rostrum vs. narrow; usually two dorsal fins vs. usually one dorsal fin; and species occur in Atlantic, Indian, and Pacific oceans vs. limited to Indo-West Pacific. *See* BATOIDEA; ELASMOBRANCHII; ELECTRIC ORGAN (BIOLOGY); PRISTIFORMES; RAJIFORMES; MYLIOBATIFORMES.

Herbert Boschung

Bibliography. H. B. Bigelow and W. C. Schroeder, in J. Tee-Van et al. (eds.), *Fishes of the Western North Atlantic*, Sears Found. Mar. Res. Mem. 1, pt. 2, 1954; L. J. V. Compagno, Checklist of Chondrichthyes, pp. 503–547 in W. C. Hamlett (ed.), *Reproductive Biology and Phylogeny of Chondrichthyes: Sharks, Batoids and Chimaeras*, Science Publishers, Enfield, NH, 2005; J. D. McEachran and N. Aschliman, Phylogeny of Batoidea, pp. 79–113 in J. C. Carrier, J. A. Musick, and M. R. Heithaus (eds.), *Biology of Sharks and Their Relatives*, CRC Press, Boca Raton, FL, 2004; J. D. McEachran and M. R. de Carvalho, Batoid fishes, pp. 507–589 in K. E. Carpenter (ed.), *The Living Marine Resources of the Western Central Atlantic: FAO Species Identification Guide for Fishery Purposes*, vol. 1, FAO, Rome, 2002; J. D. McEachran, K. A. Dunn, and T. Miyake, Interrelationships of the batoid fishes (Chondrichthyes: Batoidea), pp. 63–84 in M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson (eds.), *Interrelationships of Fishes*, Academic Press, San Diego, 1996; J. S. Nelson, *Fishes of the World*, 4th ed., Wiley, New York, 2006.

Torque

The product of a force and its perpendicular distance to a point of turning; also called the moment of the force. Torque produces torsion and tends to produce rotation. Torque arises from a force or forces acting tangentially to a cylinder or from any force or force system acting about a point. A couple, consisting of two equal, parallel, and oppositely directed forces, produces a torque or moment about the central point. A prime mover such as a turbine exerts a twisting effort on its output shaft, measured as torque. In structures, torque appears as the sum of moments of torsional shear forces acting on a transverse section of a shaft or beam. *See* COUPLE; TORSION.

Nelson S. Fisk

Torque converter

A device for changing the torque-speed ratio or mechanical advantage between an input shaft and an output shaft. A pair of gears is a mechanical torque converter. A hydraulic torque converter, with which this article deals, is an automatically and continuously variable torque converter, in contrast to a gear shift, whose torque ratio is changed in steps by an external control. *See* AUTOMOTIVE TRANSMISSION; MECHANICAL ADVANTAGE.

Converter characteristics. A mechanical torque converter transmits power with only incidental losses; thus, the power, which is the product of torque T and rotational speed N, at input I is substantially equal to the power at output O of a mechanical torque converter, or $T_I N_I = kT_O N_O$, where k is the efficiency of the gear train. This equal-power characteristic is in contrast to that of a fluid coupling in which input and output torques are equal during steady-state operations. *See* FLUID COUPLING.

In a hydraulic torque converter, efficiency depends intimately on the angles at which the fluid enters and leaves the blades of the several parts. Because these angles change appreciably over the operating range, k varies, being by definition zero when the output is stalled, although output torque at stall may be three times engine torque for a single-stage converter and five times engine torque for a three-stage converter. Depending on its input absorption characteristics, the hydraulic torque converter tends to pull down the engine speed toward the speed at which the engine develops maximum torque when the load pulls down the converter output speed toward stall.

Converter power efficiency is highest (80–90%) at a design speed, usually 40–80% of maximum engine speed, and falls toward zero as shaft speed approaches engine speed. Because of this characteristic, the mode of operation may be modified to change from torque conversion to simple fluid coupling or to direct mechanical drive at high speed.

Hydraulic action. These characteristics are achieved by the exchange of momentum between the solid parts of the converter and the fluid (**Fig. 1**). A vaned impeller on the input shaft pumps the fluid from near the axis of rotation to the outer rim. Fluid momentum increases because of the greater

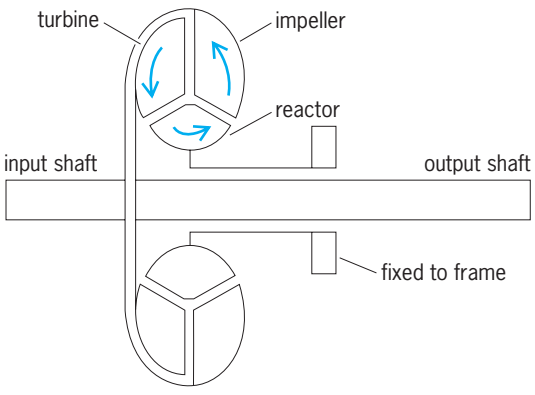


Fig. 1. Elementary hydraulic torque converter.

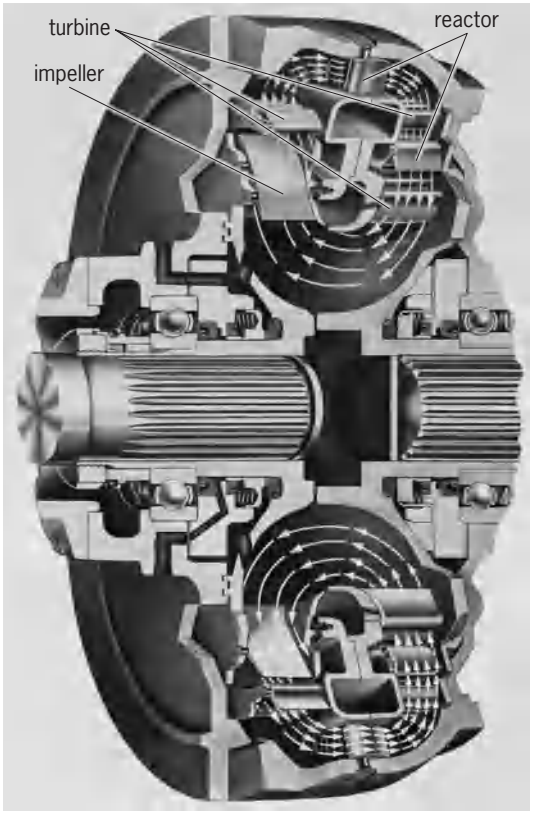


Fig. 2. Three-stage converter showing simplified fluid flow around torus. (*Twin Disc, Inc.*)

radius and the influence of the vanes. The high-energy fluid leaves the impeller and impinges on the blades of a turbine, giving up its energy to drive the turbine, which is connected to the output shaft. The fluid discharges from the turbine into a bladed reactor. The reactor blades are fixed to the frame; they deflect the fluid flow and redirect it into the impeller. This change in flow direction produced by the stationary reactor is equivalent to an increasing change in momentum which adds to the momentum imparted by the impeller to give a torque increase at the output of the converter (**Fig. 2**). *See* HYDRAULICS.

In a typical converter, as the output shaft comes up to the speed of the input shaft, efficiency decreases. Therefore, the reaction member may be mounted on a freewheel unit so that it rotates with the fluid at high speed ratio when torque multiplication is no longer possible. In addition, splitting the reaction member to give a four-element polyphase converter gives even more uniform efficiency. *See* SHAFTING.

Henry J. Wirry

Bibliography. G. Raczkowski, *Principles of Machine Dynamics*, 1979; H. F. Tucker, *Automatic Transmissions*, 1980.

Torricelli's theorem

The proposition that the speed of efflux of a liquid from an opening in a reservoir equals the speed that the liquid would acquire if allowed to fall from rest from the surface of the reservoir to the opening.

Torricelli, a student of Galileo, observed this relationship in 1643. In equation form, $v^2 = 2gh$, in which v is the speed of efflux, h the head (or elevation difference between reservoir surface and center line of opening if in a vertical plane), and g the acceleration due to gravity. (The equation is the same as that for a solid particle dropped a distance h in a vacuum.) The relationship can be derived from the energy equation for flow along a streamline, if energy losses are neglected.

An orifice (opening in the wall or bottom of a reservoir) is used as a flow-measuring device. From Torricelli's theorem, by solving for v and multiplying by the flow area, an expression for discharge Q, in volume per unit time, is obtained. In equation form, $Q = C_d A\sqrt{2gh}$, in which A is the area of opening and C_d is a dimensionless coefficient, determined experimentally, that corrects for contraction of the jet as it leaves the orifice and for energy loss due to viscosity. When h is measured, Q may be determined from the above formula. *See* BERNOULLI'S THEOREM; FLOW MEASUREMENT.

Victor L. Streeter

Torsion

A straining action produced by couples that act normal to the axis of a member. Torsion is identified by a twisting deformation.

In practice, torsion is often accompanied by bending or axial thrust as in the case of line shafting driving gears or pulleys, or propeller shafts for ship propulsion. Other important examples include springs and machine mechanisms usually having circular sections, either solid or tubular. Members with noncircular sections are of interest in special applications, such as structural members subjected to unsymmetrical bending loads that twist and buckle beams.

The shear properties of materials are determined by a torsion test. *See* SHEAR.

Cylindrical bars. The twist of a bar due to torque can be visualized as the accumulated rotational displacements of imaginary disks cut by transverse sections on which tangential forces operate. Shearing forces vary across the section and together furnish the internal resisting torque.

Torsional angle, designated θ, is the total relative rotation of the ends of a straight cylindrical bar of length L, when subjected to torque (**Fig. 1**).

Helical angle, designated ϕ, is the angular displacement of a longitudinal element, originally straight on

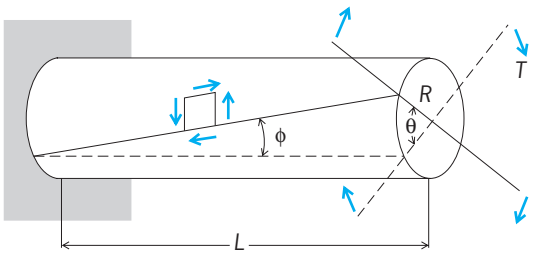


Fig. 1. Cylindrical bar in torsion.

the surface of the untwisted bar, which becomes helical after twisting (Fig. 1). Angle ϕ is the shear strain. For small twist, torsional and helical angles are related by geometry $\phi = R\theta/L$, where R is the radius of the bar.

Elastic shear stress. Within the elastic limit, shear stress S_s is found by Hooke's law, $S_s/\phi = E_s$, and is expressed in terms of the torsional angle as $S_s = (R/L)E_s\theta$, where E_s is the modulus of rigidity. *See* HOOKE'S LAW.

The shear stress varies linearly across the section, being maximum at the surface and zero at the center. For a circular section the maximum shear stress acting perpendicular to the radius at the extreme distance $R = D/2$ from the neutral axis is $S_{max} = 16T/\pi D^3$, where T is the externally applied twisting moment.

Tangential shear stresses on the section are accompanied by longitudinal shear stresses along the bar. These complementary stresses induce tensile and compressive stresses, equal to the shear intensity, at 45° to the shear stresses. The longitudinal stresses are important in laminated materials, wood, or metals with seams. Brittle materials, low in tensile strength, fracture on a 45° helicoidal surface; ductile materials fracture on transverse sections after large twist.

Resisting torque equal to the applied torque is the moment of the internal shear forces about the neutral axis expressed in terms of the sectional dimensions and the stresses. A general expression for resisting torque is $T = S_{max}J/R$, where J is the polar moment of inertia of the section. This relation is applicable to both solid and tubular circular sections which are differentiated by J. In terms of torque T, torsional angle θ is TL/E_sJ. Torsional angle per unit of length is a measure of torsional stiffness, which may limit the required dimensions of a shaft. In power transmission the torque associated with horsepower is found from hp $= TN/63{,}000$, where T is expressed in inch-pounds of moment and N is the rotation of the shaft in revolutions per minute.

Inelastic behavior in torsion. Strains exceeding the elastic limit are not completely recoverable after unloading and the behavior is inelastic. Torsional strains vary linearly from the center of the bar during elastic and plastic deformation, and the corresponding shear stresses reflect the stress-strain curve for the material (**Fig. 2**). After the extreme element reaches the yield point, continued twisting produces inelastic strains at increasing distances from the surface while the stress remains constant. When the action is fully plastic, the stresses are constant, equal to the yield point over the entire section. The fully plastic resisting torque is shown by Eq. (1), which is 1.33

$$T_p = \frac{4}{3}\frac{S_{yp}J}{R} \tag{1}$$

times that required to just produce surface yielding [S_{yp} = shear stress at yield point]. Torsional resistance increases because of strain hardening but is of interest only where large deformation can be tolerated. Elastic analysis is applicable to designs where

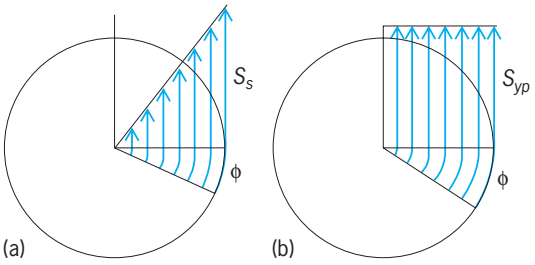


Fig. 2. Stress distribution. (*a*) Elastic torsional stress. (*b*) Fully plastic torsional stress.

permanent deformation must be avoided and where endurance (fatigue) properties limit the stresses.

Thin-walled tubes. Thin tubular members find application particularly in aircraft. Shear stresses are assumed uniform over the wall thickness when a thin-walled tube of any shape is subjected to torque at the ends. Shear force q per unit length of perimeter is constant.

Shear flow is the constant shear force q acting along the median line of the wall and is equal to the product of shear stress S times thickness t at any point; thus $q = St$ is constant. The concept of flow is drawn from the similarity of the expression for constant shear force with the constant quantity Q of a liquid passing variable sections of a channel having area A and velocity V, $Q = AV$. Resisting torque T is the summation of moments of shear forces on unit lengths ds of the wall perimeter about the center of rotation $T = 2Aq$, where A is the area enclosed by the centerline of the tube wall (**Fig. 3**). The stress at any point where thickness is t is $S = q/t = T/2At$. The torsional angle produced by applied torque T is found from Eq. (2), where S is the length of the perimeter

$$\theta = \frac{TL}{4A^2E}\int_0^S \frac{ds}{t} \tag{2}$$

and t is the variable thickness. For constant thickness, $\theta = TLS/4A_2Et$, where S is peripheral length of the centerline.

Solid noncircular sections. When a solid member with noncircular section is twisted, the sections become warped and the stresses do not vary linearly as in the case of circular sections. Evaluation of stresses and torsional twist requires the rigorous procedures

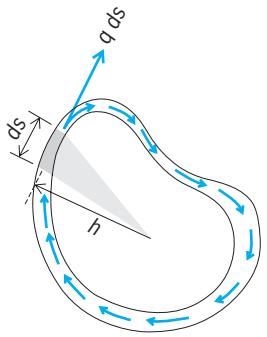


Fig. 3. Diagram showing the shear flow in a thin-walled tube.

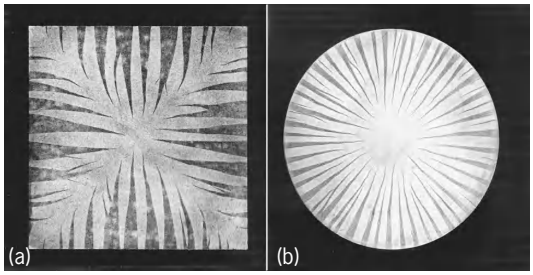


Fig. 4. Plastic strain in torsion. (*a*) Square bar. (*b*) Round bar.

of the theory of elasticity. If a grid is scribed on the surface of a square or rectangular bar and the bar is twisted, distortions of the grid indicate that maximum shear stress is at a boundary nearest the center. Contrary to theory applicable to circular sections, the stress is zero at the corners, which are the most remote elements. The location of maximum stress is indicated by points of initial plastic yielding as shown by the macrographs of a square and a round bar (**Fig. 4**). Sections were etched after yielding, thus differentiating the darker plastic zones. Formulas for maximum shear stress and torsional angle for common noncircular sections are presented in **Fig. 5**.

Helical springs subjected to axial loads involve all four possible straining actions: direct stress, transverse shear, bending, and torsional shear. For small obliquity of the coils, as in close-coiled springs, torsional shear is the most important action. When stresses and deflection are determined by formulas applicable to straight bars, a correction is necessary to account for the effect of curvature of the coils. *See* SPRING (MACHINES).

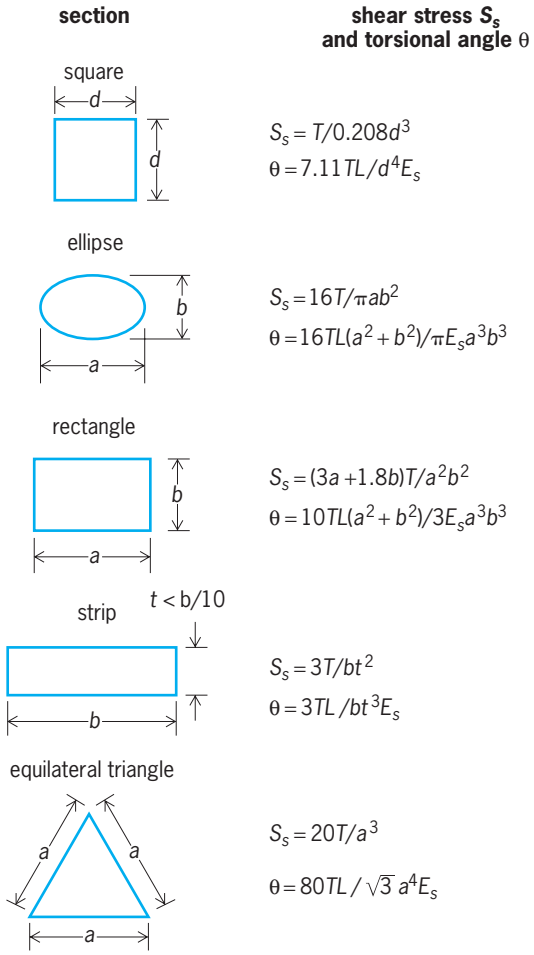


Fig. 5. Formulas for maximum shear stress and torsional angle for common noncircular sections.

Membrane analogy. Shearing stresses in sections which cannot be conveniently analyzed mathematically are determined experimentally by membrane analogy. The analogy presented by Ludwig Prandtl (1903) is based on the similarity of the equilibrium equation for a membrane with pressure on one side and the differential equation for torsional stresses. In application, a thin membrane, such as a soap film, is placed over an opening in a plate having the same geometrical shape as the section under investigation. Slight air pressure on one side deflects the film, and micrometer measurements determine the contours of equal deflection. The slope at any point and the volume enclosed by the deflected membrane can be found from these measurements. If a bar having this section is twisted, the torsional shearing stress at any point is proportional to the slope of the membrane, the stress direction is tangent to the contour, and the torque is proportional to the volume enclosed by the deflected membrane.

The method is a valuable qualitative aid in locating points of maximum stress by visualizing or observing points of maximum slope of the deflected film. The high stress at a reentrant corner, such as at a fillet of a structural angle or channel section, is indicated by a steep slope of the film. John B. Scalzi

Bibliography. J. S. Golan, *Torsion Theories*, 1987; W. D. Pilkey, *Formulas for Stress, Strain, and Structural Matrices*, 1994.

Torsion bar

A spring flexed by twisting about its axis. Design of a torsion bar spring is primarily based on the relationships between the torque applied in twisting the spring, the angle through which the torsion bar twists, and the physical dimensions and material (modulus of elasticity in shear) from which the torsion bar is made. The **illustration** shows the elements of a simple torsion bar and the important dimensions involved in its design. Equation (1) relates

$$\theta = \frac{32Fal}{\pi D^4 G} \qquad (1)$$

these dimensions. Here θ is angle of twist in radians, F is force in pounds, a is radius of arm of force in inches, l is length of torsion bar in inches, D is diameter of torsion bar in inches, and G is modulus of elasticity in shear in pounds per square inch. *See* TORQUE.

If the deflection or twist of the spring θ is large, force F must change direction if a is to remain constant. For this reason Eq. (1) is frequently written as

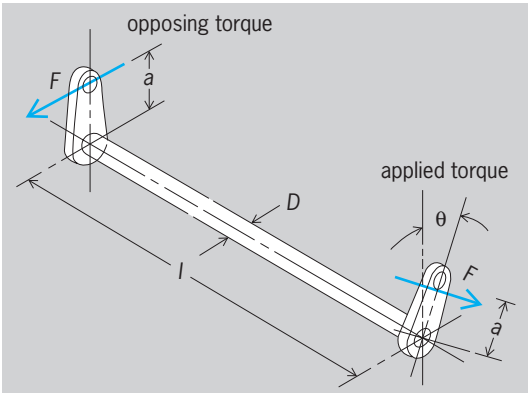


Diagram of a torsion bar.

Eq. (2), in which τ is torque in inch-pounds.

$$\theta = \frac{32\tau l}{\pi D^4 G} \qquad (2)$$

Torsion bar springs are found in the spring suspension of truck and passenger car wheels, in production machines where space limitations are critical, and in high-speed mechanisms where inertia forces must be minimized. *See* SPRING (MACHINES); TORSION.

L. Sigfred Linderoth, Jr.

Torus

A surface obtained by rotating a circle about a line that lies in the plane of the circle but that has no points in common with the circle (see **illus.**). The

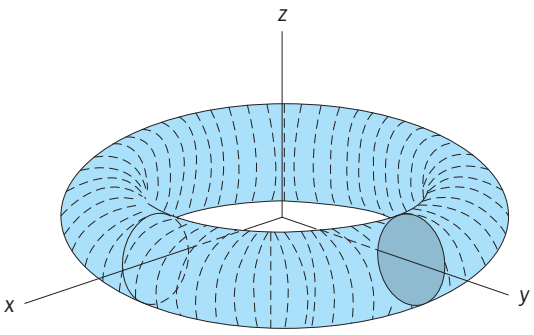


Diagram of a torus.

equations $x = u \cos v$, $y = u \sin v$, $z = [r^2 - (u - b)^2]^{1/2}$, $b > r > 0$, represent the upper half of the torus obtained by rotating about the z axis a circle of radius r whose center is the point $(b, 0, 0)$. The parameter u represents the distance of a point P of the torus from the z axis, and v is the angle of rotation. According to whether $b < u \leqq b + a$ or $b - a \leqq u < b$ or $u = b$, the corresponding point P is elliptic, hyperbolic, or parabolic, respectively, and the Gauss curvature of the surface at P is positive, negative, or zero. *See* DIFFERENTIAL GEOMETRY; MANIFOLD (MATHEMATICS); TOPOLOGY.

Leonard M. Blumenthal

Tourette's syndrome

A neurobehavioral disorder that is characterized by frequent, recurrent motor and vocal tics (described in 1885 by the French neurologist Gilles de la Tourette). The motor tics of Tourette's syndrome include brief, rapid, and darting movements of almost any muscle group, and can include eye blinking, eye rolling or deviations, nose wrinkling, facial grimacing, and head shaking. Some motor tics are more complex, are slow, and appear purposeful such as head turning, shoulder shrugging, touching, hopping, or twirling. Vocal tics are brief guttural sounds such as recurrent sniffing, throat clearing, coughing, and grunting or barking sounds. Complex vocal tics can be more meaningful and include expressions such as "No," "Stop," or "Some other time." Infrequently, persons with Tourette's syndrome repeat what they have said (palilalia), repeat what others have said (echolalia), or bark or grunt profane language (coprolalia). Tourette's syndrome has been described in nearly every country and ethnic group, with an estimated prevalence of one or two occurrences per 2000 people.

Symptoms. The motor and vocal tics begin in childhood. Within an individual the types and severity of tics will vary over time. The tics are often worse during adolescence and tend to improve during the twenties and thirties. The tic symptoms increase with stress and excitement and decrease with activities that require focused effort. While the motor and vocal tics are involuntary, they can be suppressed for brief periods of time, giving the false impression that the movements and sounds are voluntary. The characteristic waxing and waning nature of the tic movements can be confusing and often causes the disorder to be unrecognized and undiagnosed.

It is unusual for tic symptoms alone to be severe and incapacitating or to progress and worsen over the life of an individual. Often the impairment is related to associated behavioral problems, such as problems with attention, concentration, and impulsivity. Problems with recurrent, intrusive, and unwanted thoughts (obsessions) and behaviors or rituals (compulsions) can also cause significant impairment. Many other behavioral difficulties have been described in persons with Tourette's syndrome, but the scientific evidence for their association is not clear.

Inheritance pattern. There is a pattern of inheritance consistent with a single autosomal dominant gene whose expression is variable and dependent on the sex of the person. Symptoms can vary from transient tics to Tourette's syndrome and can include obsessive-compulsive symptoms. Males have the greatest risk of having some form of a tic disorder; females have a reduced risk of developing Tourette's syndrome but a greater risk than males of developing obsessive-compulsive symptoms. *See* HUMAN GENETICS; OBSESSIVE-COMPULSIVE DISORDER.

Brain involvement. The complexity of symptoms is likely related to the various brain regions implicated in the development of Tourette's syndrome.

The brain regions most likely to be involved are the basal ganglia, the frontal cortex, and the limbic system. The basal ganglia are responsible for movement control, the frontal cortex is responsible for the organization of thinking and feeling, and the limbic system is the considered the center of emotion in the brain. The dense interconnections of these brain regions may help explain why the symptoms of Tourette's syndrome can include uncontrolled motor movements and sounds in combination with unusual thoughts, feelings, and sensations. *See* BRAIN.

Treatment. Treatment can be targeted toward suppressing tics and the specific associated behavioral problems. Methods for tic suppression include medications such as haloperidol, fluphenazine, and pimozide—all major tranquilizers which exert their effect on the brain by blocking the neurotransmitter dopamine at the site of nerve-to-nerve connections. Treatments for the associated behavioral problems are dependent on the nature of the specific problem. Given that the overall impairment in persons with Tourette's syndrome is often related to the severity of the associated problems, the importance of treatment oriented toward these problems must be emphasized. *See* NERVOUS SYSTEM DISORDERS.

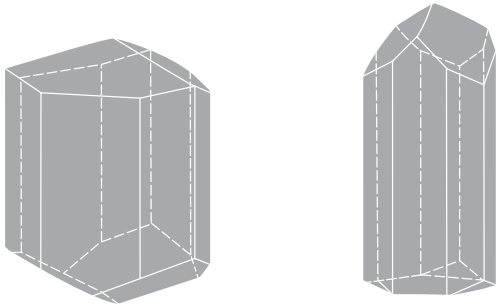
John T. Walkup; Mark A. Riddle

Bibliography. T. N. Chase (ed.), *Tourette Syndrome: Genetics, Neurobiology, and Treatment*, 1992; R. Kurlan (ed.), *Handbook of Tourette's Syndrome and Related Tic and Behavioral Disorders*, 1993.

Tourmaline

A cyclosilicate mineral family with (BO_3) triangular groups and a complex chemical composition. The general formula can be written $XY_3Al_6(OH)_4(BO_3)_3(Si_6O_{18})$, in which X = Na, Ca, and Y = Al, Fe^{3+}, Li, Mg, Mn^{2+}. The more common tourmalines are dravite (X,Y = Na,Mg), schorl (X,Y = Na,Fe), uvite (X,Y = Ca,Mg), and elbaite (X,Y = Na,Li). Fluorine commonly substitutes in the hydroxyl position. Tourmaline is a hard ($7\frac{1}{2}$ on Mohs scale), varicolored mineral which can be an important semiprecious gemstone. *See* GEM.

Tourmaline is trigonal, space group R3m, with $a = 1.584$ nanometers and $c = 0.71$ nm for an elbaite. Other members have similar cell dimensions, the variation in size dictated by the size of the ions.



Tourmaline crystal habits. (*After C. S. Hurlbut, Jr., Dana's Manual of Mineralogy, 17th ed, Wiley, 1959*)

Crystals are typically trigonal prismatic (see **illus.**) and may occur as slender needles in elbaite to thick prismatic individual crystals of schorl which have been found in pegmatites. The tourmaline crystal is polar; thus, different forms are found at the opposite ends. Because of this polarity, tourmaline is piezoelectric; that is, if pressure is exerted at one end, opposite electrical charges will occur at opposite poles. It is also pyroelectric, with the electrical charges developed at the ends of the polar axis on a change in temperature. Because of its piezoelectric property, tourmaline can be cut into gages to measure transient pressures. *See* PIEZOELECTRICITY; PYROELECTRICITY.

Color in tourmalines is largely caused by incorporation of minor amounts of cations of the first transition series in the Y position and substitution for some of the octahedrally coordinated Al. For example, some Mn^{2+} (or Mn^{3+}) in the Y position in elbaite can color it pink to burgundy red, and this mineral is called rubellite. With $Fe^{2+} > Mn^{2+}$, tourmalines of the elbaite clan can be yellow to green, and with some Fe^{3+} present as well the gemstone is blue. These minerals are often called indicolite. The absence of transition-metal chromophores results in colorless achroite. Such colored, transparent flawless elbaites constitute the bulk of gemstones, and they are culled from granitic pegmatites. In these pegmatites, tourmalines occur in late-stage large chambers or pockets, characteristically associated with platy albite (cleavelandite); lepidolite; aquamarine beryl (gemmy bluish-green); morganite beryl (gemmy pink); and rarer accessory minerals such as lithiophilite, $LiMn(PO_4)$ and bertrandite, $Be_4(OH)_2(Si_3O_7)$. Magnificent gems have been obtained from certain late-stage pegmatites such as in San Diego County, California; Oxford County, Maine; Minas Gerais district, Brazil; Madagascar; and Afghanistan. In the same pegmatites, but at an earlier stage, occurs black schorl. Schorl, of no gem value, also occurs in regionally metamorphosed schists. Uvite and dravite typically occur in metamorphic rock assemblages such as regionally recrystallized marbles which have been intruded by granites. Their colors are characteristically dull green to brown and they are of no gem value. *See* PEGMATITE; SILICATE MINERALS.

Paul Brian Moore

Bibliography. W. A. Deer, R. A. Howie, and J. Zussman, *Rock Forming Minerals: Disilicates and Ring Silicates*, vol. 1b, 1986.

Tower

A concrete, metal, or timber structure that is relatively high for its length and width. Towers are constructed for many purposes, including the support of electric power transmission lines, radio and television antennas, and rockets and missiles prior to launching.

Transmission towers. These towers are rectangular in plan and are not steadied by guy wires. A transmission tower is subjected to a number of forces: its own

weight, the pull of the cables at the top of the tower, the effect of wind and ice on the cable, and the effect of wind on the tower itself. Torsional forces on a tower caused by breakage of the cables in one span on one side of the tower must also be considered.

Radio and television towers. Such towers are either guyed or freestanding. Freestanding towers are usually rectangular in plan. In addition to their own weight, freestanding towers support the weight of the antenna and accessories and the weight of ice, unless a deicing circuit is installed. Wind forces must also be carefully considered.

Guyed towers are usually triangular in plan, with the main structural members, or legs, at the vertexes of the triangle. The legs are usually solid round steel bars. All members are galvanized and primed before erection. Television reception requires towers to be as high as possible to increase the area of coverage, and results in the use of the new high-strength steels to produce lighter-weight and taller towers. Field connections are made with galvanized bolts and lock-type nuts. The original design usually provides for increasing the height of the tower, when this is permitted by the Federal Communications Commission. *See* ANTENNA (ELECTROMAGNETISM); LAUNCH COMPLEX; TELEVISION TRANSMITTER; TRANSMISSION LINES.

Charles M. Antoni

Bibliography. D. A. Firmage, *Fundamental Theory of Structures*, 2d ed., 1980.

Towing tank

A tank of water used to determine the hydrodynamic performance of waterborne bodies such as ships and submarines, as well as torpedoes and other underwater forms. In the narrow sense, towing tanks are considered to be experimental facilities used to measure the forces, such as drag, on ship models and in turn to predict the performance of the full-scale prototype. In general, towing tanks are rectangular in planform with a uniform cross section. Different section shapes are used, ranging from rectangular to semicircular. The cross-section dimension may vary from about 8 to 52 ft (2.5 to 16 m) in width, from about 4 to 33 ft (1.5 to 10 m) in depth, and from under 100 ft (30 m) to almost 6560 ft (2000 m) in length; the size of the model varies in length from 4 to 30 ft (1.5 to 9 m).

Towing methods. The principal measurements made in a towing tank are force measurements, particularly drag or resistance of a towed ship model or other body. One of two principal systems for towing a model is used in most towing tanks. The simpler system consists of a gravity dynamometer and an endless cable attached to the model. A weight provides a constant towing force (**Fig. 1**). The time to traverse a fixed distance is measured when the model reaches a constant speed, thus establishing the speed-resistance relationship for the model. This dynamometer is simple and capable of high accuracy, but is limited to the measurement of the drag force of waterborne bodies. It is used in the smaller towing tanks in which the models are generally under 6 ft (2 m) in length.

In larger towing tanks the model is towed by a towing carriage mounted on rails at the side of the towing tank or suspended from an overhead track system (**Fig. 2**). Speed can be controlled and measured precisely on these carriages. Most carriages are equipped with a drag dynamometer as a permanent component (**Fig. 3**).

The dynamometer girder, mounted on the carriage frame, carries a long horizontal floating beam in pendulum fashion on two pairs of vertical arms terminating in flexible springs. A counterweight positioned at the upper end of a vertical swinging arm and mounted on the girder and attached to the floating beam maintains the beam in equilibrium at any position between the limit stops. The model resistance is transmitted as a horizontal force through the upper flexible link L_1 to the lower arm of the T-shaped balance, where it is balanced by the weight W. When the model resistance is not equal exactly to a unit weight W, the remainder is taken up (or applied) by the resiliency of the whole group of flexible spring supports; the exact amount of this auxiliary load or force is recorded on the drum by the link L_2 and the recording arm shown in Fig. 3.

In modern installations, the towing force is measured at the point where the towing link is connected to the towing post in the model. The measurements are usually made by measuring the strain in a flexure with strain gages, or by other means such as piezocrystals or inductance coils. The electronic signals are amplified and filtered and then sent to a computer for recording and analysis.

Law of similitude. Modern towing tank technology was established by William Froude in the 1870s, when he discovered the law of similitude for phenomena in which gravity is the predominating factor and established one of the essential principles of hydrodynamics for comparing model phenomena with the actual ship. There are three principal forces involved for a body moving through the water: inertia, gravity, and viscosity. The law of similitude requires that the ratio between inertia force and gravity force be the same for both the model and the prototype, and that the ratio between inertia force and viscous force be the same as well. *See* FROUDE NUMBER; REYNOLDS NUMBER.

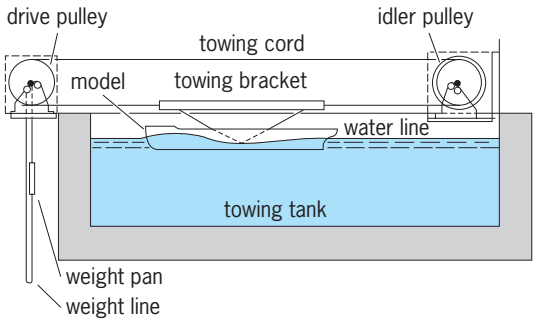


Fig. 1. Tank with model towed by falling weight.

Froude's law requires the validity of Eq. (1), where

$$\frac{V}{v} = \sqrt{\frac{L}{l}} = \sqrt{\lambda} \tag{1}$$

V and v represent the velocity of the prototype and model, respectively; L and l represent a characteristic dimension such as length of prototype and model, respectively; and λ represents the linear ratio of prototype to model.

Reynolds' law requires the validity of Eq. (2),

$$\frac{VL}{\nu_s} = \frac{vl}{\nu_m} \tag{2}$$

where ν_s and ν_m are the kinematic viscosity of the fluid for prototype and model, respectively. Because model towing tanks use fresh water and ships usually operate in seawater, ν_s and ν_m are essentially in the same order of magnitude. Thus, Froude's law requires a model velocity which is less than that of the prototype, whereas Reynolds' law requires a model velocity which is substantially higher than that of the prototype.

Froude overcame this difficulty by dividing the ship resistance into two parts: frictional resistance, and residuary resistance consisting mainly of wavemaking resistance. Frictional resistance is primarily the effect of viscosity and is thus governed by Reynolds' law. Residuary resistance is due mainly to gravitational effect and thus is governed by Froude's law. Based on this distinction, Froude developed the technique of predicting ship resistance from the resistance test of a scaled model, a technique used to this day. It can be outlined as follows. A model geometrically similar to the prototype is made and the drag or resistance measured in a towing tank at the corresponding speeds as expressed in Eq. (1). Viscous or frictional resistance r_F of the model is calculated by assuming the resistance to be the same as that of a smooth flat plate of comparable area and length. Residuary resistance r_R, which is largely a gravitational effect, is then found by subtracting the frictional resistance from the total measured resistance r_T as in Eq. (3).

$$r_R = r_T - r_F \tag{3}$$

Residuary resistance of the ship R_R is calculated by the law of comparison at corresponding speeds as in Eq. (4). The frictional resistance of the ship

$$R_R = r_R \times \lambda^3 \tag{4}$$

R_F is calculated on the same assumptions used in calculating the viscous resistance of the model. The total resistance of the ship is finally found by adding the residuary and frictional resistance as in Eq. (5).

$$R_T = R_F + R_R \tag{5}$$

Besides the resistance test, there are two other important tests: the propeller open-water test and the self-propulsion test. From the measuring of the revolutions per minute, torque, and thrust of the model propeller, the shaft horsepower required to drive the prototype at designed speed can be predicted. *See* DIMENSIONLESS GROUPS.

Force measurements. Experimentation in modern towing tanks has been extended far beyond predicting the resistance and power of a moving body in still water. Towing tanks are used to measure any combination of forces and moments upon a waterborne or submerged body under steady-state conditions. A few of the many possible types of tests involving force measurements in a towing tank are (1) lift and

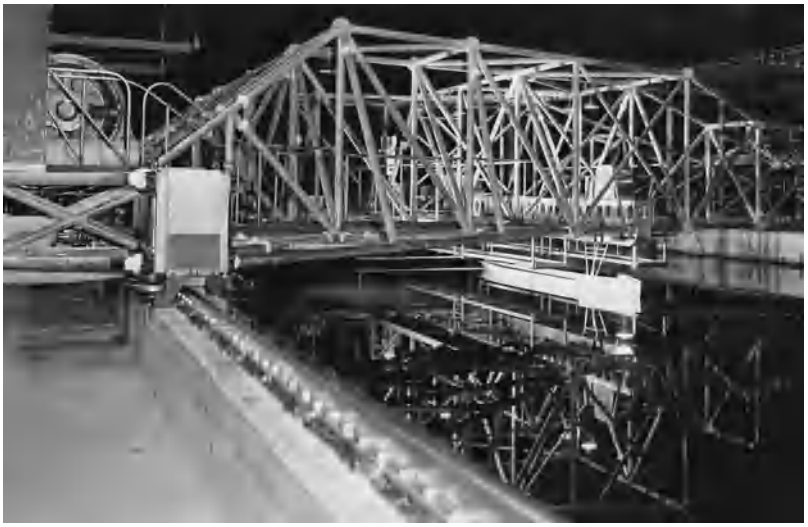


Fig. 2. Model in David Taylor Model Basin towed by carriage.

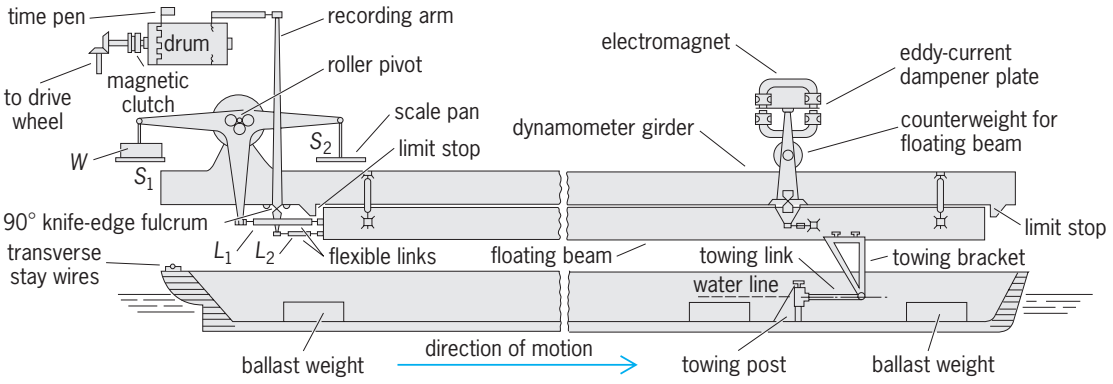


Fig. 3. Schematic arrangement of towing dynamometer.

drag of a planing surface; (2) lift, drag, and pitch moment of a submerged hydrofoil; (3) forces and moments on a submerged body towed at an angle of attack to obtain the static coefficients of a body for equations of motion; and (4) the turning moment on a ship's rudder.

Pressure and velocity measurement. A series of experiments has evolved from the measurement of pressures and velocities. The pressure measurements may be integrated over the surface to determine the force on the surface. Typical tests are the measurement of the pressure distribution on propeller blades, the duct of a ducted propeller, and the appendages of ship models. Various velocity-measuring devices have been developed, ranging from pitot tubes and hot-wire anemometers to sophisticated laser velocimetry. They are used to measure the velocity field at various locations around a model's hull, the most common of which is the plane in which the ship's propeller operates. *See* PITOT TUBE.

Lines of flow. A towing tank is also used to determine the lines of flow over portions of a hull; thus, appendages can be installed so that they will have minimum resistance and avoid or minimize the creation of cavitation. For surface ships, the conditions of comparable speed are maintained between model and prototype. Flow patterns on the hull are usually established by the emission of dyes upon a color-sensitive paint such as hydrogen sulfide upon a white lead-based paint. Flow patterns outside the boundary layer may be established by vanes or flags, which are free to pivot and orient themselves to the direction of flow. Typical tests are (1) establishment of the location of bilge keels; (2) orientation of shaft struts; and (3) determination of wave profile.

Wave experiments. Many towing tanks are provided with wavemakers, which extend the range of experimentation to the study of ship performance in head and following seas. At one end of the towing tank a wavemaker is installed which can generate waves of a more or less uniform profile with a predetermined height and length. At the other end of the tank a wave absorption beach is installed. Experiments are conducted on ship models at corresponding speeds in various head and following sea conditions, recognizing that the waves generated are much more regular than those encountered in the ocean. In addition to maintaining geometric similarity of the model and its prototype, the dynamic similarity of the system must be maintained. Measurements are made primarily of the motions of the body, particularly the rotational motions, pitch and roll, and the translational motions, heave (vertical) and surge (longitudinal). Some techniques involve free-running models with sensitive accelerometers. The carriage is used to carry the recording equipment and provide power to the model through flexible cables; these flexible cables do not exert any restraining force on the model. In some instances in connection with the measurement of motions, various components of force or moment may also be measured. *See* OCEAN WAVES.

Non-steady-state experiments. With the wide availability of electronic measuring instrumentation, the study of unsteady hydrodynamic phenomena has become increasingly important in towing tanks. Such fluctuating forces or pressures are measured as (1) pressure on a model's hull from propeller blades passing in proximity; (2) vibratory forces produced by a propeller operating in the variable velocity field behind a model; (3) route stability characteristics of a model from alternate course variations; and (4) forces and moments on a submerged body undergoing pure pitching and heaving motion. From the last measurement the coefficients for the equation of motion are obtained for a specific submarine configuration. This makes possible, through the use of the digital computer, the calculation of a submarine's motion without further experimentation and under a variety of conditions difficult to achieve in a towing tank. *See* DIGITAL COMPUTER.

Tank modifications. The towing tanks and the test equipment described have given rise to a number of special-purpose test facilities, which are modifications of the tanks described above. The most important of these are the maneuvering basin, the rotating-arm tank, seakeeping basins for both ships and offshore structures, ice tanks, and "vacuum" tanks for propulsion testing of large models. Maneuvering basins are used to measure the characteristics of ships pertaining to their maneuvering capability, such as turning radius, response to rudder deflections, and course stability. The rotating-arm tank is specialized in measuring the so-called rotary derivatives of a model. These derivatives link turning rate to forces and moments, and can be used for the determination of required rudder characteristics. Seakeeping basins are either towing tanks with a wave generation capability installed at one end, or a rectangular basin, with or without a carriage, with multiple wave flaps for the simulation of long-crested as well as short-crested irregular seas. The rectangular-basin type in particular has been instrumental in the explosive growth of the offshore oil industry. For fixed and floating offshore structures, the motion, anchoring forces, and internal forces are measured, including forces and moments so small that optical means must be sought to be able to measure them.

The most recent addition to the growing number of specialized towing tanks is the ice tank. These facilities may use actual ice or some other form of solid layer representing ice (in many cases a form of paraffin). They are used primarily to study the resistance characteristics of ships in ice, the effect of drift or pack ice on offshore structures, and the efficiency of ice breaking. *See* SHIP POWERING, MANEUVERING, AND SEAKEEPING; WATER TUNNEL (RESEARCH AND TESTING); SURFACE WAVES. Jacques B. Hadler

Bibliography. R. Bhattacharyya, *Dynamics of Marine Vessels*, 1978; D. A. Blank and A. E. Bock (eds.), *Introduction to Naval Engineering*, 2d ed., 1985; E. F. Gritzen, *Introduction to Naval Engineering*, 1980; E. Lewis (ed.), *Principles of Naval Architecture*, Society of Naval Architects and Marine Engineers, 3 vols., 1988; K. J. Rawson and E. C. Tupper, *Basic Ship Theory*, vol. 2, 4th ed., 1995.

Townsend discharge

A particular part of the voltage-current characteristic curve for a gaseous discharge device named for J. S. Townsend, who studied it about 1900. It is that part for low current where the discharge cannot be maintained by the field alone. Thus, if the agents producing the initial ionization were removed, conduction would cease.

In the lower end of this region, conduction is accomplished only by charges produced by external agents. As the electric field is increased, secondary ionization and more efficient collection of the primary ionization cause an increase in the current. After further increase in the field, the end of the Townsend region is reached. Any additional increase in the field causes a transition into a region where the discharge may be maintained by the field alone, whether it be glow, brush, or arc. *See* DARK CURRENT; ELECTRICAL CONDUCTION IN GASES; GLOW DISCHARGE.

Glenn H. Miller

Toxic shock syndrome

A serious, sometimes life-threatening disease usually caused by a toxin produced by some strains of the bacterium *Staphylococcus aureus*. The signs and symptoms are fever, abnormally low blood pressure, nausea and vomiting, diarrhea, muscle tenderness, and a reddish rash, followed by peeling of the skin.

Toxic shock syndrome was first reported in 1978 in seven pediatric patients. The disease, however, became prominent in 1980, when hundreds of cases were reported among young women without apparent staphylococcal infections. Epidemiologists observed that the illness occurred predominantly in young women who were menstruating and were using tampons, especially those that contained so-called superabsorbent synthetic materials. A toxin [toxic shock syndrome toxin number 1 (TSST-1)] that occurs in some strains of staphylococci has since been identified. These bacteria are known to proliferate in the presence of foreign particles in human infections, and it has been postulated that the tampons acted as foreign particles, allowing toxin-producing staphylococci to multiply in the vagina, where they usually compete with harmless bacteria.

Theoretically, millions of women were at risk, and yet the number of reported cases remained small. This suggests that those women who suffered from toxic shock syndrome were especially susceptible. Susceptibility may depend on their lacking antibodies to the toxin that occur in most adults. Several hundred cases of toxic shock syndrome not associated with menstruation have been reported. In these cases, which occurred in males as well as females, there was almost always an overt staphylococcal infection.

The toxin has been shown to occur in only about 1% of the staphylococcal strains studied. Moreover, there is some evidence that the syndrome may be caused also by other staphylococcal toxins, particularly enterotoxins. Cases of toxic shock syndrome that were caused by streptococci have been reported. A toxin distinct from TSST-1 appears involved. Persons with the symptoms of toxic shock syndrome should receive immediate medical care to reduce the chance of death. *See* STAPHYLOCOCCUS; TOXIN.

Jay O. Cohen

Bibliography. L. A. Cone et al., Clincial and bacterologic observations of a toxic shock-like syndrome due to *Streptococcus pyogenes, N. Engl. J. Med.*, 317:146–147, 1987; J. P. Davis et al., Toxic-shock syndrome: Epidemiologic features, recurrence, risk factors, and prevention, *N. Engl. J. Med.*, 303:1429–1435, 1980; A. L. Reingold et al., Toxic shock surveillance in the United States, 1980 to 1981, *Ann. Intern. Med.*, 96:875–880, 1982; K. N. Shands et al., Toxic-shock syndrome in menstruating women: Association with tampon use and *Staphylococcus aureus* and clinical features in 52 cases, *N. Engl. J. Med.*, 303:1436–1442, 1980; J. Todd et al., Toxic-shock syndrome associated with group-I staphylococci, *Lancet*, 2:1227–1229, 1978; W. Tyson et al., Atypical staphylococcal toxic shock syndrome: Two fatal cases, *Ped. Infect. Dis.*, 8:642–645, 1989.

Toxicology

The study of the adverse effects of chemical and physical agents on living organisms. Toxicology has also been referred to as the science of poisons. Since about 1970, there has been a substantial increase in the understanding of biological systems and the way in which chemicals can interfere with the proper functioning of such systems. At the same time, there has been a tremendous increase in the number and volume of chemicals that are in everyday use. Therefore, despite intensive research efforts, knowledge of fundamental toxicological principles and the specific effects of many individual compounds is incomplete. *See* POISON.

Factors that influence toxicity. The most important factor that influences the toxic effect of a specific chemical is the dose. All chemicals, including essential substances such as oxygen and water, produce toxic effects when administered in large enough doses. For a specific chemical, such as ethanol, the active ingredient of alcoholic beverages, the main question is the type of effects that this substance will produce and the doses at which these effects will occur. Another significant factor is the route of exposure. Living organisms may be exposed to a chemical by inhalation (into the lungs), ingestion (into the stomach), penetration through the skin, or, in special circumstances, injection into the body. In general, substances are absorbed into the body most efficiently through the lungs so that inhalation is often the most serious route of exposure. However, some chemicals do not vaporize very easily, so that other routes of exposure are more likely to occur and will be of greater toxicological significance.

A third factor is the fate of the chemical after the organism is exposed. The chemical may not be absorbed at all, limiting its possible adverse effects to the site of exposure, for example, a burn on the skin

or irritation of the throat. If it is absorbed, then it may travel throughout the body and has the potential to cause toxic effects at one or more sites remote from the site of entry. The remote sites where these adverse effects occur are called target organs.

A chemical that is absorbed and distributed throughout the body may be altered or metabolized, with the liver being the principal site of metabolism. There are two types of metabolic activity: useful chemicals are broken down into products (metabolites) that can be incorporated into the normal functioning of the organism; and unwanted chemicals are changed into forms that are more easily excreted from the body. The second type of activity can result in the formation of one or more new products, each of which may have greater or lesser toxicity than the chemical that was absorbed. When these products are less toxic, the process is known as detoxification. It serves to protect the organism from adverse toxic effects, but in some cases toxicity occurs as the dose is increased because the mechanisms for detoxification are overwhelmed.

The chemical and its metabolites can be excreted, stored, or transported in the organism and may, therefore, reach sites where toxic effects are induced. Substances that are rapidly excreted will have little opportunity to have adverse effects and thus are generally of low toxicity. Those that are excreted more slowly have the potential to cause long-term effects. Many substances are stored in the body, mainly in fat or bone, and thus can circulate throughout the organism for a long time.

Another significant variable is the time course of the exposure. A quantity of chemical administered at one time may have an effect even though the same quantity administered in small doses over time has no effect. If a chemical is completely excreted, then succeeding doses have no increased effect. However, if a residue remains, then it is possible for the second dose to add to the first and, if doses are repeated often enough, to lead to a level high enough to be toxic.

While the properties of a chemical are innate and unchangeable, those of the exposed organism may change over time. Individuals may be more susceptible at some life stages than others—for example, when they are very young or very old. The fetus is especially sensitive to chemical insult at certain developmental stages. In addition, previous exposures of the organism may make it more or less susceptible. There are also significant endogenous factors, such as the genetic makeup of the organism. Some species or individuals respond at very low doses, whereas others do not respond until very high levels are reached. The sex of the individual can also have an impact on susceptibility.

Classes of toxicity. In view of the importance of timing in producing adverse effects, toxicologists distinguish between two broad classes of toxicity, acute and chronic. Acute toxicity refers to effects that occur shortly after a single exposure or small number of closely spaced exposures. Chronic toxicity refers to delayed effects that occur after long-term repeated exposures.

Traditionally, the effect of most concern for acute toxicants (such as cyanide) is death. Consistent with this, acute toxicity is generally measured by using an assay to determine the lethal dose. In particular, rodents are given single doses and the number that have died 14 days later is recorded. The data are plotted for each dose, and the dose that is lethal for 50% of the animals (lethal dose 50 or LD_{50}) is used as the criterion for acute toxicity. *See* LETHAL DOSE 50.

Some synthetic chemicals, such as polychlorinated biphenyls (PCBs) and dichlorodiphenyltrichloroethane (DDT), exhibit their effects only after a number of repeated exposures and are considered chronic hazards, with cancer and reproductive effects being of greatest concern. To determine the dose at which chronic effects occur, rodents are exposed to daily doses of the chemical under study for long periods of time—from a few months to a lifetime. The highest dose at which no effects can be observed, the no observed effect level (NOEL), is used as a measure of chronic toxicity.

To study possible carcinogenic effects, rodents are exposed to very high doses over a lifetime in an attempt to detect low response in a small population of animals. Reproductive effects are studied by using special experimental protocols during which males and females are exposed during various periods of life, including the period of gestation for females. The experimental animals are utilized as surrogates for humans, and various techniques are employed to extrapolate results from rodents to humans. In many cases, the NOELs are divided by specific numbers, called safety factors, to estimate the levels below which effects are not likely to occur in humans. In other cases, particularly carcinogenicity, the high-dose results are extrapolated mathematically to low doses and, by using a scaling factor, to humans. These techniques have been designed to establish safe levels for humans.

Toxicology of metals. The basic principles of toxicity and toxicity assessment can be illustrated with the metals. These elements occur naturally, but the highest human exposures are usually of anthropogenic origin, especially when the metal is used in the workplace. Exposures can be due to other causes, however, such as burning coal or the use of pesticides to increase crop yields. The **table** lists metals that are toxicologically significant, but oversimplifies the situation since the types of toxic effects that occur often depend on a number of factors, such as the form of the metal (that is, inorganic or organic) and the specific type of chemical group to which the metal ion is bonded (for example, sulfate or phosphate). In addition, there are interactive effects so that deficiencies in levels of one metal may affect the toxicity of other metals.

The metals represent a very small fraction of the tens of thousands of chemicals in use, but there is much more information available about adverse effects in humans exposed to metals than about human toxicity resulting from most other chemicals in the environment. The metals that pose the greatest threats to public health are lead, mercury, and

Sources of exposure and toxicological properties of selected metals

Substance	Source of exposure	Acute effects in humans	Chronic effects in humans	Treatment
Arsenic	Occupational, pesticidal, food, drinking water, air	Fever, anorexia, skin lesions, multiorgan effects	Liver injury, peripheral vascular disease, cancer	Chelating agents for pulmonary and skin effects
Beryllium	Occupational, air (coal combustion)	Chemical pneumonia	Chronic pulmonary disease, cancer	Cease exposure
Cadmium	Occupational, food, air	Nausea and vomiting (ingestion), chemical pneumonia (inhalation)	Pulmonary disease, kidney damage, hypertension, bone fragility, cancer	Cease exposure
Chromium	Occupational, food	Respiratory irritation (inhalation)	Irritation, skin reactions, cancer	Cease exposure
Cobalt	Occupational, food	Vomiting, diarrhea, cardiomyopathy	Goiter, dermatitis, lung irritation	Cease exposure
Copper	Occupational, water	Nausea, vomiting, liver damage	Wilson's disease	Chelating agents
Gold	Medicinal	None reported	Dermatitis, kidney disease	Cease exposure
Lead	Food, water, air, lead paint, occupational	Vomiting, spasms	Neurological effects (especially in children), kidney damage, anemia	Chelating agents, removal of sources
Manganese	Occupational, food, air	Inflammation of the lung	Respiratory disease, neurological effects	L-Dopa
Mercury	Occupational, food, air	Bronchitis (inhalation); abdominal cramps, bloody diarrhea, and kidney damage (ingestion)	Neurological effects, kidney damage	Hemodialysis, chelating agents
Molybdenum	Occupational	Irritation	—	Cease exposure
Nickel	Occupational	Headache, nausea, vomiting, fever	Dermatitis, cancer	Cease exposure
Silver	Occupational	Gastrointestinal irritation	Argyria, kidney and lung damage	Cease exposure
Thallium	Occupational, pesticidal	Gastrointestinal irritation, neurological effects, cardiovascular effects	Hair loss, cataracts, neurological effects, kidney damage, loss of vision	Cease exposure
Zinc	Occupational, food, water, air	Abdominal distress, diarrhea (ingestion); fever and chills (inhalation)	Pulmonary effects	Cease exposure

arsenic, but toxic effects can be caused by other metals as well. *See* ENVIRONMENTAL TOXICOLOGY; MUTAGENS AND CARCINOGENS.

Lead. The most serious adverse effects of lead, mental retardation and learning problems, occur in young children subjected to chronic exposure, most often through ingestion of lead paints. Children with iron and calcium deficiencies absorb more lead and may suffer greater adverse effects. The principal effect of lead on adults is also neurological. The initial symptoms include nausea, vomiting, and joint pain, and at higher exposure levels may progress to toxic psychosis. Lead can also cause a variety of other effects, such as hypertension, anemia, kidney damage, and digestive problems. Lead acetate and lead phosphate are carcinogenic in animals, and it is possible that similar effects occur in humans. *See* LEAD.

Mercury. The toxicity of mercury is strongly dependent on its form. Mercury vapor causes bronchitis and pneumonitis, as well as damage to the central nervous system. Symptoms of the latter include excitability and tremors. Inorganic mercury poisoning is characterized by ulceration, bleeding, and necrosis of the gastrointestinal tract followed by severe renal effects. Chronic exposures may also result in kidney damage.

Organic mercury, particularly methylmercury, is the most important public health threat. Consumption of fish or grain containing methylmercury has led to significant epidemics in Japan, Iraq, and Pakistan, as well as lesser incidents in other countries. The major symptoms are neurological and include numbness, lack of coordination, speech impairment, and deafness. These reflect lesions in both the cerebellum and cerebral cortex. Organic mercury can also be toxic to the developing fetus. *See* MERCURY (ELEMENT).

Arsenic. The third metal most often implicated in human toxicity is arsenic. In this case, the valence form is critical to its toxicity. Trivalent arsenic, as in arsenic trioxide (As_2O_3), is usually more toxic to mammals than is pentavalent arsenic, as in arsenic pentoxide (As_2O_5).

The acute effects of high doses of arsenicals are broad and impact the upper respiratory tract and the cardiovascular, gastrointestinal, hematopoietic, and central nervous systems. A characteristic symptom resulting from ingestion is a pattern of skin abnormalities, including hyperpigmentation. Chronic, low-level exposures lead to more subtle effects on the gastrointestinal tract, nervous system, and skin. Continued intake can lead to degeneration of the liver and kidneys. Chronic exposure to arsenic is associated with cancers of the skin and lung and may be linked to cancers of internal organs. *See* ARSENIC.

Other metals. Toxic effects due to exposure to other metals may be of concern in situations where occupational or environmental exposure is high. These

other metals will be considered in alphabetical order.

1. *Beryllium*. Acute inhalation of high levels of beryllium causes severe inflammation of the upper respiratory tract and lungs. Exposure to lower doses can lead to more insidious disease, developing over weeks, with less severe pulmonary damage. A less toxic but more common effect is delayed-type allergic contact dermatitis. Chronic exposure can also lead to severe respiratory disease as well as to enlargement of organs such as the spleen and liver. Such effects may appear after a latency period of up to 20 years. Beryllium has been shown to cause cancer in laboratory animals, but this has not been confirmed in humans.

2. *Cadmium*. Exposure to high levels of cadmium over short periods of time leads to nausea, vomiting, and cramps. Inhalation of cadmium fumes can produce acute pneumonitis and pulmonary edema. Chronic exposure can result in kidney damage (renal tubular disease), chronic obstructive pulmonary disease, bone fragility, and emphysema. Ingested cadmium is carcinogenic in animals, but such carcinogenicity has not been confirmed in humans.

3. *Chromium*. Chromium intoxication depends on the valence form, with the hexavalent compounds being more toxic. These hexavalent compounds are irritating and corrosive but are slowly absorbed, so that chronic effects are localized, often limited to skin ulceration and gastrointestinal tract irritation. Chronic inhalation of chromium is associated with cancer of the lung in humans.

4. *Cobalt*. Cobalt is an essential nutrient and of low toxicity at usual levels. However, a number of fatalities due to cardiomyopathy occurred in people consuming beer containing a cobalt additive. Very high acute exposures cause vomiting and diarrhea, and high chronic exposures are linked to goiter. In addition, chronic occupational exposures may lead to respiratory irritation and dermatitis.

5. *Copper*. Copper is also an essential nutrient, and will not produce intoxication unless there is exposure to high levels. Acute symptoms of copper ingestion are nausea and vomiting due to vascular congestion of the gastrointestinal tract. Very high doses can lead to death from vascular collapse. In addition, liver necrosis may result in such cases. Chronic exposure to low levels of copper does not cause adverse effects except in rare individuals whose genetic makeup leads to excess copper accumulation (Wilson's disease).

6. *Gold*. Gold is toxic only after long exposures such as those associated with chronic intake of gold used medicinally. Effects include dermatitis, stomatitis, and injury to kidney tubules.

7. *Manganese*. Manganese toxicity has been seen only in miners and mill workers. Acute exposure can lead to a lung disease known as manganese pneumonitis. Chronic exposure can result in severe effects on the central nervous system, effects ascribed to interference with metabolism of amines. Symptoms include psychotic episodes, speech impairment, hypokinesia, and liver cirrhosis.

8. *Molybdenum*. Human poisoning due to molybdenum, an essential mineral, has rarely been observed and appears limited to irritation at the sites of exposure.

9. *Nickel*. The most serious effect of nickel exposure is lung and nasal cancer associated with chronic inhalation by refinery workers. Acute effects from ingestion of high doses of nickel include headache, nausea, and vomiting as well as pulmonary symptoms. Continued exposure can lead to lesions in lung and brain. The most common effect of low-level exposure in the general population is contact dermatitis.

10. *Silver*. The toxic effects of elemental silver are limited since it is poorly absorbed. Ingestion of very large amounts of silver can lead to acute gastroenteritis. Chronic occupational exposure can lead to argyria, a condition characterized by local or generalized changes in skin pigmentation. In rare instances, the respiratory tract and kidneys are also affected.

11. *Thallium*. The principal symptoms of acute thallium poisoning are gastroenteritis, headache, and rapid heartbeat followed by neurological effects such as paralysis and psychic disturbances. Lower levels of exposure can lead to lack of coordination, numbness, and tremor. Chronic effects include liver necrosis, nephritis, cataracts, neurological damage, and pulmonary edema. Hair loss is a characteristic symptom of long-term thallium exposure.

12. *Zinc*. The most significant toxic effect of zinc is zinc fume fever, fever and chills that can result from acute inhalation of zinc oxide fumes. In addition, ingestion of high doses of soluble salts of zinc can lead to acute gastroenteritis. Chronic inhalation of zinc can cause pulmonary irritation. *See* BERYLLIUM; CADMIUM; CHROMIUM; COBALT; COPPER; GOLD; MANGANESE; MOLYBDENUM; NICKEL; SILVER; THALLIUM; ZINC. Michael Kamrin; Robert W. Leader

Bibliography. L. Friberg et al. (eds.), *Handbook of the Toxicology of Metals*, 2d ed., 1986; M. A. Kamrin, *Toxicology: A Primer on Toxicology Principles and Applications*, 1988; C. D. Klaasen, M. O. Amdur, and J. Doull (eds.), *Casarett and Doulls' Toxicology*, 3d ed., 1986; F. C. Lu, *Basic Toxicology*, 1985.

Toxin

Properly, a poisonous protein, especially of bacterial origin. However, nonproteinaceous poisons, such as fungal aflatoxins and plant alkaloids, are often called toxins. *See* AFLATOXIN; ALKALOID.

Bacterial exotoxins. These proteins of disease-causing bacteria are usually secreted and have deleterious effects. Several hundred are known that account to a variable extent for the damage caused by bacterial diseases and infections. In some extreme cases a single toxin accounts for the principal symptoms of a disease, such as diphtheria, tetanus, and cholera. Bacteria that cause local infections with pus (such as *Streptococcus pyogenes*, *Staphylococcus aureus*, and *Clostridium perfringens*) often produce many toxins that affect the tissues around the

infection site or are distributed to remote organs by the blood. *See* CHOLERA; DIPHTHERIA; STAPHYLOCOCCUS; TETANUS.

Toxins may assist the parent bacteria to combat host defense systems, to increase the supply of certain nutrients such as iron, to invade cells or tissues, or to spread between hosts. Sometimes the damage suffered by the host organism has no obvious benefit to the bacteria. For example, botulinal neurotoxin in spoiled food may kill the person or animal that eats it long after the parent bacteria have died. In such situations it is assumed that the bacteria benefit from the toxin in some other habitat and that the damage to vertebrates is accidental. In complex situations the role played by a toxin in disease or in assisting the bacteria is best studied by comparing the courses of artificial infections established with genetically matched pairs of bacteria that differ only in the ability to synthesize a single toxin. *See* FOOD POISONING.

Intracellular action sites. Certain bacterial and plant toxins have the unusual ability to catalyze chemical reactions inside animal cells. Such toxins are always composed of two functionally distinct parts termed A and B, and they are often called A-B toxins. The B part binds to receptor molecules on the animal cell surface and positions the toxin upon the cell membrane. Subsequently, the enzymically active A portion of the toxin crosses the animal cell membrane and catalyzes some intracellular chemical reaction that disrupts the cell physiology or causes cell death. A and B may be adjacent regions of a single polypeptide chain (for example, in diphtheria toxin and *Pseudomonas aeruginosa* exotoxin A); two protein chains linked by disulfide bonds (for example, the plant toxins abrin and ricin); a complex of separate protein subunits encoded by different genes (for example, cholera toxin, pertussis toxin, and shigella toxin); or two unlinked proteins that only come together on the target cell's surface (for example, anthrax toxin and botulinal C2 toxin). The last type are also called binary toxins.

About one-half of the known A-B toxins catalyze a class of chemical reaction known as adenosine diphosphate (ADP) ribosylation whereby key cellular enzymes acquire a large chemical tag, that is, ADP-ribose, that disturbs their functions. Many of the target enzymes are proteins that bind guanosine triphosphate (GTP), called G proteins, and normally function in the regulation of cell physiology, shape, and growth.

Several of the A-B toxins increase the cell content of an intracellular messenger molecule, cyclic adenosine monophosphate (cyclic AMP). In some cases the toxins stimulate the activity of the target cell's enzyme adenylate cyclase which synthesizes cyclic AMP. In other cases the toxin itself is an adenylate cyclase. In phagocytes an increased cyclic AMP level translates into reduced migration to the site of infection, reduced phagocytosis, and reduced intracellular killing of bacteria.

When the intracellular reaction of an A-B toxin blocks protein synthesis, the target cell dies. Diphtheria toxin's fragment A inactivates a soluble protein key in protein synthesis: one molecule of fragment A is enough to kill an animal cell in this way. Several A-B toxins of plant origin, such as abrin, ricin, and viscumin, inactivate ribosomes. Other plant seeds, including wheat, barley, rye, and corn, contain A chains similar to the A chain of ricin, but no B chain. These A chains inactivate ribosomes in the test tube but have very limited abilities to enter cells and fortunately are not toxic. Much research has been directed at the formation of artificial hybrids containing the A portion of diphtheria toxin or one of the plant proteins, combined with a molecule that binds only to cancer cells such as a surface-specific antibody. In certain cases these hybrid immunotoxins are able to selectively kill tumor cells. *See* IMMUNOLOGIC CYTOTOXICITY.

Membrane damage. A large group of toxins (more than 200 are known) breach the normal barrier to free movement of molecules across cell membranes. In sufficient concentration such cytolytic toxins cause cytolysis, a process by which soluble molecules leak out of cells, but in lower concentration they may cause less obvious damage to the cell's plasma membrane or to its internal membranes. The cytolytic agents are commonly detected by their ability to lyse red blood cells. Some are enzymes that hydrolyze cell membrane phospholipids, such as *Clostridium perfringens* alpha toxin and *Staphylococcus aureus* beta toxin; others insert themselves into the membrane and form channels through which small molecules may pass, such as *Staphylococcus aureus* alpha and delta toxins, mellitin of bee venom, and streptolysin O and streptolysin S of *Streptococcus pyogenes*. *See* CELL MEMBRANES.

Neurotoxins. Tetanus and botulinal neurotoxins owe their extraordinary potencies to the fact that their target cells are neurons; both toxins block the transmission of nerve impulses across synapses. Tetanus toxin is transported by axons to the central nervous system where the blockage, mainly of inhibitory synapses, results in spastic paralysis, in which opposing muscles contract simultaneously. The botulinal neurotoxins principally paralyze neuromuscular junctions and cause flaccid paralysis, in which the muscles cannot contract. Snake and invertebrate venoms often contain less potent neurotoxins, such as the crotoxin of the Brazilian rattlesnake and the bungarotoxins produced by the Formosan banded krait. Also, many poisonous species contain low-molecular-weight (nonprotein) substances that have specific effects on the nervous system and are useful in brain research, such as tetrodotoxin of the puffer fish, black widow spider venom, and the algal saxitoxin of red tides.

Fever-producing toxins. Gram-negative bacteria, such as *Salmonella* and *Hemophilus*, have a toxic component in their cell walls known as endotoxin or lipopolysaccharide. Among other detrimental effects, endotoxins cause white blood cells to produce interleukin-1 (formerly known as endogenous pyrogen), a hormone responsible for fever, malaise,

Approximate median lethal doses of some toxins per kilogram of body weight

Toxin	Dose	Test animal
Tetanus	1 nanogram	Mouse, probably human
Botulinal neurotoxin	1 nanogram	Mouse, human
Shigella	1 nanogram	Monkey, human
	1 microgram	Mouse
Diphtheria	100 nanograms	Human
	1.6 milligrams	Mouse
Ricin	1 microgram	Human
Staphylococcus alpha	50 micrograms	Mouse
Neurotoxins of snakes and invertebrates	50 micrograms	Mouse

headache, muscle aches, and other nonspecific consequences of infection. The exotoxins of toxic shock syndrome and of scarlet fever induce interleukin-1 and also tumor necrosis factor, which has similar effects. *See* ENDOTOXIN; FEVER; SCARLET FEVER; TOXIC SHOCK SYNDROME.

Immunity to toxins. Toxoids are toxins that have been exposed to formaldehyde or other chemicals that destroy their toxicities without impairing immunogenicity. When injected into humans, toxoids elicit specific antibodies known as antitoxins that neutralize circulating toxins. Such immunization (vaccination) is very effective for systemic toxinoses, such as diphtheria and tetanus. It has resulted in the virtual eradication both of diphtheria and of its causative organism, *Corynebacterium diphtheriae*. Immunization with a toxoid is less successful in controlling diarrheal diseases caused by toxins produced within the intestine that are not exposed to antitoxins in the blood. *See* ANTIBODY; DIARRHEA; DIPHTHERIA; IMMUNITY; VACCINATION.

Relative toxicities. Toxicities vary with the route of administration and can differ greatly with the test animal. It may be difficult to predict human toxicities solely from measurements with other animals. For example, mice are many-thousand-fold less sensitive than humans to diphtheria toxin because their cells lack a surface component needed for the entry of fragment A (see **table**).

D. Michael Gill

Bibliography. R. J. Collier and D. A. Kaplan, Immunotoxins, *Sci. Amer.*, 251:56–64, 1984; D. M. Gill, Bacterial toxins: A table of lethal amounts, *Microbiol. Rev.*, 46:86–94, 1982; R. F. Keeler, N. Mandara, and A. T. Tu, *Natural Toxins: Toxicology, Chemistry and Safety*, 1992; A. I. Laskin and H. A. Lechevalier (eds.), *Handbook of Microbiology*, vol. 8, 1987; J. L. Middlebrook and R. B. Dorland, Bacterial toxins: Cellular mechanisms of action, *Microbiol. Rev.*, 48:199–221, 1984; C. A. Mims, *Pathogenesis of Infectious Disease*, 5th ed., 2001.

Toxin-antitoxin reaction

In serology, the combination of a toxic antigen with its corresponding antitoxin. If the antitoxin is derived from any species other than the horse, precipitation occurs over a wide range of reactant ratios, as in other antigen-antibody reactions. With horse antitoxin, flocculation occurs only if toxin and antitoxin are near equivalence, a twofold excess of either reactant giving soluble complexes. In most instances, the reaction results in partial or complete neutralization of the toxic activity of the antigen. *See* ANTIGEN; ANTITOXIN; NEUTRALIZATION REACTION (IMMUNOLOGY); TOXIN.

The Danysz reaction occurs when an exact equivalence of toxin is added to antitoxin, not in one portion, but in successive increments. The mixture will remain somewhat toxic, although it becomes neutral within a few days. This is the result of the variable combining proportions of antigen and antibody, which result in a binding of excess antibody by the initial increments of toxin giving soluble molecular complexes composed of one molecule of toxin and eight molecules of antibody (TA_8). As a result, some of the subsequent increments of toxin remain free. In time, the system rearranges to give the neutral equivalence composition, one molecule of toxin and two of antitoxin (TA_2). *See* ANTIBODY; SEROLOGY.

Henry P. Treffers

Toxoplasmea

A class of the subphylum Sporozoa. Inadequate information about the structure and life cycles of the members of this class has, for many years, resulted in the organisms being classified as parasites of unknown nature. Electron microscope studies have shown structural similarities between this group and the Telosporea, but characteristics of the Telosporea such as the presence of spores (oocysts) and sexual reproduction have only recently been reported for the Toxoplasmea. These findings will necessitate reclassification of the Toxoplasmea. *See* SPOROZOA; TELOSPOREA; TOXOPLASMIDA.

Morphology. The organisms are small and crescent-shaped. They move by body flection or gliding and have no flagella or pseudopodia. Characteristic structures are the two-layered pellicle and underlying longitudinal microtubules, micropyle, a conoid, paired organelles, and micronemes.

The most distinguishing characteristic of the Toxoplasmea is the unique means of reproduction. Although reproduction by binary fission has been reported for these organisms, electron microscope studies indicate that endodyogeny is the

sole method. Endodyogeny is an internal budding wherein two daughter cells are produced within a mother cell which is destroyed in the process.

Life cycle. Only two stages are known in most animals. One stage, the proliferative form or trophozoite, occurs singly or in groups within host cells. The other stage, the so-called cyst, consists of a large number of organisms which, with minor differences, are structurally similar to the proliferative forms.

The life cycle of one member of this group has recently been described. Although proliferative forms and cysts of *Toxoplasma gondii* are found in many animals, the complete life cycle has been found only in the cat. A resistant *Isospora*-like oocyst can transmit the organism from cat to cat. Asexual and sexual stages develop in the intestinal epithelium. Further study may elucidate similar life cycles for other members of the Toxoplasmea. *See* PROTOZOA.

Harley G. Sheffield

Bibliography. K. M. Adam et al., *Medical and Veterinary Protozoology*, rev. ed., 1980; J. N. Farmer, *The Protozoa: Introduction to Protozoology*, 1980; R. R. Kudo, *Protozoology*, 5th ed., 1977; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Toxoplasmida

An order of the class Toxoplasmea. Four genera, *Toxoplasma, Besnoitia, Sarcocystis*, and *Encephalitozoon*, make up the order. The organisms are parasites of vertebrates. *Toxoplasma* is often found encysted in nerve tissue, *Besnoitia* in connective tissue, and *Sarcocystis* in muscle. Very little is known about *Encephalitozoon*, but the parasite has been found in the brain of rabbits. *See* TOXOPLASMEA.

Only one or two stages of the life cycle of most members are known. A proliferative form may be found singly or in small groups within host cells. The other stage, the cyst, is found intra- or extracellularly.

Typical structures appear in the proliferative form of *T. gondii*. The organism is surrounded by a two-layered pellicle. A micropyle is located in the pellicle lateral to the nucleus. A cone-shaped structure, the conoid, is situated at the anterior end. Long clublike paired organelles extend posteriorly from the conoid area. The nucleus and adjacent Golgi zone are in the posterior half of the organism. Mitochondria, granules, vesicles, and ribosomes fill the cytoplasm.

Toxoplasma produces a serious, although rare, disease in humans. Toxoplasmosis occurs throughout the world. Serological studies indicate that many people have been exposed to the organism, although clinical disease is rare. Congenital toxoplasmosis may result in abortion or deformity, and ocular toxoplasmosis may destroy the eye.

While the mechanism of transmission of *Sarcocystis, Besnoitia*, and *Encephalitozoon* is unknown, that of *Toxoplasma* has been described. This parasite is a coccidian of cats, although proliferative forms and cysts have been found in many animals, including humans. The exogenous form is an *Isospora*-like oocyst which is passed in the feces. The oocyst is infective to cats, to most laboratory animals, and perhaps to humans. Probably the more common means of human infection is through ingestion of raw or poorly cooked meat.

Harley G. Sheffield

Bibliography. R. L. Barnes, *Invertebrate Zoology*, 6th ed., 1994; J. A. Pechenik, *Biology of the Invertebrates*, 4th ed., 2000.

Trace fossils

Fossilized evidence of animal behavior, also known as ichnofossils, biogenic sedimentary structures, bioerosion structures, or lebensspuren. The fossils include burrows, trails, and trackways created by animals in unconsolidated sediment, as well as borings, gnawings, raspings, and scrapings excavated by organisms in harder materials, such as rock, shell, bone, or wood. Some workers also consider coprolites (fossilized feces), regurgitation pellets, burrow excavation pellets, rhizoliths (plant root penetration structures), and algal stromatolites to be trace fossils. *See* STROMATOLITE.

Trace fossils are important in paleontology and paleoecology, because they are fossils that provide information about the presence of unpreserved soft-bodied members of the original communities, life habits of fossil organisms, evolution of certain behavior patterns through geologic time, and biostratigraphy of otherwise unfossiliferous deposits. Trace fossils also are useful in sedimentology and paleoenvironmental studies, because they are sedimentary structures that are preserved in place and are very rarely reworked and transported, as body fossils of animals and plants commonly are. This fact allows trace fossils to be regarded as reliable indicators of original conditions in the sedimentary environment. The production of trace fossils involves disruption of original stratification and sometimes results in alteration of sediment texture or composition.

Occurrence. Trace fossils occur in sedimentary deposits of all ages from the late Precambrian to the Recent. Host rocks include limestone, sandstone, siltstone, shale, coal, and other sedimentary rocks. These deposits represent sedimentation in a broad spectrum of settings, ranging from subaerial (such as eolian dunes and soil horizons) to subaqueous (such as rivers, lakes, swamps, tidal flats, beaches, continental shelves, and the deep-sea floor). *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS.

Preservation. Organisms may produce fossilizable traces on the sediment surface (epigenic structures) or within the sediment (endogenic structures). Trace fossils may be preserved in full three-dimensional relief (either wholly contained within a rock or weathered out as a separate piece) or in partial relief (either as a depression or as a raised structure on a bedding plane). Simply because a trace fossil is preserved on a bedding plane does not indicate that it originally was an epigenic trace.

Adolf Seilacher, a German paleontologist, proposed the following terminology to describe the

Fig. 1. Domichnial dwelling burrow with pelleted wall, probably created by a burrowing crustacean. *Ophiomorpha*, Cretaceous, North Dakota. (*From W. Häntzschel, in C. Teichert, ed., Treatise on Invertebrate Paleontology, pt. W, revised, University of Kansas Press, 1975*)

preservational mode of trace fossils: full relief (enclosed entirely within the sediment); positive epirelief (ridge, mound, or other raised structure on the upper surface of a bed); negative epirelief (groove, pit, or other indentation on the upper surface of a bed); positive hyporelief (raised structure on the sole of a bed); and negative hyporelief (indentation on the sole of a bed). Anders Martinsson, a Swedish paleontologist, proposed an alternative system to characterize the preservation of trace fossils: endichnia (enclosed entirely within the sediment); epichnia (exposed in relief on the upper surface of a bed); hypichnia (exposed in relief on the sole of a bed); and exichnia (entirely removed from the sediment in which it was produced).

Diagenetic alteration of sediment commonly enhances the preservation of trace fossils by differential cementation or selective mineralization. In some cases, trace fossils have been preferentially replaced by chert, dolomite, pyrite, glauconite, apatite, siderite, or other minerals. *See* DIAGENESIS.

Classification and nomenclature. The study of trace fossils is known as ichnology. The prefix "ichno-" (as in ichnofossil and ichnotaxonomy) and the suffix "-ichnia" (as in epichnia and hypichnia) commonly are employed to designate subjects relating to trace fossils. The suffix "-ichnus" commonly is attached to the ichnogenus name of many trace fossils (as in *Dimorphichnus* and *Teichichnus*).

In the nineteenth century, many trace fossils were mistakenly identified as fossil plants, so they were given the genus and species names of plants in accordance with established principles of Linnean taxonomy. Subsequent recognition that these fossils actually were biogenic sedimentary structures that represented animal activity in the sediment has not stopped the practice of assigning formal taxonomic names to trace fossils. Usually, geologists differentiate between trace fossils and body fossils by the terms "ichnogenus" and "ichnospecies" when speaking of trace fossils. Rules of the International Code of Zoological Nomenclature generally apply to trace fossils at the genus and species level. Although some workers have proposed higher taxonomic levels for trace fossils, such as ichnofamilies or ichnophyla, none of these have gained universal acceptance. *See* TAXONOMY.

Fig. 2. Highly branched, fodinichnial mining burrow, possibly produced by a deposit-feeding crustacean or worm. (*a*) *Chondrites*, Tertiary, Austria. (*b*) *Chondrites*, Tertiary, Spain. (*From W. Häntzschel, in C. Teichert, ed., Treatise on Invertebrate Paleontology, pt. W, revised, University of Kansas Press, 1975*)

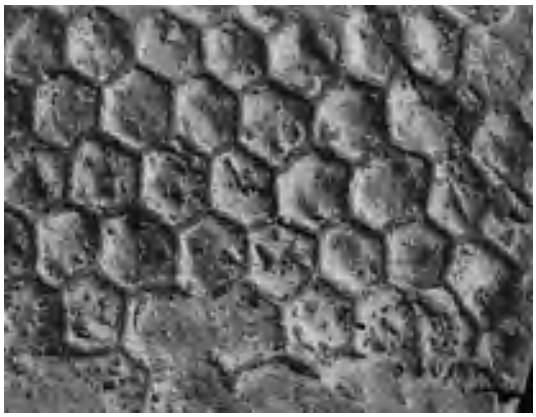


Fig. 3. Agrichnial farming trace of an unknown organism, composed of a hexagonal, meshlike network of tunnels. *Paleodictyon*, Tertiary, Poland. (*Photograph by W. Häntzschel*)

Fossil behavior. Trace fossils provide tangible information of the activities of ancient organisms, because they represent particular behavior patterns related to dwelling, feeding, locomotion, and resting. Ichnologists have established several behavioral categories of trace fossils, including the following eight groups that are most widely recognized today (**Figs. 1–10**): domichnia (dwelling structures, such as permanent burrows or agglutinated tubes); fodinichnia (burrows produced in the process of mining the sediment for food); agrichnia (burrows produced in order to farm or trap food inside the sediment); praedichnia (traces of predation); pascichnia (feeding trails, either within the sediment or on the sediment surface); repichnia (locomotion trails and trackways); fugichnia (escape traces, usually produced by an animal crawling out from beneath a rapidly deposited pile of sediment); and cubichnia (resting or nesting traces).

Fig. 4. Agrichnial farming traces of unknown organisms, including a double-spiral tunnel (*Spirorhaphe*) and a meshlike network of tunnels (*Paleodictyon*). Tertiary, Austria. (*Photograph by W. Häntzschel*)

Fig. 5. Loosely meandering, agrichnial farming trace of an unknown organism. *Cosmorhaphe*, Tertiary, Poland. (*From W. Häntzschel, in C. Teichert, ed., Treatise on Invertebrate Paleontology, pt. W, revised, University of Kansas Press, 1975*)

Environmental implications. Trace fossils are useful to geologists as indicators of ancient environments of deposition. Recurrent assemblages of trace fossils that represent certain environmental conditions, such as water depth, salinity, or character of the sea floor, are known as ichnofacies. Ichnofacies are named after common ichnogenera that exemplify this association. *See* FACIES (GEOLOGY).

Adolf Seilacher established a bathymetric zonation of universal ichnofacies representing unconsolidated sediments in marine settings, which can be found throughout the geologic column all over the world. The *Skolithos* ichnofacies typically represents nearshore, often intertidal, environments characterized by well-sorted clastic sediments that are dominated by primary sedimentary structures. Most of the trace fossils are domichnia, repichnia, and fugichnia. The *Cruziana* ichnofacies represents offshore settings, generally within wave base (that is, in water shallow enough for waves to move sediment grains on the sea floor). The trace fossils include an abundance of domichnia, repichnia, fugichnia, cubichnia, pascichnia, or fodinichnia. The *Zoophycos* ichnofacies usually represents quiet-water conditions far from shore, often on a submarine slope. The trace fossils are characterized by a low-diversity assemblage of fodinichnia. The *Nereites* ichnofacies represents fine-grained, distal turbidites that were deposited in relatively deep water. The trace fossils consist mainly of pascichnia and agrichnia. Another deep-water trace fossil association, simply referred to as the deep-sea ichnofacies, represents pelagic sedimentary environments, and it is characterized by a moderate-diversity assemblage of fodinichnia that were deeply emplaced within the sediment. *See* MARINE SEDIMENTS; TURBIDITE.

Other marine ichnofacies have been established for substrates that were not unconsolidated sediment. The *Glossifungites* ichnofacies represents very firm substrates (highly compacted but uncemented sediment). The *Trypanites* ichnofacies represents fully lithified substrates (cemented

Fig. 6. Praedichnial boring drilled in a bivalve shell by a carnivorous gastropod. *Oichnus*, Recent, Texas. (*Photograph by A. A. Ekdale*)

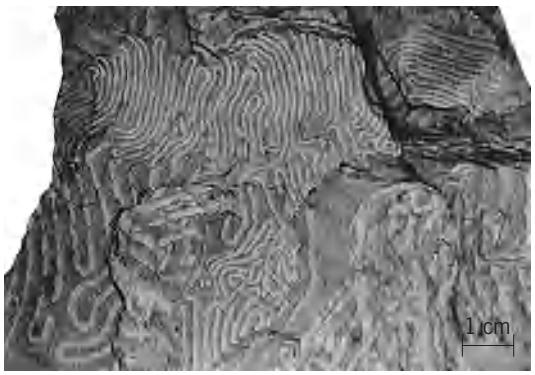


Fig. 7. Tightly meandering, pascichnial grazing trail, created by an unknown worm. *Helminthoida*, Tertiary, Austria. (*From W. Häntzschel, in C. Teichert, ed., Treatise on Invertebrate Paleontology, pt. W, revised, University of Kansas Press, 1975*)

Fig. 8. Repichnial trail of an arthropod, possibly a trilobite. *Cruziana*, Cambrian, Alberta. (*Photograph by J. P. A. Magwood*)

sedimentary rock and calcareous shell material). The *Teredolites* ichnofacies represents wood substrates. In all three cases, the dominant trace fossils are domichnia, namely borings, which were excavated by organisms with very special adaptations for penetrating the harder substrates.

Seilacher grouped all trace fossil associations in continental settings into one ichnofacies, the *Scoyenia* ichnofacies. Although it is widely recognized that several nonmarine ichnofacies exist in the geologic record, no precise delineation of these varied ichnofacies has achieved universal acceptance. Some workers have established various local ichnofacies that have yet to be accepted as having universal application.

Ichnofabric. The activities of burrowing and boring organisms can profoundly affect many aspects of the texture and internal structure of a sedimentary deposit, as sediment grains are sorted, modified, and redistributed by infaunal (living within the sediment below the sediment surface) animals. Sediment fabrics that result from bioturbation (vertical mixing of sediment by the burrowing activities of animals) and bioerosion activities are called ichnofabrics. In many situations, such as in pelagic carbonate deposits, continuous sedimentation and simultaneous bioturbation allow for the superimposition of different suites of organism traces, thus producing composite ichnofabrics (**Fig. 11**). Ichnofabric analysis can shed light on the paleoecology of the infaunal community, including trophic relationships and tiering structure. Ichnofabrics also allow the stability and firmness of the original substrate to be interpreted by examining the distinctness and degree of deformation of trace fossils. Interstitial oxygen conditions in the original sediment may be deciphered from the abundance and preservational modes of deep-tier burrows. Early diagenetic processes, including differential cementation and secondary mineralization, can enhance the preservation and visibility of ichnofabrics.

Sedimentologic implications. Trace fossils reflect the interplay among the three important sedimentologic processes of deposition, erosion, and burrowing of the sediment by organisms (causing disturbance or obliteration of primary stratification). Slow, continuous deposition, as occurs in the oxygen-rich water of offshore shelf and deep-sea environments, usually is accompanied by a total burrowing of the sediment. Numerous trace fossils, especially domichnia, fodinichnia, and pascichnia, characterize such situations. In contrast, slow, continuous deposition in eutrophic lakes or restricted marine basins that contain oxygen-depleted bottom water will yield laminated sediment without trace fossils, because the available oxygen is insufficient to support bottom-dwelling animals. Rapid, continuous deposition, as occurs in prograding beaches and laterally accreting point bars, commonly is reflected by sparse trace fossils (mainly fugichnia, cubichnia, and domichnia) superimposed on a sedimentary fabric of primary bed forms and primary sedimentary structures.

Discontinuous deposition usually results from the alternation of slow and rapid sedimentation events, as exemplified by turbidites, or from the alternation of rapid sedimentation and erosion events, as occurs during major storms along a marine coastline. In the former case, the fine-grained units of a turbidite represent a lengthy period of slow deposition, during which numerous kinds of organisms lived on and

Fig. 9. Unnamed escape structure (fugichnial trace) in sandstone. Cretaceous, Utah. (*Photograph by A. A. Ekdale*)

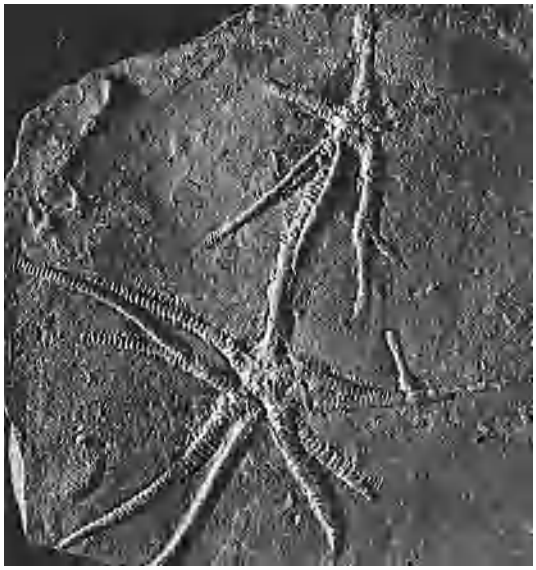


Fig. 10. Cubichnial resting trace of an ophiuroid brittle star. *Asteriacites*, Jurassic, Germany. (*Photograph by W. Häntzschel*)

Fig. 11. Complex, composite ichnofabric created by numerous successive phases of burrowing in a fine-grained pelagic carbonate deposit, which has been weakly cemented to become chalk. Upper Cretaceous, Denmark. (*Photograph by A. A. Ekdale and R. G. Bromley*)

in the sea floor, creating a wide variety of fossilizable traces (especially pascichnia, fodinichnia, and agrichnia). The coarser-grained unit of the same turbidite represents sudden deposition by a turbidity current, and this unit is characterized by domichnia of organisms that colonized the new sediment immediately after the turbidite event. Thus, turbidite sequences actually contain two separate generations of trace fossils: a predepositional trace fossil association in the fine-grained units and a postdepositional trace fossil association in the coarse-grained units.

In the latter case, laminated-to-burrowed sedimentary sequences represent the effects of storms, which erode the sea bottom and resuspend the sediment. As a storm subsides, sediment is redeposited, and the organism community is reestablished. The vertical transition from laminated sediment layers to burrowed sediment layers reflects a declining sedimentation rate, and the top of the sequence is bounded by an erosional unconformity that marks the next erosive storm event. *See* FOSSIL; PALEOECOLOGY; PALEONTOLOGY; SEDIMENTOLOGY.

A. A. Ekdale

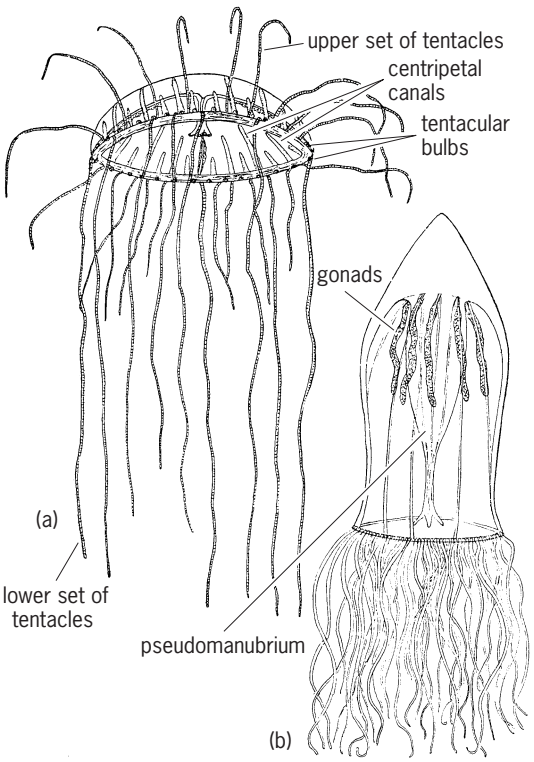
Bibliography. R. G. Bromley, *Trace Fossils: Biology, Taphonomy and Applications.*, 2d ed., Chapman and Hall, 1996; A. A. Ekdale, *Palaeogeog. Palaeoclimatol. Palaeoecol.*, 50:63–81, 1985; A. A. Ekdale, *Palaios*, 3:464–472, 1988; A. A. Ekdale, R. G. Bromley, and S. G. Pemberton, *Ichnology: Trace Fossils in Sedimentology and Stratigraphy*, 1984; W. Häntzschel, in C. Teichert (ed.), *Treatise on Invertebrate Paleontology*, pt. W, revised, University of Kansas Press, 1975.

Trachylina

An order of jellyfish of the class Hydrozoa of the phylum Cnidaria. These jellyfish are of moderate size. They differ from other hydrozoan jellyfish in having balancing organs which develop partly from the digestive epithelium and in having only a small polyp stage or none at all. Many authorities now recognize three distinct orders of trachylines, Limnomedusae, Trachymedusae, and Narcomedusae, and in this case the older term Trachylina is abandoned.

Limnomedusae have a small polyp stage which produces other polyps and medusae by budding. The tentacles are hollow. The best-known Limnomedusae are *Gonionemus* and *Craspedacusta*. *Gonionemus* is widely used in zoology classes as a representative jellyfish and is described in most textbooks. It is often abundant in salt-water ponds and bays. *Craspedacusta* lives in fresh water and occurs in nearly all parts of the world.

Trachymedusae (see **illus.**) and Narcomedusae are jellyfish of the open seas. Their tentacles, unlike



Trachymedusae, (*a*) *Olindias*, Bermuda. (*b*) *Aglantha*, Puget Sound. (*After L. H. Hyman, The Invertebrates, vol. 1, McGraw-Hill, 1940*)

those of Limnomedusae, have a solid core consisting of a single row of endodermal cells, and there is no polyp stage.

Narcomedusae differ from both Limnomedusae and Trachymedusae in having broad and often lobed stomachs; there are no peripheral canals, and tentacles are attached above the margin of the bell, which is heavily grooved and lobed. *See* HYDROZOA.

Sears Crowell

Bibliography. S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; F. S. Russell, *The Medusae of the British Isles*, 1954.

Trachyte

A light-colored, aphanitic (very finely crystalline) rock of volcanic origin, composed largely of alkali feldspar with minor amounts of dark-colored (mafic) minerals (biotite, hornblende, or pyroxene). If sodic plagioclase (oligoclase or andesine) exceeds the quantity of alkali feldspar, the rock is called latite. Trachyte and latite are chemically equivalent to syenite and monzonite, respectively. *See* SYENITE.

Texture. The extremely fine-grained texture and more or less glassy material are due to rapid cooling and solidification of the lava. Large crystals (phenocrysts) are commonly sprinkled liberally through the dense rock, giving it a porphyritic texture. These may be well formed and 1–2 in. (2.5–5 cm) wide. They appear as glassy crystals of sanidine, and in addition small mafic phenocrysts may be present. In latite the phenocrysts are largely plagioclase. As the quantity of glass increases, these porphyric rocks pass into vitrophyre; and as the abundance of phenocrysts increases, these rocks pass into trachyte porphyry. *See* PORPHYRY; VOLCANIC GLASS.

The detailed features of trachyte are best studied microscopically. Sanidine and orthoclase are dominant over oligoclase in normal (potash) trachyte. In alkali (soda) trachyte, both alkali feldspar and mafics are soda-rich.

Composition. Brown biotite mica is the common mafic. It occurs as flakes which may be more or less resorbed by the liquid in the late stages of solidification so that only patches of dusty iron oxide remain. Normal trachyte commonly carries somewhat corroded and resorbed hornblende or diopside. Alkali trachyte usually contains soda-rich amphibole (riebeckite, arfvedsonite, and barkevikite) or pyroxenes (aegirine-augite or aegirite). Zoned crystals with diopsidic cores and progressively more soda-rich margins are common.

Either free silica (quartz, tridymite, or cristobalite) or feldspathoids (leucite, nepheline, or sodalite) may be present in small amounts. With increase in free silica, the rock passes into rhyolite; and with increase in feldspathoids, it passes into phonolite. Accessory minerals as tiny grains and crystals are magmetite, ilmenite, apatite, zircon, and sphene.

Structure. Streaked, banded, and fluidal structures due to flowage of the solidifying lava are commonly visible in many trachytes and may be detected by a parallel arrangement of tabular feldspar phenocrysts. A distinctive microscopic feature is trachytic texture in which the tiny, lath-shaped sanidine crystals of the rock matrix are in parallel arrangement and closely packed. This rather uniform pattern is locally interrupted where the laths more or less deviate or wrap around the phenocrysts. Orthophyric texture is common where tiny feldspar crystals show a stumpy or square outline.

Occurrence and origin. Trachyte is not an abundant rock, but it is widespread. It occurs as flows, tuffs, or small intrusives (dikes and sills). It may be associated with alkali rhyolite, latite, or phonolite.

Trachyte is commonly considered to have been derived from a basaltic magma by differentiation, a process involving removal in large quantities of early formed crystals rich in iron, magnesium, and calcium. A factor of importance in the formation of some trachyte is contamination of the original magma by incorporation of foreign rock material. The chemical transformation of andesite to trachyte may have occurred (in the solid state) where calcium was removed and sodium added metasomatically. This may explain the origin of some keratophyres (a variety of soda-rich trachyte). *See* IGNEOUS ROCKS; MAGMA; SPILITE.

Carleton A. Chapman

Tractor

A wheeled, self-propelled vehicle for hauling other vehicles or equipment and for operating the towed implements; also, a crawler which runs on an endless, self-laid track and performs similar functions.

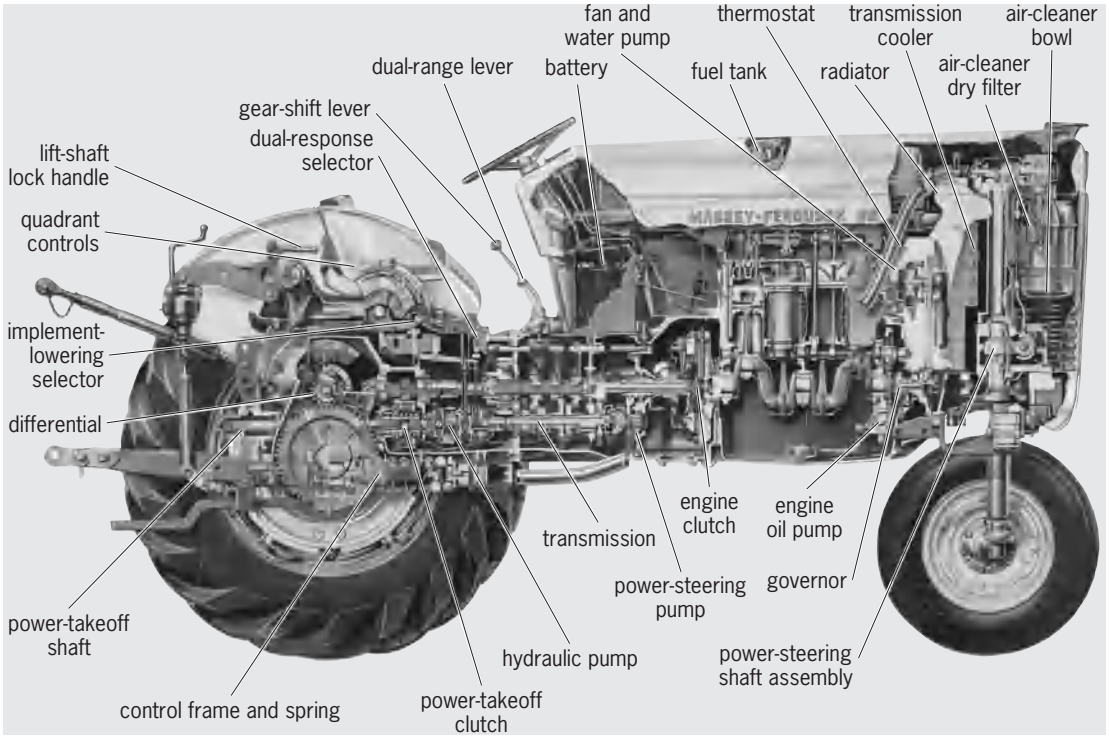
Farm tractor. A farm tractor is a multipurpose power unit. It has a drawbar for drawing tillage tools and a power takeoff device for driving implements or operating a belt pulley as shown in the **illustration**.

The acreage to be worked, type of crops grown, and the terrain all impose their requirements on tractor design. Accordingly, models vary in such details as power generated, weight, ground clearance, turning radius, and facilities for operating equipment. All models can, however, be grouped under four general types: four-wheel, row-crop or high-wheel, tricycle, and crawler.

Power delivered at the place where it is useful is what counts; therefore, tractors are rated by the horsepower they deliver at the drawbar and at the belt. On small models, the drawbar and belt horsepower may run as low as 10 (7.5 kW); on large models the drawbar horsepower runs as high as 132 (98 kW), while belt horsepower reaches about 144 (107 kW).

The major components are engine, clutch, and transmission. These components are intimately related and designed to work in conjunction with each other to accomplish specific work.

Unlike passenger-car engines, which are of the high-speed type, tractor engines are relatively low-speed; their maximum horsepower is generated at crankshaft speeds in the neighborhood of 2000 rpm. These engines have one, two, three, four, six, or eight



Four-wheel tractor showing major components. (*Massey-Ferguson, Inc.*)

cylinders and operate on gasoline, kerosine, liquid petroleum gas, or diesel fuel. They are of the spark-ignition or diesel type, operating on the four-stroke-cycle principle, and are cooled by water or air.

Power is transmitted to the rear wheels or to all four wheels. Drive to the front wheels is mechanical or hydrostatic, its purpose being to increase drawbar pull at the will of the operator. Transmissions have 3, 4, 5, 6, 8, 10, or 12 forward speeds and one or two reverse gears. Clutchless hydraulic transmissions are also used, making it possible to shift gears while in motion. *See* CLUTCH; TORQUE CONVERTER.

Vehicle speeds are low, ranging from slightly more than 1 mi/h(1.6 km/s) to about 18 mi/h (13 km/h) in high gear.

Between transmission and engine is a single- or multiple-disk clutch of the general automotive type, manually or hydraulically controlled.

Power is taken off the transmission by shaft for operating equipment. The power takeoff may be manually controlled or hydraulically engaged and disengaged independently of vehicle speed. It may be run at one or two set speeds, depending upon the nature of the work it is intended to do. With a pulley attached to the power takeoff, a similar choice of speeds is afforded for belt work.

Hydraulic systems are used for control of both rear- and front-mounted implements. The rear-mounted implement may be attached to the two arms of the drawbar and to a third installed arm to give a three-point hitch. An engine-driven hydraulic pump and cylinder built into the tractor provide the power to raise or lower the arms and thus to lift the implement to carry it, or lower it to the ground for work, at the will of the driver. The hydraulic mechanism may be so designed as to transfer weight from the front wheels to the rear wheels of the tractor as the load demands to give better traction when drawing multiple plows. Some systems are designed to control the depth of plow penetration automatically; others will disengage the clutch automatically if the mounted implement strikes an obstruction. The hydraulic cylinder may be attached to the implement, or may be part of it.

With the exception of crawlers, tractors are steered by turning a wheel. Hydraulic power devices are also used as manual assists.

A tractor has no frame and no springs. The supporting structure is a housing, or housings, for the transmission, clutch, hydraulic mechanism, drive shaft, and differential, to which are bolted the engine crankcase, or frame for carrying the engine, and the housing for the rear axle. The cushioning of the tractor load depends on the tires.

A typical front axle is pivoted at its center to accommodate for rough ground. Made in sections, it can be extended to change the width between wheels for straddling crop rows. The rear tread width can also be adjusted by assembling the wheel disk and rim in different positions.

Braking is by means of drum or disk brakes on the rear wheels, or by brakes on the differential. Both types permit the brakes to be applied independently on either side to assist in making sharp turns.

Crawler-type tractors in the smaller models are also used for farm work. In lieu of rubber-tired wheels, the crawler propels itself on two endless tracks made of hinged steel plates. Each track envelops a front and rear wheel; each is carried on a roller track frame which supports the weight of

the tractor. The track is made to move by engaging teeth on the periphery of the rear wheels. Dual clutches permit power to be applied independently to each track. Steering is accomplished by manipulating levers which control the clutches to send power to one track while keeping the other motionless. *See* AGRICULTURAL MACHINERY.

Industrial tractor. The basic design of an industrial tractor for hauling and for operating construction equipment departs little from that of a farm tractor, and differences in design of models fit the vehicle to its intended work.

Because high ground clearance is not needed for industrial work, the tractor is commonly built with a lower center of gravity and is capable of traveling a few miles per hour faster than a farm tractor. If its use is confined to hauling, it may not be equipped with hydraulic power. If it is to be used for operating a scraper, backhoe, or front-end loader, its structure may be heavier and more rugged. *See* BULK-HANDLING MACHINES.

Philip H. Smith

Tractrix

A plane curve for which the length of any tangent between the curve and a fixed line is constant c. If the x axis is the fixed line, its differential equation is Eq. (1). With ψ, the inclination of the tangent, as

$$\left(\frac{dy}{dx}\right)^2 = \frac{y^2}{c^2 - y^2} \tag{1}$$

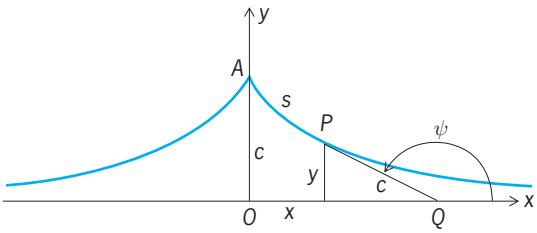
parameter, this yields Eqs. (2). The tractrix has the

$$\begin{aligned} x &= c \log \tan {}^1\!/_2 \,|\psi| + c \cos \psi \\ y &= c \sin \psi \end{aligned} \tag{2}$$

x axis as asymptote and cusps at $(0,\pm c)$. The arc AP (see **illus.**), measured from a cusp, is as in Eq. (3).

$$s = -c \log \sin \psi = c \log \frac{c}{y} \tag{3}$$

A taut string PQ, moving in a horizontal plane so that Q describes a straight line, will pull a heavy particle at P along a tractrix. The tractrix is an involute



Tractrix (upper half).

of a catenary. A tractrix revolved about its asymptote generates a pseudospherical surface of constant negative gaussian curvature $-1/c^2$; its area is $4\pi c^2$, the same as a sphere of radius c. *See* CATENARY.

Louis Brand

Traffic-control systems

Systems that act to control the movement of people, goods, and vehicles in order to ensure their safe, orderly, and expeditious flow throughout the transportation system. Each of the five areas of transportation—roadways, airports and airways, railways, coastal and inland waterways, and pipelines—have unique systems of control.

Roadway traffic control. Roadway traffic-control systems are intended to improve safety, increase the operational efficiency and capacity of the roadway, and contribute to the traveler's comfort and convenience. They range from simple control at isolated intersections using signs and markings, to sophisticated traffic-control centers which have the ability to react to changes in the traffic environment. Traffic-control systems are used at roadway intersections, on highways and freeways, at ramp entrances to freeways, and in monitoring and controlling wider-area transportation networks. Intelligent roadway traffic control, now known as intelligent transportation systems (ITS), is a very sophisticated form of traffic control for roadway and other areas.

Traffic-control devices. Traffic-control devices are the primary means of communication with road users, including drivers, pedestrians, and bicyclists. Devices placed on, adjacent to, or above the roadway fulfill three general functions: guidance, regulation, and warning. Guidance devices provide information and direction, including mileage to destinations, points of interest, and route designations. Regulatory devices provide information concerning applicable traffic regulations and laws, including speed limits and turn prohibitions. Warning devices alert road users to potentially hazardous conditions. Traffic-control devices can also be thought of as active or passive. Passive devices, such as signs and markings, provide the same guidance at all times. Active devices, such as traffic signals and changeable message signs, provide guidance to drivers in response to changes in the roadway environment.

Traffic signs are the most common traffic-control devices. Signs are usually mounted above or beside the roadway and convey information by means of words or symbols. Specific colors and shapes have been reserved for each functional type of sign. In the United States, guidance signs are rectangular, with white letters and symbols on a green background. Warning signs are usually diamond shaped, with black markings on a yellow background. Regulatory signs, also rectangular, have black markings on a white background. The message or symbol upon the sign and the size and placement of the sign itself are carefully determined to provide consistent, understandable information to road users.

Changeable message signs present lighted, changeable messages to travelers on an as-needed basis. They are often used to provide drivers advanced warning of congestion, construction work zones, or other hazardous conditions. Changeable message signs may be permanently mounted beside

or above the roadway, or upon trailers or trucks for temporary use.

Traffic markings provide guidance to traffic, supplement information provided by signs, and channel vehicle movements without the use of barriers or other physical equipment. Typical roadway markings include stop bars, lane lines, channelizing lines, and other guidance markings. Markings may be made with reflectorized paint or with raised reflective markers.

Traffic islands are defined areas between traffic lanes for the control of vehicle movements or the accommodation of pedestrians. Traffic-control islands may be designated by paint, raised bars, curbs, guideposts, or other devices. Islands may be classified according to their principal function as pedestrian refuge islands, divisional islands, or traffic-channelizing islands. Pedestrian refuge islands provide a place of safety for pedestrians who cannot cross an entire roadway width at one time because of oncoming traffic or changing traffic signals. Traffic-divisional islands separate opposing- or same-direction traffic along the course of a highway. Divisional islands may be used to guide traffic around an obstruction in the roadway. When they are continuous, divisional islands are referred to as medians. Traffic-channelizing islands control and direct vehicle operators into the proper channels for their intended routes. Channelization is particularly useful in controlling the flow of traffic at roadways and intersections where there would otherwise be broad expanses of pavement.

Traffic signals include all power-operated devices (except signs) for regulating, directing, or warning motorists or pedestrians of traffic conditions. Although more expensive than signs, signals provide more positive guidance and increased flexibility. Traffic signals consist of illuminated, colored displays. The tricolor displays used at intersections are the most common; however, flashing yellow beacons are used to attract drivers' attention to warning signs, and red-green displays may be used on freeway ramps.

Traffic-control systems. Traffic conditions at specific locations, in conjunction with those in the wider area, determine the type and complexity of traffic control needed. The simplest roadway traffic-control systems consist of a combination of passive traffic-control devices, such as signs and markings. Increasingly complex traffic-control systems incorporate a variety of passive and active traffic-control devices, traffic monitoring technologies, and computer control systems.

Intersection traffic-control systems may employ several different control strategies. In the simplest system, signals operate in isolation and are pretimed, employing one or more preset signal patterns throughout the day. As the systems become more complex, signals along a route or in a network of intersections may be coordinated to improve traffic flow. All or part of the signal cycle at some locations may be demand actuated; for example, a left-turn arrow may be provided when sensors indicate that a vehicle is present in the turn lane.

Freeway traffic-control systems may incorporate a number of control strategies for both main lanes and high-occupancy-vehicle (HOV) lanes. These include ramp metering, lane controls, and in some cases, the use of changeable speed limits. Changeable message signs may be used to provide drivers with information on conditions, including congestion, incidents, construction, bad weather, and other hazardous situations.

Traffic monitoring systems provide information on traffic volume and speed on roadways. Inductive-loop detectors are the most common type of monitoring system. However, video-based systems for direct surveillance and those employing image analysis to extract traffic characteristics (speed and volume) are growing in use. In special circumstances, radar and sonic-based systems are also used. *See* IMAGE PROCESSING; INDUCTANCE; RADAR.

Intelligent transportation systems. Intelligent transportation systems, previously known as intelligent vehicle-highway systems (IVHS), involve the application of advanced technologies to surface transportation problems. A wide array of transportation areas are involved, including traffic and transportation management, travel demand management, advanced public transportation management, electronic payment, commercial vehicle operations, emergency services management, and advanced vehicle control and safety systems.

Advanced traffic management systems with traffic or transportation management centers (TMCs) operate in a number of urban areas. They provide integrated centers from which to efficiently manage transportation systems and mitigate the impact of accidents and other incidents. Traffic monitoring systems relay information on flow conditions to the traffic management center, where operators may manually adjust signal cycles in response to incidents, notify the appropriate emergency response personnel in case of incidents, and provide information to drivers through changeable message signs and highway advisory radio (HAR). In-vehicle navigation systems provide map and route guidance information. *See* HIGHWAY ENGINEERING.

Airport and airway traffic control. The federal government has designated airspace as either uncontrolled or controlled. In uncontrolled airspace, pilots may conduct flights without specific authorization. In controlled airspace, pilots may be required to maintain communications with the appropriate air-traffic control facility to receive authorization and instruction on traversing, taking off from, or landing, in that controlled area. Air-traffic control systems for controlled areas may be divided loosely into en route and terminal systems.

Terminal control. The level of control at terminals is determined by the designation of the airport and the type of service provided.

Runway lighting and approach lighting are provided for night operations and use during inclement weather when instrument flight rules (IFR) must be

followed. Colored lighting on runways and taxiways is used to guide ground movements. Runway approach lighting provides a visual reference for the glide path of the designated landing runway.

Instrument landing systems (ILS), used primarily during instrument flight rules conditions, provide electronic guidance to the pilot on the final approach to the runway. They provide both lateral (azimuth) and vertical (elevation) guidance to descending aircraft. *See* INSTRUMENT LANDING SYSTEM (ILS).

Microwave landing systems (MLS) are technically sophisticated instrument landing systems designed to overcome the limitations of instrument landing systems. The microwave landing system includes instrumentation for providing pilots with azimuth, elevation, and precision distance measurements to the end of the runway, and uses a time-reference scanning beam, allowing curved or multipath approaches to the active runway. The microwave landing system has seen very limited use in the United States and is being phased out. *See* MICROWAVE LANDING SYSTEM (MLS).

Precision-approach radar (PAR), located in the airport traffic-control tower, provides landing assistance to pilots. The pilot receives confirming radio information on lateral and vertical deviations from a predetermined glide path to the designated runway, and can then manually adjust the aircraft approach path as necessary. Precision-approach radar has been implemented only at military airports and a few very large civilian airports. *See* PRECISION APPROACH RADAR (PAR).

Airport surface detection equipment (ASDE) allows controllers to monitor all aircraft activity on the ground, including all runways, taxiways, and parking ramps. It is a relatively new development, available at only a few large airports in the United States. *See* AIRPORT SURFACE DETECTION EQUIPMENT.

En route control and guidance. Air route traffic control centers (ARTCCs) monitor and direct the movement of aircraft along the airways or to specific destinations. Each ARTCC controller is assigned a specific geographical area. Aircrafts are released as they enter adjoining areas. Each aircraft is tracked by using both radar and voice communications. Air route surveillance radar (ARSR), which is installed nationwide, allows controllers to precisely locate each aircraft along an airway. *See* SURVEILLANCE RADAR.

Air-traffic control radar beacon systems (ATCRBS) augment the ARSR. ATCRBS equipment includes an interrogator and transponder to provide more positive location information than the passive radar of the ARSR. The interrogator transmits a discretely coded signal requesting all transponders on that mode to respond. The controller views the coded aircraft target on the radar scope, including the plane's identification and altitude on an alphanumeric display.

En route guidance is provided by the very high-frequency multidirectional radio system, known as VORTAC. VORTAC is a combination of VOR (VHF Omnidirectional Range) and TACAN (Tactical Air Navigation). For altitudes of 24,000–60,000 ft (7315–18,228 m), airspace is under the direction of controllers using instrument flight rules, known as air positive control (APC). In heavily traveled corridors (for example, in the northeast), control is exercised above 18,000 ft (5486 m). *See* RHO-THETA SYSTEM; TACAN.

Inertial navigation systems are used by large commercial transports and some military aircraft for long-range travel, especially transoceanic flights. An on-board inertial device computes latitudes and longitudes, allowing pilots to pinpoint their locations along the airways and to provide controllers with accurate position reports. *See* INERTIAL GUIDANCE SYSTEM.

The Global Positioning System (GPS) is often used in aircraft tracking and navigation systems. Aircraft receiving corrections from the Differential Global Positioning System (DGPS) can pinpoint their positions to an accuracy of a few meters. The Global Positioning System has applications to both air and surface surveillance. *See* SATELLITE NAVIGATION SYSTEMS.

Advanced Air Traffic Management System. The Advanced Air Traffic Management System (ATMS) is an automated system that balances air-traffic departures with airspace system capabilities. Flight information and traffic counts are entered into a central flow-control facility, where they are compared to threshold capacity levels at various airports. Traffic managers are alerted to predicted congestion situations in specific areas of the national airspace. Managers can take action to mitigate airport and airway congestion and subsequent delay problems by directing aircraft to different routes and airports, or by holding aircraft on the ground. *See* AIR NAVIGATION; AIR-TRAFFIC CONTROL; AIR TRANSPORTATION.

Railroad traffic control. Railroads operate high-speed freight and passenger services essentially over an exclusive right of way. Railroads use both semaphore and light signals for traffic control. Semaphores convey visual messages to train operators according to predetermined rules indicating how the train is to be operated in specified areas.

Automatic block signaling prevents rear-end and head-on collisions on signal tracks. In this system, track sections are divided into blocks. Only one train is permitted to occupy a block at any time. Blocks are monitored by automatic circuitry that controls traffic signals, indicating the appropriate clear or stop signals to following or approaching trains. Similar block systems are used for subway systems.

The need to keep rail freight cars moving and to determine the location of individual cars has led to the development of an extensive traffic-control system both within terminals and en route. Centralized traffic-control systems may control hundreds of miles of track signals and switches. A dispatcher at a central location monitors the location of trains by means of visual displays of colored lights on a large track diagram, and can operate the switches and signals at key points from the central control console.

Automatic car identification systems electronically scan and record the information stored on an optically coded plate attached to the side of each freight

car. This information is transmitted to each railroad and to a centralized computer system called the Telerail Automated Information Network II (TRAIN II). This network connects individual railroads and permits a balancing of full and empty cars among the railroads and geographical regions. *See* RAILROAD CONTROL SYSTEMS; RAILROAD ENGINEERING.

Vessel traffic control. The volume of traffic on waterways is generally quite low. Marine aides, therefore, function more for informational, advisory, and guidance purposes than as positive traffic-control devices. Lighted or unlighted buoys indicate navigable areas in coastal waters and within waterways. Navigators accurately determine their positions by referencing sequentially numbered and colored-coded buoys to navigation charts. *See* BUOY; PILOTING.

Lightships and lighthouses with fog signals and radio beacons are placed as markers at prominent points during periods of limited visibility. Each station has a characteristic identification signal. Vessels can determine their positions in coastal waters by timing the interval between flashes of the fog light and the sounding of the horn signal. *See* LIGHTHOUSE.

Radar devices have become more common, even on smaller ships. They provide navigation and traffic control under conditions of poor visibility when it is difficulty to determine the location of the ship and to avoid other ships. Navigation systems often employ the Differential Global Positioning System.

The Vessel Traffic Service (VTS) is available in selected areas. Services may range from the provision of single advisory messages to extensive management of traffic communication and radar services. The Vessel Traffic Service is particularly useful in approaches to ports, access channels, and other areas with high traffic densities; for traffic carrying dangerous cargoes, in especially narrow channels; and in areas of environmental sensitivity. *See* MARINE NAVIGATION; NAVIGATION.

Pipeline traffic control. The 450,000 mi (750,000 km) of pipelines in the United States are a major part of the nation's transportation network, carrying about 25% of all intercity freight-ton mileage. The primary goods moved through pipelines are oil and oil by-products, natural gas, and fertilizers. Solid products, such as coal, may be moved through slurry pipelines.

The movement of goods in pipelines is controlled by systems of valves, pumps, and compressors. Pipe diameters generally range from 6 in. (15 cm) to 30–40 in. (75–100 cm). Pumping stations are generally positioned every 30–70 mi (50–110 km) along the pipeline, depending on the terrain and viscosity of the materials transported.

Slurry pipelines maintain a homogeneous mixture of a solid product and liquid to flow at approximately 3–5 mi/h (5–8 km/h). The speed is carefully calculated for the particular material being transported. If the material moves too fast, the walls of the pipe may be abraded. If it moves too slowly, the solids drop out of suspension. Large amounts of water are required for the operation of coal slurry pipelines.

The Trans-Alaska Pipeline System transports crude oil at about 7.5 mi/h (12 km/h), carrying 2×10^6 barrels/day (3×10^5 m^3/day). This requires 12 pumping stations along the 800-mi (1300-km) route. Much of the pipeline is above ground to avoid melting the permafrost. In those locations where below-ground installation was necessary because of the terrain or to permit animal crossing, refrigerant lines were located alongside the pipe to keep the warm moving oil (at about 135°F or 57°C) from melting the permafrost. *See* PIPELINE; TRANSPORTATION ENGINEERING. James Costantino; Donna C. Nelson

Bibliography. K. J. Button and D. A. Hensher (eds.), *Handbook of Transport Systems and Traffic Control*, 2001; J. D. Fricker and R. K. Whitford, *Fundamentals of Transportation Engineering: A Multimodal Approach*, 2003; N. J. Garber and L. A. Hoel, *Traffic and Highway Engineering*, 3d ed., 2001; S. Ghosh and T. Lee, *Intelligent Transportation Systems: New Principles and Architectures*, 2000; M. Kutz (ed.), *Handbook of Transportation Engineering*, 2003; J. M. Sussman, *Perspectives on Intelligent Transportation Systems*, 2005.

Trajectory

The curve described by a body moving through space, as a meteor moving through the atmosphere, a planet around the Sun, a projectile fired from a gun, or a rocket in flight. In general, the trajectory of a body in a gravitational field is a conic section—ellipse, hyperbola, or parabola—depending on the energy of the motion. The trajectory of a shell or rocket fired from the ground is a portion of an ellipse with the Earth's center as one focus; however, if the altitude reached is not great, the effect of gravity is essentially constant, and the parabola is a good approximation. *See* BALLISTICS. John P. Hagen

Tranquilizer

A psychopharmacologic drug that tends to calm overexcited individuals, producing a state of "tranquility." The term tranquilizer was originally applied to two groups of drugs. Members of one group, the antianxiety drugs, were called minor tranquilizers, and members of the second group, the antipsychotic drugs, were called major tranquilizers. Although drugs in both groups may have a calming effect in appropriate doses, it is now clear that the two groups are quite different.

Antianxiety drugs. This group of drugs, also known as anxiolytics, comprises three subgroups: propanediols, of which meprobamate (Miltown) is the best known; alcohol (ethanol) and the barbiturates, such as phenobarbital; benzodiazepines, of which diazepam (Valium) is the best known.

Like all drugs, antianxiety drugs have a range of effects at different doses, including—in the low-to-high order of doses at which the effects generally occur—(1) a period of transient excitement for some

persons; (2) a reduction in anxiety, the effect for which the drugs are prescribed; (3) antagonism of certain kinds of seizure activity in the brain, a so-called anticonvulsant activity; (4) a sedative or sleep-inducing effect; and (5) a generalized reduction in muscle tension, a muscle relaxant effect.

An important difference among the three groups of antianxiety drugs is based on this dose dependency: for the first two groups, the difference in doses that reduce anxiety and those that produce sedation is small; a major advantage of the third group is that this difference is large, so that they reduce anxiety with relatively less sedation. Furthermore, any sedation that may occur initially with a benzodiazepine tends to wane with repeated administration (an effect known as tolerance), whereas its antianxiety effect is less likely to do so. This phenomenon contributes to an increased separation of antianxiety and sedative effects. This differential, coupled with the relative lack of serious complications resulting from overdose, accounts for the enormous popularity of the benzodiazepines in the clinical management of anxiety.

The benzodiazepines modify brain activity by interacting with gamma-aminobutyric acid (GABA), a naturally occurring substance that inhibits the activity of brain cells by increasing their permeability to chloride ions; these ions decrease the likelihood that the cell will be responsive to input from other cells. The benzodiazepines increase the tendency for GABA to act at its own receptors, and thus have an indirect inhibitory role that is GABA-dependent. Ethanol and the barbiturates, on the other hand, act directly on the membrane to increase chloride permeability.

Gamma-amino butyric acid and changes in chloride permeability are also involved in the muscle relaxant, anticonvulsant, and sedative effects of these drugs. It is not clear, however, whether these interactions can account directly for their antianxiety effects. Meprobamate, for example, reduces anxiety but has not been shown to augment GABA; increased GABA activity may therefore be a sufficient but not necessary condition for anxiety reduction. On the other hand, another drug (valproic acid) can augment the activity of GABA but has not been shown to have clinical antianxiety effects; the benzodiazepines may therefore engage some process that acts in concert with GABA but that is not engaged by valproic acid. *See* AFFECTIVE DISORDERS.

Antipsychotic drugs. The control of psychotic symptoms by these drugs wrought a revolution in mental health care when they were introduced in the mid-1950s: chronic treatment with antipsychotics reversed, for the first time, the progressive increase in hospitalization of patients. Because of the symptom control that antipsychotics provide, it is possible to manage the majority of psychotic patients (mostly schizophrenics) on an outpatient basis. *See* SCHIZOPHRENIA.

There are a large number of antipsychotic drugs, all of which are equally useful in managing psychotic symptoms. However, the antipsychotics vary widely in potency, that is, the amount of drug that is required to produce effective symptom control. Furthermore, different side effects are associated with different potencies: a low-potency drug such as chlorpromazine (Thorazine), for example, is more likely to produce sedation and certain other side effects than is a high-potency drug such as haloperidol (Haldol); haloperidol is more likely to produce disturbances of movement than is chlorpromazine.

The major complication of treatment with antipsychotic drugs is the array of movement disturbances that they can induce. Most of these disturbances can be controlled by concurrent medication with other drugs. For example, one type of reaction mimics the signs of Parkinson's disease; it occurs in about 30% of patients, although only transiently in many cases. This type of reaction can be successfully managed by adjunctive treatment with one of the several drugs useful in the control of parkinsonian symptoms. On the other hand, another type of reaction that may occur in up to 25% of patients in the course of several years of treatment, tardive dyskinesia, has thus far proven to be refractory to adjunctive medication. Therefore, the patient may face a trade-off between disturbed movement and the debilitating symptoms of psychosis.

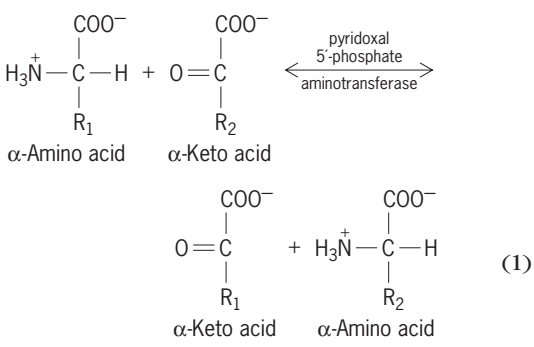
These disturbances are clearly related to a drug-induced attenuation in the activity of a substance called dopamine, which occurs naturally in various areas of the brain and is critically involved in the mediation of normal movement. It is less clear, however, as to whether the reduction in psychotic symptoms produced by these drugs is also due to attenuated dopamine activity. *See* PSYCHOPHARMACOLOGY; PSYCHOSIS.

Peter L. Carlton

Bibliography. R. J. Baldessarini, *Chemotherapy in Psychiatry*, 1985; H. Y. Meltzer (ed.), *Psychopharmacology: The Third Generation of Progress*, 1987.

Transamination

Transfer of an amino group from one carbon chain to another without the intermediate formation of ammonia. Typically, an α-amino acid serves as the donor of the amino group and is converted to an α-keto acid, and an α-keto acid serves as the acceptor of the amino group and is converted to an α-amino acid, as in reaction (1), where R_1 and R_2 represent two different side chains.



(1)

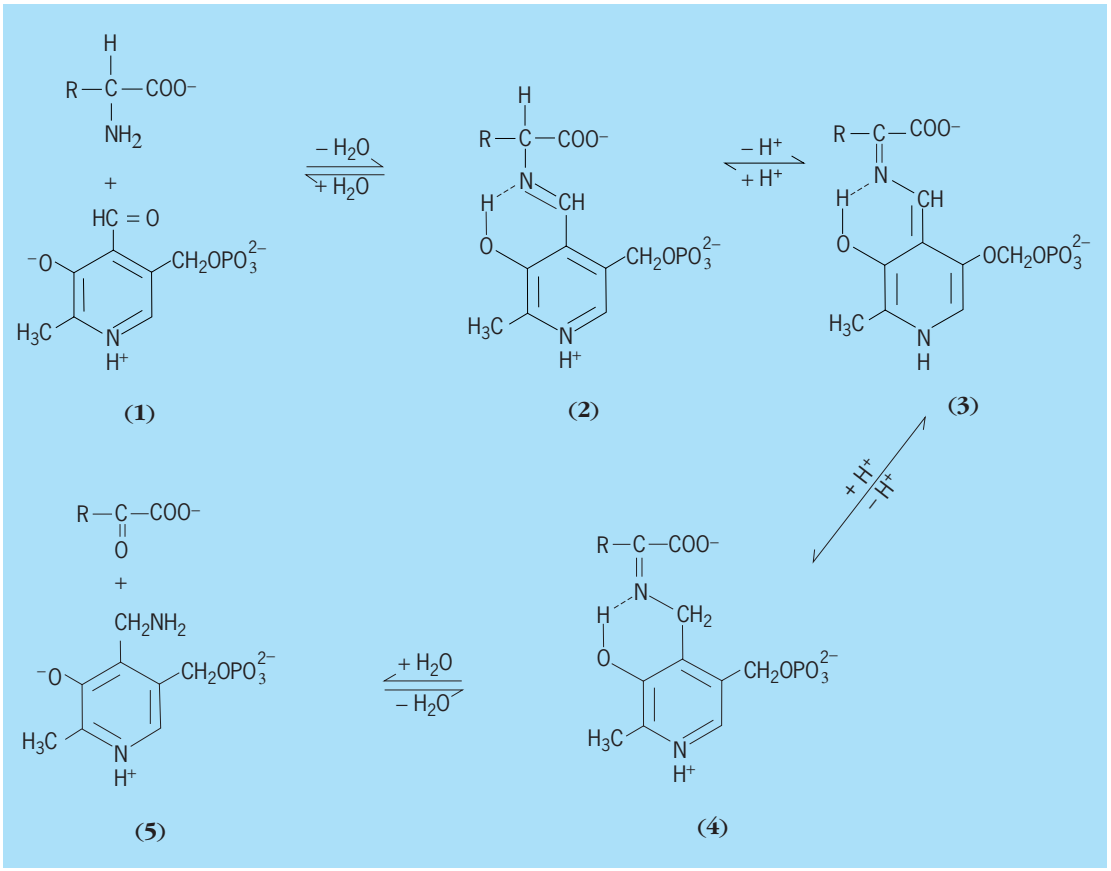


Fig. 1. Pathway of reactions for a transamination.

Enzyme-catalyzed reactions of this type were first discovered by A. E. Braunstein and M. G. Kritzman in 1937. Transamination reactions are catalyzed by enzymes called aminotransferases (or transaminases) that require pyridoxal 5′-phosphate as a cofactor. The pyridoxal 5′-phosphate functions alternately as an amino group acceptor [in the aldehyde form, PLP, as in reaction (2a)] and as an amino group donor [in the amine form, PMP, as in reaction (2b)], where

$$\text{Amino acid}_1 + \text{PLP} \rightarrow \text{keto acid}_1 + \text{PMP} \qquad (2a)$$

$$\text{Keto acid}_2 + \text{PMP} \rightarrow \text{amino acid}_2 + \text{PLP} \qquad (2b)$$

subscripts 1 and 2 indicate different amino (or keto) acids. **Figure 1** illustrates the chemistry of these reactions.

Most physiologically important aminotransferases have a preferred amino acid/keto acid substrate and utilize α-ketoglutarate/glutamate as the counter keto acid/amino acid. An example is aspartate aminotransferase (usually abbreviated AST for aspartate transaminase or SGOT for serum glutamate-oxaloacetate transaminase), which accepts aspartate or oxaloacetate as substrate and uses glutamate or α-ketoglutarate as cosubstrate (**Fig. 2**).

Amino acid metabolism. In human tissues, the amino acids alanine, aspartate, glutamate, tyrosine, serine, valine, isoleucine, and leucine are actively transaminated. Histidine, phenylalanine, methionine, cysteine, glutamine, asparagine, and glycine also may undergo transamination in human tissues; these amino acids are primarily metabolized by other types of reactions under normal physiological conditions, but intermediates in the degradation pathways of phenylalanine, cysteine, glutamine, asparagine, and glycine (that is, tyrosine formed from phenylalanine, cysteinesulfinate formed from cysteine, glutamate formed from glutamine, aspartate formed from asparagine, and serine formed from glycine) are actively transaminated. Threonine, lysine, tryptophan, proline, and arginine do not directly participate

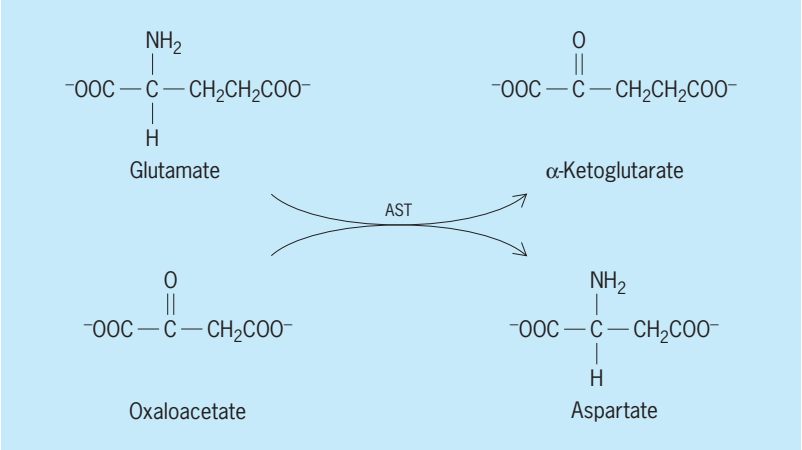


Fig. 2. Transamination reaction catalyzed by aspartate aminotransferase (AST).

in transamination reactions in mammalian tissues; nevertheless, intermediates in the degradation pathways of threonine (glycine), lysine (α-aminoadipate), tryptophan (3-hydroxykynurenine and alanine), and proline/arginine (ornithine and glutamate) do undergo transamination for removal of amino groups from the carbon chain. The interconversion of ornithine and glutamate γ-semialdehyde represents a more unusual type of transamination reaction in which one of the reactants is not an α-amino acid/α-keto acid couple. In this specific case, the δ-amino group of ornithine is transferred by ornithine δ-aminotransferase. *See* AMINO ACIDS.

Amino acid catabolism. Transamination reactions play a prominent role in the synthesis of the dispensable (or nonessential) amino acids in higher organisms. Amino groups are directly transferred to the keto acids pyruvate, oxaloacetate, and α-ketoglutarate to form alanine, aspartate, and glutamate, respectively. Transamination is also involved in the pathways of de novo synthesis of serine, glycine, glutamine, asparagine, proline, and arginine.

Aspartate. Because α-ketoglutarate is used so widely as the acceptor of amino groups in transamination reactions, the α-amino groups of numerous amino acids are funneled through glutamate in the process of amino acid catabolism. Aspartate aminotransferases are widespread in tissues, so amino groups are readily interchanged between glutamate and aspartate. Amino groups from various amino acids are funneled, via transamination reactions, to aspartate for entrance into the urea cycle. Aspartate is also an important donor of amino groups for purine nucleotide synthesis and of both carbon and nitrogen for pyrimidine nucleotide synthesis. The mitochondrial and cytosolic isozymes of AST play important roles in the movement of carbon and reducing equivalents across mitochondrial membranes by the malate-aspartate shuttle. During ureagenesis (urea synthesis) and gluconeogenesis (production of glucose from amino acid carbon chains or other nonglucose substrates such as lactate and glycerol) in the liver, aspartate carries carbon and reducing equivalents for hepatic gluconeogenesis as well as nitrogen for ureagenesis from the mitochondria to the cytosol. *See* LIVER.

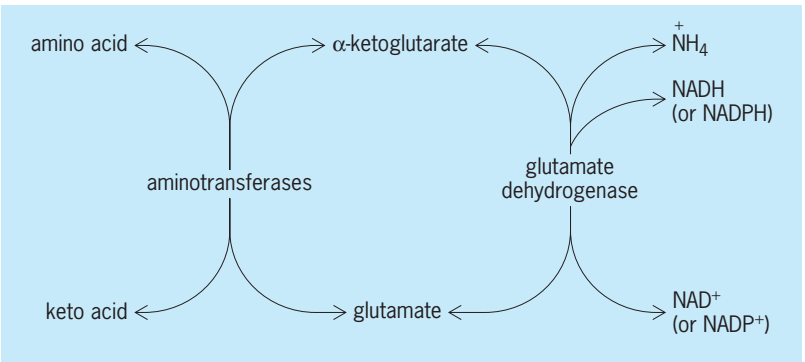
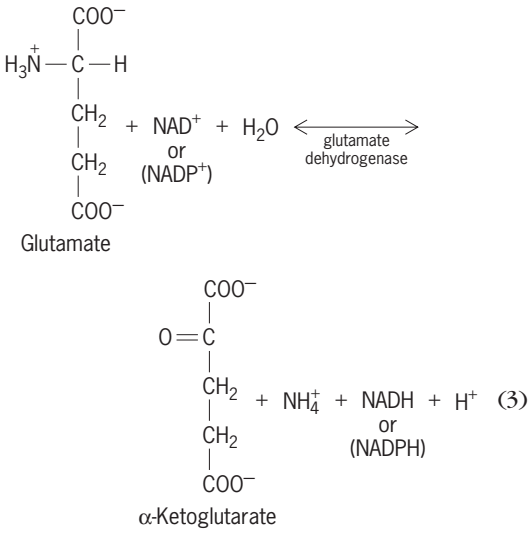


Fig. 3. Transamination reaction in which glutamate/α-ketoglutarate serves as cosubstrate coupled with the glutamate dehydrogenase reaction.

Alanine. Alanine aminotransferase (usually abbreviated ALT for alanine transaminase or SGPT for serum glutamate-pyruvate transaminase) catalyzes a reaction close to equilibrium in the liver, skeletal muscle, and the small intestine. Both alanine and glutamine are important carriers of amino groups derived from muscle amino acid catabolism to the liver for urea synthesis. Alanine is also released by the small intestine as a result of its catabolism of glutamine, glutamate, and aspartate. The liver metabolizes alanine by transamination, and uses the carbon chain for gluconeogenesis and the amino group for urea synthesis.

Glutamate. The movement of nitrogen between organic and inorganic pools is mediated by the glutamate/α-ketoglutarate couple and a reaction catalyzed by glutamate dehydrogenase, as shown in reaction (3).



Glutamate dehydrogenase is a mitochondrial enzyme that catalyzes a near-equilibrium reaction in tissues with high activity (liver, kidney, brain), and it can operate to either incorporate ammonia into or release ammonia from the α-amino acid pool. The direction of flux depends on the provision and removal of reactants. The equilibrium nature of the glutamate dehydrogenase and transamination reactions in these tissues acts to maintain intracellular ammonia levels in a narrow range.

Transamination reactions in which the glutamate/α-ketoglutarate couple serve as cosubstrate coupled with the glutamate dehydrogenase reaction allow the α-amino group of most amino acids to be ultimately released as ammonia (**Fig. 3**). This ammonia may be incorporated into glutamine, which donates amide groups for various synthetic reactions, or incorporated into urea via carbamoyl phosphate synthesis. Ammonia/ammonium is also excreted in the urine as such and plays an important role in acid-base balance. *See* NITROGEN EXCRETION; URINE.

Aminotransferases as clinical indicators. Plasma or serum levels of ALT and AST are frequently used as clinical indicators of tissue damage. These enzymes leak out of damaged cells. Because alanine aminotransferase is highest in liver, a high plasma level of

this aminotransferase is usually an indicator of liver damage. High levels of aspartate aminotransferase are less specific because this enzyme can be released by damaged liver, muscle, heart, kidney, and brain tissue. *See* ENZYME; PROTEIN METABOLISM.

Martha H. Stipanuk

Bibliography. A. J. Cooper and A. Meister, An appreciation of Professor Alexander E. Braunstein: The discovery and scope of enzymatic transamination, *Biochimie*, 71:387–404, 1989; M. H. Stipanuk and M. Watford, Amino acid metabolism, in M. H. Stipanuk (ed.), *Biochemical and Physiological Aspects of Human Nutrition*, 2d ed., Saunders, Philadephia, 2006.

Transcription

The process that occurs within a living cell in which an enzyme makes a ribonucleic acid (RNA) copy of the deoxyribonucleic acid (DNA) that contains the genetic information. The resulting RNA is often used as a template to make proteins by the cellular protein-synthesizing (translation) machinery. Transcription is an essential step in the growth and differentiation of all cells and contributes to almost every aspect of the development of an organism. Because transcription is so fundamental to life, there is significant conservation of proteins and enzymatic steps involved in transcription in cells of bacterial, archaeal, and eukaryotic origin. An RNA polymerase catalyzes synthesis of RNA from DNA by binding to a sequence termed the promoter located adjacent to or upstream of the coding portion of the linked gene. Genetic and biochemical studies have defined distinct steps in the transcriptional process. The first step of the transcription cycle is initiation, which involves melting (strand separation) of the template DNA and formation of the first chemical bond that makes the RNA. Like a revving engine, the RNA polymerase carries out several rounds of synthesis of short RNA transcripts (termed abortive transcripts) until it becomes disengaged from the promoter (promoter escape) and moves along the DNA, synthesizing RNA at an estimated rate of 1500 bases per minute. This step of the transcription cycle is called elongation, which is followed by termination and re-initiation. Transcription is a highly regulated process; each step requires participation of accessory proteins and is subject to control by transcription factors that respond to intra- and extracellular signals. *See* DEOXYRIBONUCLEIC ACID (DNA); ENZYME; GENE; PROTEIN; RIBONUCLEIC ACID (RNA).

Bacterial transcription. In bacteria, a promoter is defined by two DNA sequence motifs that are located at 10 and 35 bases upstream of the transcriptional start site and are bound by a protein called the sigma factor. The sigma factor is a promoter selectivity factor; most gene promoters in *Escherichia coli* are bound by sigma 54, but additional sigma factors exist that direct transcription from a subset of bacterial genes. In addition to the sigma factor, the bacterial RNA polymerase comprises four subunits. In the presence of ribonucleotide triphosphates, the promoter-bound enzyme alone can initiate RNA synthesis. The transcription from certain genes is turned on and off in response to various signals. This involves proteins called transcriptional activators and repressors that associate with the promoter DNA and influence the activity of the RNA polymerase, resulting in increased or decreased rates of RNA synthesis. In bacteria, transcription factors are commonly partnered with another protein that acts as a sensor of environmental stimuli; this two-component system coordinately regulates RNA synthesis of multiple genes in response to a particular signal. Bacterial genes that encode proteins in a common pathway are transcribed as part of a single large RNA transcript from an operon (a group of distinct genes that are expressed and regulated as a unit) that is translated into individual proteins. In bacteria, transcription and translation are coupled, and therefore regulation of transcription ensures the coordinated synthesis of proteins. *See* BACTERIAL GENETICS; OPERON.

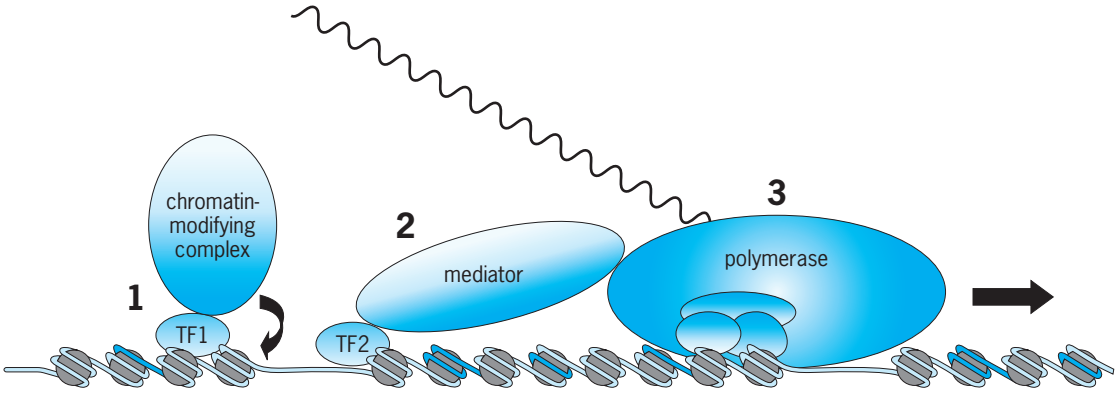
Eukaryotes. The large size of the eukaryotic genome and the complex programs of gene expression required to build multicellular organisms necessitated evolution of more elaborate transcriptional machineries and regulatory schemes, although the basic principles and mechanisms of RNA synthesis remain similar to those in bacteria. Eukaryotes encode three RNA polymerases that transcribe distinct classes of genes. Although the composition of each enzyme is more complex (12 subunits) than the bacterial polymerase, the largest subunits of each eukaryotic RNA polymerase, which carry the copying function, closely resemble the subunits of the bacterial enzyme. RNA polymerase I (Pol I) transcribes a single gene present in many copies that encodes the ribosomal RNA precursor. This takes place in a specialized structure within the cell nucleus called the nucleolus. RNA polymerase III (Pol III) transcribes genes coding for small RNAs such as transfer RNA. RNA polymerase II (Pol II) is responsible for transcribing mostly protein-encoding genes (~30,000 in mammals) and a variety of small noncoding RNA genes. Transcription by Pol II and Pol III takes place in the nucleus, and the RNA is transported to the cytoplasm where protein synthesis takes place. Thus, unlike bacteria, transcription in eukaryotes is not coupled to translation. *See* CELL NUCLEUS; CYTOPLASM.

Regulatory DNA. Eukaryotic gene promoters contain multiple DNA sequence motifs that serve as docking sites for the transcription machinery. To initiate transcription, proteins that specifically associate with the promoter region bind to DNA and in turn recruit other proteins in the transcription machinery that include RNA polymerase to form a stable protein–DNA complex called the pre-initiation complex. The promoter sequences and the factor requirement for each RNA polymerase are distinct; however, mechanistically they work in a similar manner. Each enzyme is brought to the promoter by contacting a protein complex that recognizes and binds

the promoter sequence. The TATA box-binding protein (TBP) serves a function similar to that of the bacterial sigma factor, and it is required for transcription by all three eukaryotic RNA polymerases. TBP was discovered as a transcription factor that binds to an adenine/thymine–rich sequence (TATA box) found in many protein-coding gene promoters transcribed by Pol II. It was later found to be a part of several distinct multiprotein complexes, each serving a specific function for each class of RNA polymerase by binding to promoter DNA. Further, TBP associates with transcription factor IIB (TFIIB) or B-related factor (BRF) to recruit Pol II and Pol III, respectively. The functional importance of these proteins in transcription is underscored by the fact that transcription in archaea requires two proteins related to TBP and TFIIB in addition to RNA polymerase. As in bacteria, the transcription machinery assembled at the eukaryotic gene promoter undergoes each step in the transcription cycle. The DNA template is melted, and the polymerase begins to catalyze the RNA-synthesizing reaction, which initially results in abortive transcripts. A chemical modification on the polymerase results in its dissociation from the promoter-bound transcription factors and transition to the elongation phase that requires a new set of proteins called elongation factors that associate with the enzyme complex. As the polymerase reaches the end of the gene, termination of transcription occurs, and the enzyme dissociates from the template to be recycled for the next round of transcription (re-initiation). It should be noted that the newly synthesized RNA must be appropriately modified and processed prior to being transported to the cytoplasm. Many proteins carry out distinct RNA-processing reactions, and it is now known that these proteins act in concert with the transcription machinery to produce mature RNA in a highly coordinated fashion. *See* NUCLEIC ACID.

Regulation by chromatin. A typical eukaryotic nucleus has a diameter of 6–10 micrometers and contains DNA of about 2 m (6.5 ft) in length. To fit inside the nucleus, the DNA is subject to an extraordinary degree of compaction. Histones are highly conserved eukaryotic proteins that serve to help package DNA. About 146 base pairs of DNA wrap around a core of eight histone proteins to form nucleosomes, which associate with additional proteins and undergo further compaction to form a chromatin fiber. Chromatin is inhibitory not only to transcription but to any cellular process involving DNA such as replication, repair, and recombination. In order to synthesize RNA or to "activate a gene," transcription machineries must contend with chromatin, a process not commonly found in bacteria. Eukaryotes make proteins that specialize in locally altering the structure of chromatin to permit transcription factors and RNA polymerase to bind to DNA and synthesize RNA. Two distinct classes of protein complexes participate in this process: chromatin-modifying complexes that contain enzymes that modify histones in a way to loosen the compacted structure of the nucleosomes; and chromatin-remodeling complexes that use the energy of adenosine triphosphate (ATP) hydrolysis to slide or displace histones from chromatin. In this way these protein complexes facilitate binding and assembly of the transcription initiation complex at the promoter and synthesis of RNA. *See* CHROMOSOME; NUCLEOPROTEIN; NUCLEOSOME.

Gene activation. How is a gene turned on from a transcriptionally silent state in chromatin? Current models based on well-studied genes propose the following scenario (see **illustration**). A transcriptional activator that is capable of recognizing its target DNA sequence when it is associated with chromatin binds to a regulatory region of a gene. The activators are sequence-specific DNA-binding transcription factors that when bound to DNA influence directly or indirectly the RNA polymerase transcription machinery to enhance transcription from the linked gene. The DNA-bound transcriptional activators initially make contact with chromatin-modifying



Transcriptional activation of a eukaryotic gene. Shown is a transcriptional regulatory region (promoter) of a gene; the DNA is wrapped around the nucleosomes. (Step 1) A DNA-binding transcription factor (TF1) binds to a sequence upstream of the transcriptional start site and recruits a chromatin-modifying or -remodeling complex, which alters the structure of the nearby chromatin (curved arrow). (2) This facilitates binding of another transcription factor (TF2) to its target site, which in turn recruits a transcriptional cofactor complex (mediator). (3) The mediator interacts with components of the basal transcription machinery, and a pre-initiation complex is formed at the promoter. Synthesis of RNA (wavy line) by the polymerase ensues (horizontal arrow). The relative importance of these steps and the factors involved in activation (or repression) is specific for each gene.

and/or -remodeling proteins and recruit them so that the nearby region in the chromatin may be altered. This facilitates recruitment of other transcription factors and ultimately the basal transcription machinery, including RNA polymerase to the promoter region to initiate RNA synthesis. Transcriptional activators also interact with intermediary factors or cofactors that do not contact DNA directly but associate with other proteins and components of the transcriptional machinery to assemble a specific protein–DNA complex at the promoter. The participation of dozens of transcription factors allows for combinatorial control and regulatory diversity that is critical for gene-specific regulation of transcription. Transcriptional repressors function to inhibit expression of certain genes or maintain them in a silent state. The mechanisms of repression may involve interfering with the function of activators or the basal machinery, or recruitment of co-repressors that modify histones in order to induce more compacted chromatin structure.

Disease. Modulation of transcription impacts on many areas of medicine, both as a cause and as a cure of disease. Mutations in transcription factors or in gene regulatory sequences have been implicated in numerous developmental and neurological disorders. An example is Rett syndrome, which involves mutations in a protein that regulates transcription. It is now becoming clear that these mutations subtly alter the global program of gene expression in a way that kills a small but vitally important subset of neurons in the brain. Cancer may be thought of as a transcriptional disease, and many cancer-causing proteins (oncoproteins and mutated tumor suppressors) function as transcription factors. The majority of human tumor cells carry mutations that directly affect the function of two critical transcription factors, p53 and the retinoblastoma protein. In normal cells, these proteins provide a protective role by preventing damaged cells from proliferating. They achieve this by blocking the transcription of genes needed for cell duplication and by activating genes that act as a brake to proliferation. Many pharmacological drugs interact directly with DNA-binding transcription factors, leading to activation or repression of specific target genes. One example is the steroid hormones, molecules that bind to a family of receptor proteins that serve as promoter-specific transcription factors. The enzyme complexes that alter chromatin structure are also important targets for anticancer drugs, and they work because rapidly proliferating tumor cells are more sensitive to the changes in gene expression that result from the drug treatment than normal cells. *See* MUTATION; ONCOGENES.

Summary. Transcription is a process essential to any living cell in which the genetic information encoded in the DNA is copied to RNA by the enzyme RNA polymerase. Studies show that the basic mechanism of transcription is essentially conserved among bacteria, archaea, and eukaryotes. However, eukaryotes possess a larger complement of transcription factors and more elaborate machineries and regulatory mechanisms to deal with transcription of a larger number of genes. To maintain the differentiated state, a mammalian cell may transcribe only one-half of its gene repertoire at any given time. The chromatin plays a critical role in keeping the remaining genes silent and permitting specific genes to be activated during processes such as differentiation or the cell cycle. More than 95% of the mammalian genome seems not to code for any genes, and yet they harbor DNA sequences that control transcription of many genes from a long distance. Future studies will be geared toward uncovering the vast amounts of regulatory information (instructions) encoded within eukaryotic genomes to help define the transcriptional regulatory networks responsible for cell differentiation and development of an organism.

Naoko Tanese; Angus C. Wilson

Bibliography. S. Hahn, Structure and mechanism of the RNA polymerase II transcription machinery, *Nat. Struc. Mol. Biol.*, 11(5):394–403, 2004; J. T. Kadonaga, Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors, *Cell*, 116(2):247–257, 2004; M. Levine and R. Tjian, Transcription regulation and animal diversity, *Nature*, 424(6945):147–151, 2003.

Transducer

A device that converts variations in one energy form into corresponding variations in another, usually electrical form. Measurement transducers or input transducers may exploit a wide range of physical, chemical, or biological effects to achieve transduction, and their design principles usually revolve around high sensitivity and minimum disturbance to the measurand, that is, the quantity to be measured. Output transducers or actuators are designed to achieve some end effect, for example, opening of a valve or deflection of a control surface on an aircraft. Actuators, therefore, normally operate at high power levels. The term sensor is often used instead of transducer, but strictly a sensor does not involve energy transformation; the term should be reserved for devices such as a thermistor, which is not energy-changing but simply changes its intrinsic electrical resistance in response to changes in temperature.

Vast numbers of transducers are in common use in the domestic environment, that is, in central heating systems, ovens, and refrigerators; in transport; in all forms of industry; and in medicine (see **table**). Extremely sophisticated devices are employed for the measurements needed in the pursuit of scientific research.

Both input and output transducers, together with the instrumentation to which they are connected, may be called upon to respond to both quasistatic (that is, slowly varying) signals or dynamic (that is, more rapidly varying) signals. This means that the transducer, together with its instrumentation system, must be designed to meet such a specification. Some prior knowledge is therefore required of the type of signal to be transduced, and the bandwidth

Principles of transducers and examples of their applications

Variable	Transduction principle	Example
Temperature	Thermoelectric	Thermocouple; thermopile
	Thermocapacitive	Change of capacitance (of, for example, ceramic capacitors)
	Thermoresistive	Thermistors; platinum resistance thermometer; semiconductor (silicon temperature sensors)
	Thermal expansion	Glass thermometer; bimetallic strip; quartz crystal (frequency change)
	Thermal radiation	Thermopile; thermistor; photovoltaic semiconductor; photoconductive semiconductor
	Thermochemical	Liquid crystals (temperature-sensitive chemical changes)
	Pyroelectric	Lead zirconate titanate (polarization change with temperature)
Displacement		
Low-mechanical-impedance types	Resistive	Potentiometer (rotational and linear)
	Capacitive	Change in capacitance
	Inductive	Change in inductance
	Transformer	Alternating-current-excited differential transformer
	Electromagnetic	Movement of coil through a magnetic field (no dc response)
	Optical	Moiré fringe; interferometry; reflection change
	Ultrasonic	Pulse transit time; phase shift
High-mechanical-impedance types	Piezoresistance	Very small displacements
	Mechanoresistance	Metal foil and wire strain gage; mercury gage
Velocity	Magnetic induction	Moving coil (gives output proportional to velocity)
	Doppler effect	Ultrasonic; optical (electromagnetic)
Acceleration	$f = ma$	Measurement of the force due to the movement of a known mass (sensing via piezoelectric, magnetostrictive, or mechanoresistive transducers)
Force	Elastic displacement	Primary: displacement of an elastic member Secondary: displacement transducer (as listed above)
	Piezoresistive	Semiconductor (for example, silicon)
	Piezoelectric	Barium titanate (ferroelectric ceramic, lacks dc response)
	Magnetostrictive	Induced voltage type (lacks dc response); variable inductance type
	Balance (feedback)	Electromagnetic force balance (null-type system)
Pressure	Elastic displacement	Diaphragm deflection (Bourdon tube, bellows, and so forth) measured with a displacement transducer
	Force balance	Manometer; electromagnetic force balance
Pressure (acoustic)	Capacitive	Moving plate (capacitor)
	Inductive	Moving coil
	Piezoelectric	Lead zirconate titanate
	Piezoresistive	Silicon diaphragm
Chemical	Gases; volatile chemicals	Piezoelectric; metal oxide; flame ionization; photionization
	Chemicals in solution	Absorption; fluorescence; luminescence
	Biochemical	Redox system; immunooptical; evanescent wave; enzyme-linked immunosystem
Atmosphere	Pollution monitoring	Light detection and ranging (LIDAR)
Flow	Pressure difference	Orifice type
	Mechanical	Rotameter: driving an electromagnetic emf-generating secondary transducer
	Thermal	Heat transport from aheated element (for example, self-heated thermistor)
	Electromagnetic	Electromagnetic flow (emf) generated by conducting liquid flowing through a magnetic field
	Ultrasonic	Pulse transit time; Doppler shift; cross correlation
	Optical	Laser Doppler shift
	Indicators	Dye dilution; thermal dilution; conductivity dilution
	Transit time	Marker transit time (bubbles, solids and so forth in carrying fluid)
Electromagnetic radiation	Gamma rays, x-rays, and so forth	Germanium or silicon detectors (cooled); silicon carbide detectors

of the transducer and instrumentation system must be suitably matched to this signal.

Transducers are often described in terms of their sensitivity to input signals; more properly this should be called responsivity. This is simply defined as the ratio of the output signal to the corresponding input signal. Once again, the responsivity of a transducer must be matched to the expected levels of signal to be transduced. *See* SENSITIVITY (ENGINEERING).

Displacement transducers. The measurement of displacement is very often part of the transduction process involved in the measurement of other quantities, such as acceleration, force, or pressure. Displacement transducers can be divided into two categories, low- and high-mechanical-impedance devices. The low-mechanical-impedance displacement transducer (large deflection, low force) is often based on the measurement of changes in electrical

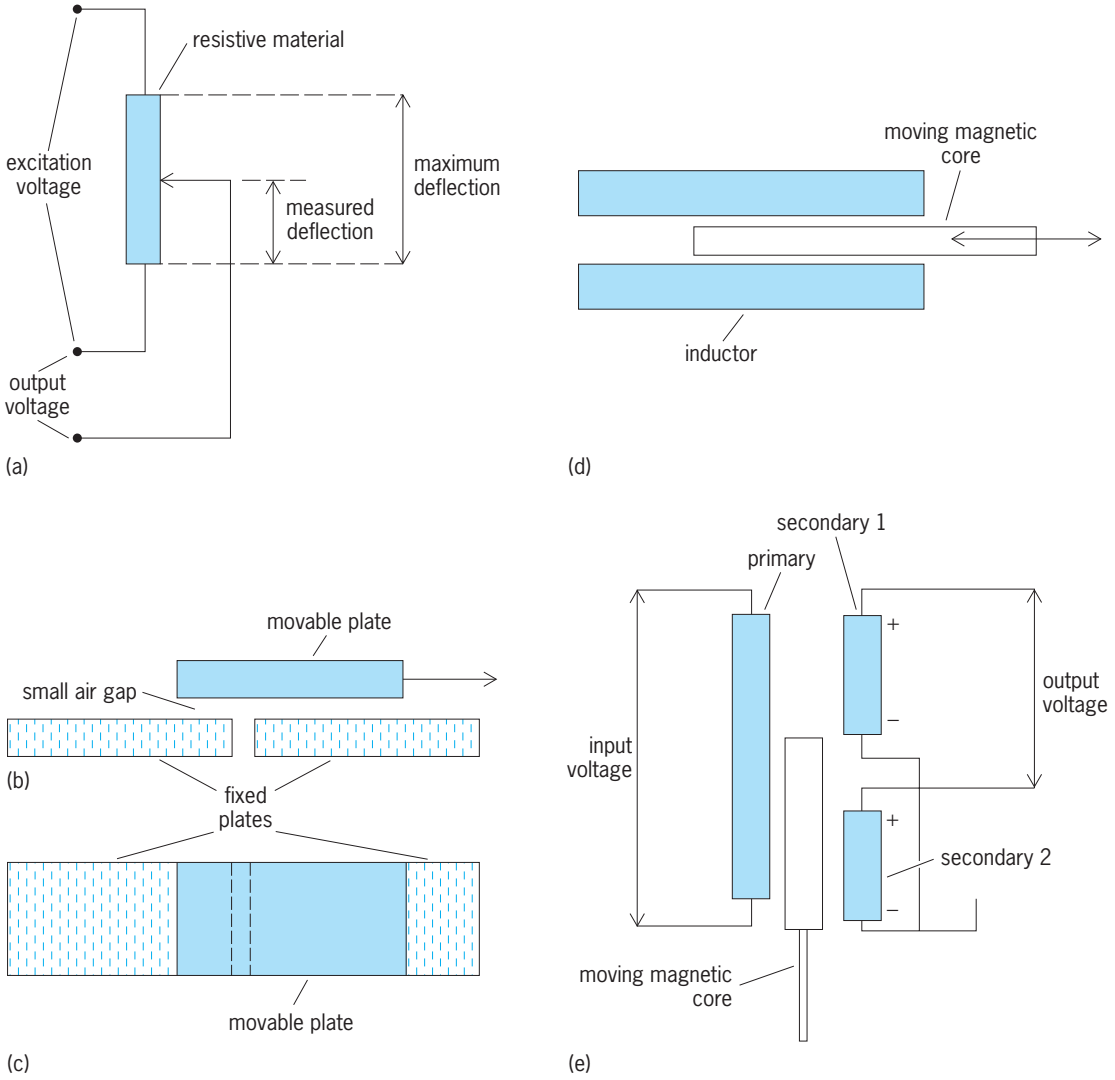


Fig. 1. Examples of displacement transducers. (*a*) Potentiometric transducer for linear deflection. Multiturn devices for rotational deflection measurement are also in common use. (*b*) Elevation and (*c*) plan view of capacitance transducer for linear deflection. Output voltage is from the moving plate with respect to the static plates. (*d*) Simple variable inductance for transducing deflection. (*e*) Circuit diagram of a linear variable differential transformer (LVDT) transducer.

resistance, capacitance, or inductance, and is usually employed for relatively large displacements, typically between a few millimeters (a fraction of an inch) up to several meters. For these devices, a restoring force is usually provided by a relatively soft spring in order not to disturb the measurand. Both linear (or translational) and rotational devices are readily available (**Fig. 1**). *See* CAPACITANCE; CAPACITANCE MEASUREMENT; INDUCTANCE MEASUREMENT; MECHANICAL IMPEDANCE; POTENTIOMETER; RESISTANCE MEASUREMENT; TRANSFORMER.

High-mechanical-impedance devices are associated with the measurement of very small displacements and might be based on, for example, the deformation of very stiff springs on which are mounted resistance strain gages. The materials used for the springs should have a linear stress–strain relationship over the range of stresses to be encountered. Other material properties, such as stability over the temperature range to be encountered, low creep, low relaxation, and low hysteresis, are also essential. Various specialty alloys have been developed for these purposes. Other high-mechanical-impedance transducers may be based on materials that are piezoresistive or piezoelectric. Piezoresistive materials such as heavily doped silicon can be used as the basis for strain gages that have very high sensitivities. *See* METAL, MECHANICAL PROPERTIES OF; PIEZOELECTRICITY; SPRING (MACHINES); STRAIN GAGE.

Force transducers. The measurement of force is very often accomplished by allowing an elastic member (spring or cantilever beam) to deflect and then measuring the deflection by using some form of displacement transducer. Two methods of achieving very high stiffness are available. One is to use an elastic member with an intrinsically small deflection for a given force. Very often, strain gages are employed to detect the small displacements involved (**Fig. 2**). The alternative method is to use feedback techniques so that the deflection caused by the

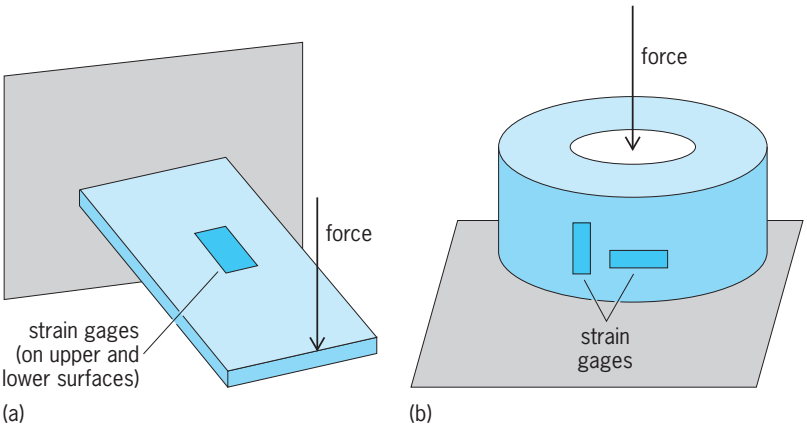


Fig. 2. Examples of force transducers. (*a*) Cantilever beam using strain gages to measure force. (*b*) Compression column plus strain gages to measure force.

application of the force generates a signal that restores the part of the transducer originally deflected to its quiescent position. Force balance systems, as these are often termed, have the advantage of good linearity. *See* LINEARITY.

Acceleration transducers. Transducers designed to measure acceleration are frequently based on the simple equation $f = ma$, where f is force, m is mass, and a is acceleration. Thus, if the force due to the movement of a known mass can be measured, it is possible to derive the acceleration. Very often, the measurement technique employed uses piezoelectric, magnetostrictive, or mechanoresistive materials (**Fig. 3**). Acceleration transducers or accelerometers are frequently employed for the measurement of vibration. *See* ACCELEROMETER; FORCE; MAGNETOSTRICTION.

Velocity transducers. Linear velocity transducers can be based on the magnetic induction effect or on the exploitation of the Doppler effect. In this latter case, the energy form used can be ultrasonic or electromagnetic. The velocity of rotating shafts can also be measured by an optical encoder with a suitable light source and detector. By choosing an appropriate pattern, the output data can be produced in binary form suitable for direct input to a computer system. This is one of the few examples of a digital transducer. (In general, the measured world is entirely analog and the transducers used to monitor it are also analog.) Rotational velocity transducers can also be based on the magnetic induction effect, using a coil rotating in the field of a permanent magnet and slip rings to detect the voltage proportional to velocity. Another approach is based on alternating currents: an exciting coil is fed with a constant-frequency signal and a detector coil picks up a voltage proportional to the rate of rotation of a cylinder, which is subject to eddy-current effects. *See* DOPPLER EFFECT; EDDY CURRENT; ELECTROMAGNETIC INDUCTION; SERVOMECHANISM.

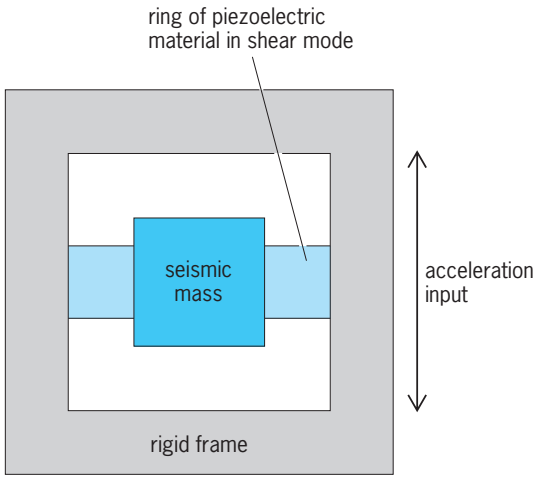


Fig. 3. Simplified diagram of a piezoelectric accelerometer.

Pressure transducers. Pressure measurement is often achieved by detecting the deflection of an elastic diaphragm or tube (for example, a Bourdon tube). The requirements in terms of material properties for such diaphragms or tubes are similar to those of restoring springs in high-mechanical-impedance displacement transducers. Semiconductor fabrication techniques have been applied to pressure transducers. A diaphragm of monocrystalline silicon possesses most of the elastic properties required, and a strain-gage bridge exploiting the piezoresistive properties of silicon can be fabricated on the diaphragm. This form of construction lends itself to microminiaturization. *See* PRESSURE MEASUREMENT; PRESSURE TRANSDUCER.

Temperature transducers. The most familiar instrument for measurement of temperature is the mercury-in-glass thermometer. Here, the measurement is really that of the expansion of the mercury as temperature increases. This expansion is magnified by employing a capillary tube in which to observe the changes in volume with temperature. This instrument has certain disadvantages as a transducer in that no electrical signal is produced. When such a signal is required, measurements may be based on either changes in the electrical resistance of certain materials or upon the thermoelectric effect.

Temperature transducers based on changes in the electrical resistance of platinum can be used over a very wide range (−200 to 900°C or −330 to 1650°F), and they are extremely stable and linear. They can be fabricated as a thin film on a ceramic substrate and are usually used with a bridge circuit. However, they produce a relatively small signal in response to temperature change (0.4 Ω/°C or 0.2 Ω/°F). *See* BRIDGE CIRCUIT.

A much higher output signal for a given temperature change is produced by devices known as thermistors, which are fabricated from semiconducting materials (for example, oxides of cobalt, nickel, and manganese). However, these are inherently nonlinear in terms of resistance change with temperature and have a relatively smaller useful range (−100 to 200°C or −150 to 390°F). They can be made fairly small and therefore have short thermal time constants, making it possible to follow rapid fluctuations in temperature. The nonlinear behavior of

thermistors can easily be overcome when they are operated in conjunction with a microprocessor. *See* MICROPROCESSOR; THERMISTOR.

Much smaller devices can be produced from thermocouples. These are transducers based on junctions between two dissimilar wires, one of which is held at a constant reference temperature and the other used to effect the measurement. The disadvantage of requiring a reference temperature can be simply overcome by electronic compensation. The disadvantage of the thermocouple is its very low output signal for a given temperature change (10–80 microvolts/°C or 5–40 μV/°F). However, a very wide temperature range is available. *See* THERMOCOUPLE.

A silicon diode has a temperature coefficient of about −2 mV/°C (−1 mV/°F) and can therefore be used as a temperature transducer, although in practice each *pn* junction needs individual calibration. However, silicon diodes have the advantage of being very inexpensive. Many other configurations of silicon-based temperature transducers have been developed using the properties of silicon, such as the temperature dependence of carrier transport, and sometimes employ numerous transistors on a single chip. *See* JUNCTION DIODE; TEMPERATURE MEASUREMENT; THERMOMETER.

Flow transducers. The measurement of flow is of considerable importance in industrial processes, and mass-flow measurement is often even more important since it may be used as the basis for metering valuable commodities (for example, oil and gas) for subsequent payment.

A wide variety of ingenious devices have been produced for flow measurement based on, for example, pitot tubes and orifice plates. The principle of both these devices is that of a pressure difference that can be transduced into a usable voltage signal. *See* FLOW MEASUREMENT.

Measurement of mass flow can be achieved by simultaneous measurement of both flow and density, but mass-flow transducers have been developed in which the liquid is made to flow through a cylinder that is forced into vibration. Differences in mass flow alter the vibrational properties of the cylinder, and these may be detected.

Chemical transducers. Transducers for a wide range of chemical species are available, but probably the most widely applied is the pH transducer for the measurement of hydrogen-ion concentration. The traditional method has relied on a glass membrane electrode used to make up an electrochemical cell. *See* HYDROGEN ION; ION-SELECTIVE MEMBRANES AND ELECTRODES; PH.

Measurements of the partial pressure of oxygen (pO_2) may be accomplished by the use of a Clark oxygen cell, which comprises a gas-permeable membrane controlling the rate of arrival of oxygen molecules at a noble-metal cathode that is held at 600–800 mV potential with respect to the anode. The ensuing reduction process gives rise to a cathode current from which oxygen concentration can be derived.

Other electrochemical transducers are based on similar principles and are used in such applications as voltametry, polarography, and amperometry. Chemical transduction is also possible by adsorbing a species onto a surface and detecting its presence by mass change, electrical property change, color change, and so on. *See* ELECTROCHEMICAL TECHNIQUES; POLAROGRAPHIC ANALYSIS; TITRATION.

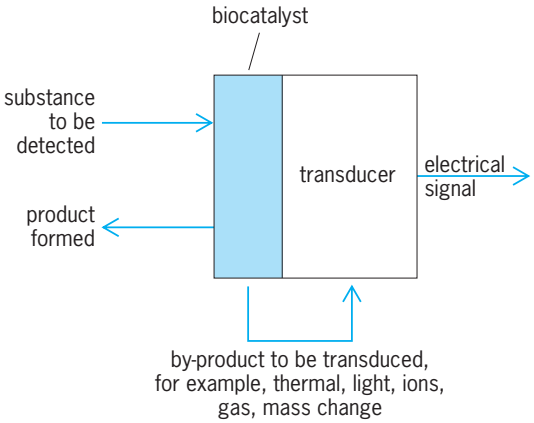


Fig. 4. Schematic diagram of a biosensor, showing the principle of operation.

Biological transducers. Measurements of the partial pressure of oxygen and the partial pressure of carbon dioxide (pCO_2) are also of particular importance in the context of blood gas analysis in medicine, and by using the Clark cell they can be performed without removing the blood from the body and noninvasively, that is, without puncturing the skin. However, it is necessary to raise the temperature of the surface of the skin onto which the Clark electrode is placed by a few degrees in order to promote permeation of oxygen (O_2) or carbon dioxide (CO_2) molecules to the cell membrane.

There have been remarkable advances in the area of biological transducers or biosensors (**Fig. 4**). Examples are the ion-selective field-effect transducer (ISFET), the insulated-gate field-effect transducer (IGFET), and the chemically sensitive field-effect transducer (CHEMFET). Many of these devices are still in the research stage, but it is almost certain that they will attain major importance in medicine, introducing a new approach to long-term patient monitoring in which a host of biochemical measurements are made at the bedside rather than in the pathology laboratory. The major advantage that this provides is speed of diagnosis and treatment. Similar devices will play a major role in measurements associated with biotechnology. Already, pH-sensitive ISFETs are widely used. The application of nanotechnology to biosensors is beginning to make possible a wide range of new applications in medical and biological research. *See* BIOELECTRONICS.

Advanced devices. Initially, the microelectronics revolution had little impact on transducer development. Efforts made to remedy this have yielded interesting results.

Solid-state and smart transducers. Considerable advantage can be obtained in many cases by placing the sensing process as close to the signal conditioning

as possible. This is particularly true of signals at very high frequencies or very low levels (because of the possibility of noise corruption in connecting cables). There have been proposals for implementing this, and silicon-based transducers are ideal for this application. These might make use of the piezoresistive, photovoltaic, or Hall effects for transduction. *See* HALL EFFECT; PHOTOVOLTAIC EFFECT.

A smart transducer or smart sensor is a device that not only undertakes measurement but also can adapt to the environment in which it is placed. Suchadaptation may range from simple changes in the characteristics of the transducer in response to changes in temperature, to more complex procedures such as adaptation of the transducer's performance to conform to overall system requirements. This may call for changes in linearity, sensitivity (or responsivity), stiffness, and so forth. In integrated transducers, much of the signal processing that might previously be done remotely is brought into the transducer packaging. This approach is most beneficial when high-frequency signals are involved, such as in an ultrasonic transducer array.

Systems of transducers. Systems have been designed that contain several transducers of various types that respond to differing parameters of the same measurand or display relatively broadband selectivity to the measurands. The cross-sensitivity between outputs of such transducers contains valuable information that may allow unique patterns of response to be obtained from such systems, and inherently poorly selective transducers to function as a group to yield highly selective information.

Fiber-optic transducers. The development of inexpensive fiber-optic materials for communications has led to an examination of the potential for using these devices as the basis for transduction. Two major types of devices have resulted: fiber-optic transducers for physical variables and similar devices devoted to chemical and biological determinations. The advantages of the all-optical transducer are its lack of susceptibility to electrical interference and its intrinsic safety. Small deformations of an optical-fiber waveguide cause a change in the light transmission of the fiber, and this has been exploited to produce force and pressure transducers. Alternatively, miniature transducers based on color chemistry can be fabricated at the end of a fiber and the color change can be sensed remotely. Devices of this type have been developed for measuring pH, the partial pressures of oxygen and carbon dioxide, and glucose. *See* FIBER-OPTIC SENSOR.

Sensors for extreme environments. Silicon carbide (also known as carborundum) is a semiconductor originally developed for high-power electronics applications. It displays a wide band gap compared with silicon which enables it to operate as a semiconductor at temperatures up to 1000°C (1800°F), whereas silicon is limited to a maximum of about 175°C (350°F). The binding energy between the carbon and silicon atoms in the lattice is much greater than that between silicon atoms, thus making silicon carbide radiation hard, leading to its use in radiation detectors. These properties also make the material useful for sensor applications in areas such as car exhaust and aerospace engine gas sensing and in the vents of volcanoes as a means of early warning of major activity. Silicon carbide has extremely low chemical activity and is thus potentially very useful for medical sensors and for the chemical engineering industry. The current challenge is to develop complex, inexpensive logic circuits for use with such sensors. Major progress has been made with wafers with few defects, and diameters up to 10 cm (4 in.) are available.

Micro-electro-mechanical systems. The most important recent technological development in the area of transducers, sensors, and actuators is micro-electro-mechanical systems (MEMS). There are a wide variety of MEMS devices, mostly fabricated in silicon. In general, MEMS-based transducers use conventional techniques to achieve transduction; for example, accelerometers are often based on a proof mass and cantilever principle with a geometry similar to that of Fig. 2*a*. At micrometer dimensions, care must be taken to ensure that a complete understanding of the mechanical behavior of such a transducer has been gained, since additional nonlinear behavior sometimes arises at these dimensions. Inexpensive silicon MEMS gyroscopes are in the second generation of commercial production, with emphasis on volume manufacture. *See* GYROSCOPE.

An injectable microsensor has been developed that is capable of transmitting signals from the motor cortex of the human brain using wireless technology. The development of such sensors is likely to be a major trend in the area of MEMS and indeed NEMS (nanometer electromechanical systems).

Considerable effort is directed at making MEMS devices that are compatible with complementary metal-oxide-semiconductor (CMOS) technology, and micromachined accelerometers giving 125 mV per *g* and a bandwidth of 2.5 kHz have been described. There is also interest in nanometer-scale devices using silicon-on-insulator (SOI) technology, giving rise to active devices working at below 1 V. MEMS-based microphones are capacitive microphones fully integrated with CMOS amplifiers and dc-to-dc converters. *See* INTEGRATED CIRCUITS; MICROPHONE; TRANSISTOR.

Microactuators using a comb drive or interdigitated electrode geometry are employed for accelerometers, gyroscopes, and micromirrors. Considerable activity is focused on the surface micromachining of micromirror hinges. Displacement sensors based on integrated optics and micro-opto-electro-mechanical systems (MOEMS) now play a major role in adaptive optical devices. *See* ADAPTIVE OPTICS; INTEGRATED OPTICS; MICRO-ELECTRO-MECHANICAL SYSTEMS (MOEMS).

Silicon micromachining is also used for constructing scanning confocal optical microscopes and, in a similar manner, modified atomic force microscopes have been described for potential use in data storage. These devices have a storage capacity of 50 times that of the CD-ROM. *See* CONFOCAL MICROSCOPY; SCANNING TUNNELING MICROSCOPE.

Gas-flow sensors have been described using air-foil plates, 30 μm thick, made of single-crystalline silicon. Other gas sensors have been micromachined to give rise to shear-stress sensors. *See* FLOW MEASUREMENT.

Micro-gas turbines based on hydrogen combustors micromachined from silicon have been manufactured. They were produced using the deep reactive ion etching (DRIE) technique, and a 150-W electrical power microcombustor has been demonstrated, where the device itself is only 0.066 cm^3 in volume. *See* GAS TURBINE.

The polymerase chain reaction (PCR) is one of the most widely used techniques for the replication of deoxyribonucleic acid (DNA) samples. PCR using micromachined structures gives rise to improved performance and much reduced reagent use. This is one of many developments leading to the so-called laboratory-on-a-chip. In addition, work on microvalve arrays has been described and, in a similar area, MEMS technology has been used to produce microneedles and microchannel-based electrical impedance spectrometers. *See* GENE AMPLIFICATION.

Considerable work is also being done on improving large-volume production techniques for MEMS-based devices. One approach is to attempt to integrate everything onto one device, as in the MEMS-based microphone. A second approach is to separate out the microelectronics, usually CMOS-based, from the microsensors and microactuators, which usually require materials that are not permitted in integrated-circuit production facilities. The two devices are then brought together, in one case using flip-chip bonding techniques. Improved production is also being sought through the use of microstamping techniques. Work is also going on to look at ways of sealing microtransducers and microsensors; in one approach, chemical vapor deposition techniques are used. *See* CONTROL SYSTEMS; CRYSTAL GROWTH; ELECTRONIC PACKAGING; INSTRUMENTATION AMPLIFIER; MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); MICROSENSOR; PHYSICAL MEASUREMENT.

Peter A. Payne

Bibliography. R. C. Asher, *Ultrasonic Sensors for Chemical and Process Plant*, IOP Publishing, Bristol, 1997; E. O. Doebelin, *Measurement Systems: Application and Design*, 5th ed. McGraw-Hill, New York, 2003; P. Hauptmann, *Sensors: Principles and Applications*, Prentice Hall, Hemel Hempstead, UK, 1993; G. T. A. Kovacs, *Micromachined Transducers Sourcebook*, WCB/McGraw-Hill, Boston and London, 1998; H. K. P. Neubert, *Instrument Transducers: An Introduction to Their Performance and Design*, 2d ed., Oxford University Press, Oxford, 1975; S. Soloman, *Sensors Handbook*, McGraw-Hill, New York, 1998; T. Togawa, T. Tamura, and P. Å. Öberg, *Biomedical Transducers and Instruments*, CRC Press, Boca Raton, 1997; P. F. Turner, I. Karube, and G. S. Wilson (eds.), *Biosensors: Fundamentals and Applications*, Oxford University Press, New York, 1987; M. J. Usher and D. A. Keating, *Sensors and Transducers: Characteristics, Applications, Instrumentation, Interfacing*, 2d ed, Macmillan,London, 1996.

Transduction (bacteria)

A mechanism for the transfer of genetic material between cells. The material is transferred by virus particles called bacteriophages (in the case of bacteria), or phages. The transfer method differentiates transduction from transformation. In transformation the genetic material (deoxyribonucleic acid) is extracted from the cell by chemical means or released by lysis. *See* BACTERIOPHAGE; DEOXYRIBONUCLEIC ACID (DNA); TRANSFORMATION (BACTERIA).

Transduction has been demonstrated only in certain bacterial species, such as *Salmonella*, *Escherichia*, *Staphylococcus*, and *Haemophilus*. Even within these species it occurs only within certain strains. There is, however, no reason to assume that it is not more prevalent, and it may be that only technical limitations, such as the appropriate phage, prevent its demonstration in other forms.

The transduction mechanism has two features to distinguish it from the more usual mechanism of gene recombination, the sexual process. The most striking feature is the transfer of genetic material from cell to cell by viruses. The second distinguishing feature is the fact that only a small part of the total genetic material of any one bacterial cell is carried by any particular transducing particle. However, in general transduction, all of the genetic material is distributed among different particles.

Transducing bacteriophage. Transduction is not accomplished by all bacteriophages. It is done by some that are classified as "temperate." When such temperate bacteriophages infect sensitive bacteria, some of the bacteria respond by producing more bacteriophage particles. These bacteria donate the transducing material. Other bacteria respond to the infection by becoming more or less permanent carriers of the bacteriophage, in a kind of symbiotic relationship; these are called lysogenic bacteria. Bacteria in this latter class survive the infection, and it is among these that transduced cells are found. The proportion of bacteria in any culture that responds to infection in either manner can be influenced by the particular environment at the time of infection. Thus, when the goal is to obtain transducing activity, the culture is influenced to grow more bacteriophage. When the study of the transduction is desired, the culture is influenced to give more lysogenic responses. *See* LYSOGENY.

General transduction. General transduction is unrestricted as to the genetic property which can be transferred. The only limitation results from the fact that the frequency of cells transduced for any particular property is rather low. It is convenient to express the transducing activity of a preparation in terms of the number of transductions produced per particle capable of acting as an infective agent, or the number of transductions per viable viral particle. For any character, this number is of the order of one

transduction for every million particles. Since only relatively few bacteriophage particles can infect a single bacterium, the number of transduced bacteria for the studied property is not much larger than one in a million. However, with bacterial populations it is relatively simple to find bacteria of a unique kind even at such a low frequency. Bacteria can be grown to numbers a thousand times more than a million. If the property transduced can be disclosed by selection, there is no problem in finding the transduced cells.

Linked transduction. Although the general rule is that individual properties are transduced independently of each other and occasionally replace their genetic homologs, instances of simultaneous or linked transduction of two properties are found, indicating that the transducing fragments encompass more than one gene. Transduction has afforded investigators an opportunity to look carefully at small regions of genetic material. These studies have shown that very often genes controlling sequential biosynthetic steps are linked; that is, they are very close together on the bacterial chromosome. Transduction has also been of use in further demonstrating that the gene controlling an individual biosynthetic step can be subdivided by mutation and then reconstituted by genetic recombination. Thus many mutations, all affecting the synthesis of a specific enzyme, can be shown to lie at different points on the same gene. They can transduce each other to a nonmutant form. *See* BACTERIAL GENETICS.

Restricted transduction and oncogenesis. In addition to general transduction, certain phages carry out a more restricted kind of transduction. The restriction is that they carry only a specific section of bacterial genetic material; they transduce only a few genes. In this restricted case it has been shown that the transducing particle has exchanged a piece of its own genome for a piece of bacterial genome. The particular piece of bacterial genome is the one that the phage is linked to when it is carried by the bacteria.

Transduction thus affords investigators further insight into the nature of viruses and the fine structure of genetic material. *See* VIRUS.

Retroviruses also carry out specific or restricted transduction. It has long been known that these viruses can cause the formation of tumors (oncogenesis) in animals. (The first Rous sarcoma, a tumor virus that affects chickens, was discovered in 1905.) It is now known that these viruses exchange a small portion of their genome for a mutant cellular gene that has a role in gene regulation or replication. These viruses carrying mutant genes infect cells, causing them to be transformed into tumor cells. *See* ANIMAL VIRUS; RETROVIRUS; ROUS SARCOMA.

Transduction and evolution. Analysis of the DNA sequences of many different microorganisms has revealed that they are mosaics of different genomes, indicating that there was a great deal of horizontal gene transfer in their evolution. Transduction was probably the major vehicle for such exchanges, as many bacteria contain a myriad of pieces of phage genes.

Norton D. Zinder

Transfer cells

Plant cells characterized by the elaboration of an unlignified, secondary cell wall to form fingerlike projections or wall ingrowths which protrude into the cytoplasm of the cell (**Figs. 1** and **2**). These ingrowths are enveloped by plasma membrane (Fig. 2), forming a wall-membrane apparatus. The extent of wall ingrowth development varies from small, conical projections, through long, unbranched or branched structures, to Y-shaped flanges. In all cases their formation increases the surface-volume ratio of the cell, which implicates them in short-distance solute transport. The surface area of the plasma membrane may be increased up to 20 times that of a similar-sized cell lacking wall ingrowths. Most transfer cells elaborate numerous mitochondria (Fig. 2), many in close association with the wall ingrowths, suggesting that the transport process is energy-requiring. An enlarged, lobed nucleus (Fig. 2) and numerous endoplasmic reticulum cisternae and ribosomes are often characteristic of the cytoplasm of transfer cells. Plasmodesmata (cytoplasmic channels for intercellular transport) connect transfer cells to other transfer cells and to parenchyma cells (Fig. 2).

Location and function. The location of transfer cells within the plant provides circumstantial evidence for their involvement in solute transport. They are situated at many sites in the plant where secretion or absorption takes place. Transfer cells are most frequently associated with the conducting elements of the xylem and phloem (elements used for long-distance transport in the plant), where they may play a role in either loading or unloading solutes from these cells. Minor veins of angiosperm leaves have particularly well-developed phloem transfer cells, whereas stem nodes often possess numerous xylem transfer cells. With an increasing number of ultrastructural studies of plant roots, it is apparent that transfer cells are more prevalent in this organ than previously recognized. *See* PHLOEM; XYLEM.

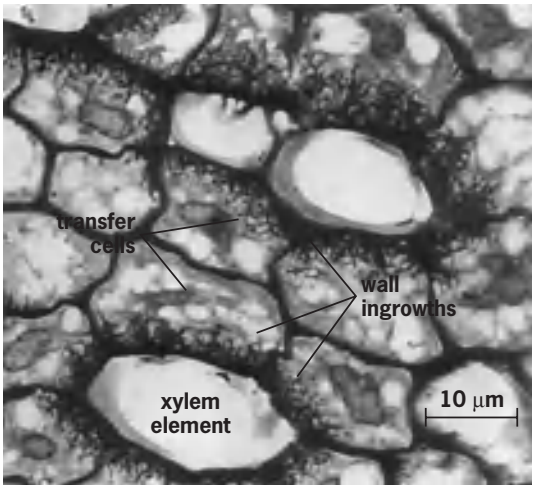


Fig. 1. Light micrograph of transfer cells adjacent to xylem elements in *Hieracium florentinum* rhizome. Long, branched wall ingrowths are evident.

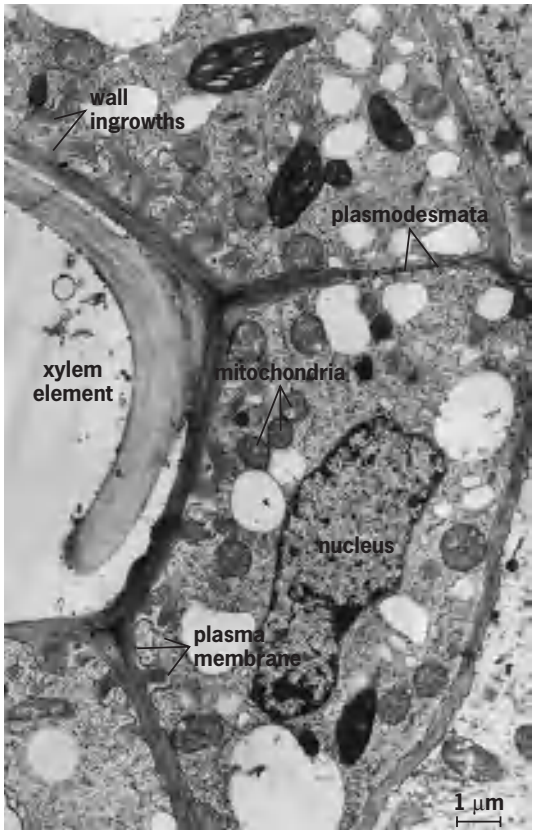


Fig. 2. Electron micrograph of transfer cells adjacent to a xylem element in *Hieracium florentinum* rhizome. Wall ingrowths enveloped by plasma membrane are located next to xylem element.

Four main categories of secretion or absorption involving transfer cells are generally recognized: (1) absorption of solutes from the external environment, with the epidermis of submerged leaves and the hydropotes ("water-drinking" glands) on the surface of water lily leaves being examples; (2) secretion of solutes to the external environment, as illustrated by various secretory glands and nectaries; (3) absorption of solutes from internal, extracytoplasmic compartments, as exemplified by transfer cells associated with tracheary elements of the xylem; and (4) secretion of solutes into internal, extracytoplasmic compartments, as shown by the secretion of materials from the tapetum into the locule of the anther in flowers. *See* SECRETORY STRUCTURES (PLANT).

Transport within the plant can occur from cell to cell either through plasmodesmata or through extracytoplasmic structures such as the cell wall and intracellular spaces. The former mode of transport is termed symplastic, while the latter is apoplastic. Transfer cells are thought to be involved in exchanges of solutes between these two systems. Wall ingrowths in transfer cells are porous enough to allow movement of lanthanum nitrate (a large molecule frequently used to show transport in the apoplast) and therefore presumably many other molecules of lower molecular weight. The plasma membrane provides the greatest resistance to solute movement. The occurrence of adenosine triphosphatases associated with the plasma membrane enveloping the wall ingrowths implies that active transport of solutes may be involved.

The exact mechanism by which transfer cells mediate exchange between the symplast and apoplast is, however, unknown. The early suggestion that these cells might increase absorption or secretion simply as a consequence of their increased surface area is probably too simplistic in some cases. Emphasis is being placed on the effects of wall ingrowth shape on transport processes. Wall ingrowths, by providing membrane-bound cytoplasmic clefts between them, may allow for the development of a standing osmotic gradient within the clefts. Primary carriers (porters) may then carry ions such as hydrogen (H^+), potassium (K^+), and chloride (Cl^-) across the membrane, thereby establishing electrical gradients. The presence of adenosinetriphosphatases on this membrane surface would be compatible with this energy-requiring step. Secondary porters driven by electrochemical gradients established by the primary porters may carry other ions or organic molecules across the membrane.

Regardless of the mechanism involved in transmembrane transport into or out of transfer cells, these cells have been shown to be capable of absorbing substances from the apoplast. Among several examples, hydropotes on water lily leaves have been shown by autoradiography to absorb sulfate ions from an external bathing medium, and transfer cells on leaves of carnivorous plants have been shown by autoradiographic and x-ray microanalytical techniques to absorb products from the digestion of insects. *See* ION TRANSPORT; PLANT TRANSPORT OF SOLUTES.

Differentiation. Although transfer cells are spectacular examples of cell differentiation within the plant, there is little firm evidence as to factors controlling their differentiation. In some cases, wall ingrowths are formed, or are increased in number, at the onset of solute transport, thus suggesting that solute presence might trigger parenchyma cells to form wall ingrowths. For example, in young pea leaves the appearance of wall ingrowths in phloem parenchyma cells coincides with export of photosynthates from the leaf. However, in other situations wall ingrowths develop during the normal ontogeny of the cell or tissue in question whether or not solutes are present for transport. An example of this was described in root nodule development in the pea. In effective nodules, wall ingrowths form in xylem parenchyma and in the pericycle prior to the onset of nitrogen fixation and the formation of nitrates for transport. Ineffective nodules, which are incapable of fixing nitrogen and therefore do not produce nitrates, still develop transfer cells in both the xylem parenchyma and pericycle. Hormonal or other morphogenetic stimuli may play a role in inducing wall ingrowth formation in parenchyma cells, but these factors are largely unknown. *See* PARENCHYMA.

Occurrence in plant groups. Transfer cells occur in diverse groups of plants, including algae, fungi, mosses, ferns, gymnosperms, and angiosperms.

Their development in presumably unrelated plant groups suggests that plants evolved this mechanism more than once to solve short-distance transport problems. The most highly evolved plant group, the angiosperms, probably has the highest frequency of species with transfer cells, but there are many angiosperm families which normally do not develop this cell type. However, the genetic information for transfer cell development may be present in species even though transport cells do not usually develop. This is shown most clearly in the roots of some species which do not normally develop transfer cells but do so when they become infected by nematodes. Root cortical cells in proximity to the head of the nematode enlarge, develop wall ingrowths, and become giant transfer cells. In some angiosperm families the presence or absence of transfer cells may be a useful taxonomic character when correlated with other morphologic characters. *See* CELL WALLS (PLANT); PLANT CELL.

R. L. Peterson

Bibliography. E. G. Cutter, *Plant Anatomy*, pt. 1: *Cells and Tissues*, 2d ed., 1978; N. Harris and K. J. Oparka (eds.), *Plant Cell Biology: A Practical Approach*, 1994; J. S. Pate and B. E. S. Gunning, *Annu. Rev. Plant Physiol.*, 23:173–196, 1972.

Transform fault

One of the three fundamental types of boundaries between the mobile lithospheric plates that cover the surface of the Earth. Whereas spreading centers mark sites where crust is created between diverging plates, and subduction zones are where crust is destroyed between convergent plates, transform faults separate plates that are sliding past each other with neither creation nor destruction of crust. The primary tectonic feature of all transform faults is a strike-slip fault zone, a generally vertical fracture parallel to the relative motion between the two plates that it separates. Strike-slip fault zones are described as right-lateral if the far side is moving right relative to the near side (for example, the Queen Charlotte zone; **Fig. 1**), left-lateral if it is moving to the left (for example, the North Caribbean zone; Fig. 1). Not all such fault zones are plate-bounding transform faults. Small-scale strike-slip faulting is a common secondary feature of many subduction zones, especially where plate convergence is oblique, and of some spreading centers, especially those with propagating rifts; it also occurs locally deep in plate interiors. The distinguishing characteristic of a transform fault is that both ends extend to a junction with another type of plate boundary. At these junctions the divergent or convergent motion along the other boundaries is transformed into purely lateral slip. *See* EARTH CRUST; PLATE TECTONICS; SUBDUCTION ZONES.

Types. Transform faults are most readily classified by the types of plate boundary intersected at their ends, the variety of lithosphere (oceanic or continental) they separate, and by whether they are isolated or are part of a multifault system. The common oceanic type is the ridge-ridge transform, linking two literally offset axes of a spreading center (for example, Clipperton and Siqueiros transforms; Fig. 1). Also common are transform faults that link the end of a spreading center to a triple junction, the meeting place of three plates and three plate boundaries. For example, the Panama transform links a spreading axis on the mid-oceanic ridge between the diverging Cocos and Nazca plates to a Cocos-Nazca-Caribbean plate triple junction at the continental margin of Central America. *See* LITHOSPHERE; MID-OCEANIC RIDGE.

Other types are long trench-trench transforms at the northern and southern margins of the Caribbean plate, and the combined San Andreas/Gulf of California transform, which separates the North American and Pacific plates for 1500 mi (2400 km) between triple junctions at Cape Mendocino (California) and the mouth of the Gulf of California (Fig. 1). Strike-slip faulting in the Gulf of California (and on the northern Caribbean plate boundary) occurs along several parallel (en echelon) zones linked by short spreading centers, and the overall structure is more properly called a transform fault system; similar fault patterns are found at many ridge-ridge transforms. The San Andreas part of this plate boundary exhibits another type of transform fault system, one with several simultaneously active zones that overlap, rather than replacing each other in stepwise fashion; this pattern may be characteristic of wholly continental transforms.

Geology of oceanic transforms. The structure of ridge-ridge transforms on mid-oceanic ridges varies to some extent with their length, ranging 6–600 mi (10–1000 km), and with the rate of slip of their strike-slip faults 0.8–8 in. (20–200 mm) per year, but the structure depends mainly on the geologic history of changing plate motions. In the absence of such changes, where transform faults separate plates that have maintained the same motion for millions of years, the characteristic structure is a transform valley parallel to the direction of relative plate motion, and thereby transverse to the mid-oceanic ridge. The valley floor is occupied by a strike-slip fault zone, a band of shattered rock only 300–600 ft (100–200 m) wide, that is often marked by a groove or rift in the sea floor. There is some correlation of valley depth and transform length, and most ridge-ridge transforms longer than 60 mi (100 km) have valleys deeper than the 3300–6600-ft-deep (1000–2000-m) axial rift valleys typical of the crests of slow-spreading ridges. Slow-slipping transform valleys are generally deepest at their ends, at their orthogonal intersections with axial rift valleys, whereas the deepest parts of fast-slipping valleys are usually in their midsections, their ends being partly filled with lava that spills over from the intersecting spreading axes. *See* LAVA.

Transform valleys are structural troughs opened by a small component of extension across ridge-ridge transform faults. Where the faults are strictly parallel to relative plate motion, the origin of this extensional stress is probably thermal contraction of the young lithosphere accreted at the intersecting spreading axes. Larger components of extension, creating deeper transform valleys, can result from

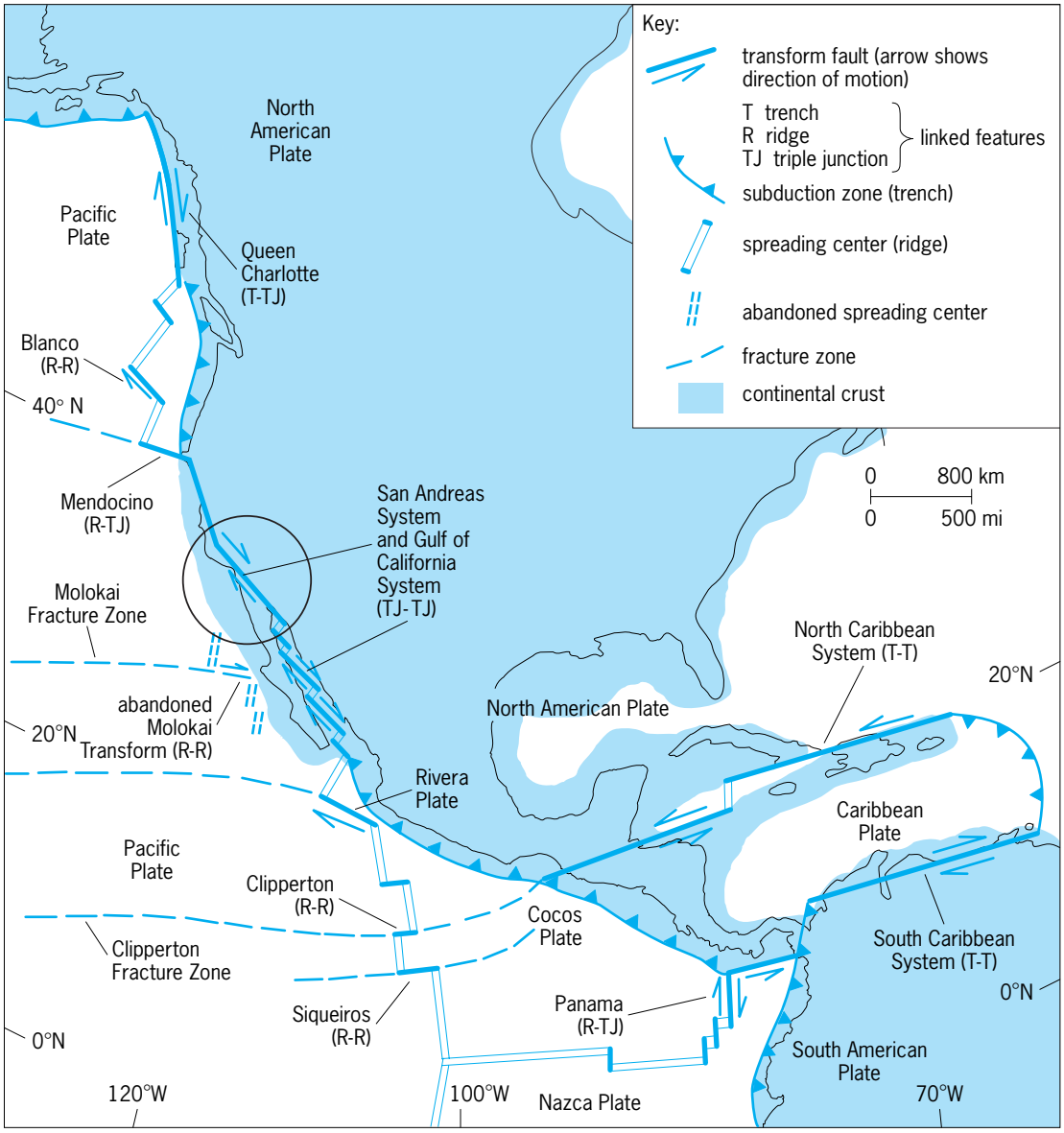


Fig. 1. Various types of transform fault mapped in parts of North and Central America. Area inside circle is shown in detail in Fig. 3.

small angular changes in the direction of plate motion; conversely, opposite changes can add a component of valley-closing compression. The motions of most oceanic plates are changing continuously, albeit slowly, affecting all the transform faults along their boundaries. A small rotation in the direction of plate motion has an opposite effect on the stresses at adjacent left-lateral and right-lateral transforms (**Fig. 2**). Such a motion has affected adjacent transform faults on part of the East Pacific Rise, where relative plate motion (spreading direction) has rotated anticlockwise by about 5° in the past 5 million years. The resulting convergence across the left-lateral Clipperton transform has closed its transform valley and thrust up a median ridge 2000 ft (600 m) high, with the strike-slip fault zone along its crest. The same rotation has caused extension at the right-lateral Siqueiros transform (Fig. 1). Some transforms react to a change of this sort by opening wider and deeper transform valleys, often accompanied by uplift of the valley margins by 0.6–1.8 mi (1–3 km) to form structures known as transverse ridges. In some cases, opening transform faults by adding a component of extension allows seawater to penetrate deep into the lithosphere, chemically altering the upper mantle to a low-density rock (serpentinite) that rises along the fault zone, forming median ridges similar in shape but quite different in origin to those at compressed transform faults. In a few examples, crustal divergence across the transform fault allows magma to leak out, building yet another type of median ridge. The Siqueiros transform shows a more common response to the change in plate motion. Instead of maintaining the same strike-slip fault zone, and adding an extensional component to its lateral motion, it developed a set of new fault zones, each parallel to the new plate motion, and slightly oblique to the overall trend of the transform fault system.

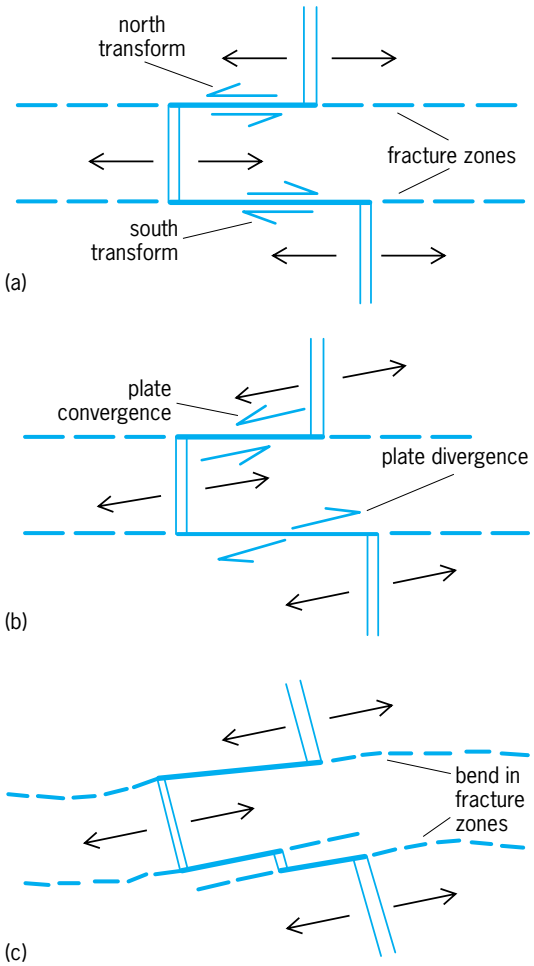


Fig. 2. The effect of changing plate motion on transform faults and their fracture zones. (*a*) East-west plate motion; offset spreading axes with a left-lateral North Transform and a right-lateral South Transform. (*b*) Change in motion; a small rotation of the direction of plate motion adds components of plate convergence to North Transform and divergence to South Transform. (*c*) Adjustment to change; fault is segmented into two fault zones aligned parallel to the new plate motion, which are linked by a short new spreading axis.

Magma does leak out within the fault system, but only at the short new spreading axes which link the parallel fault zones. *See* SERPENTINITE.

The geology of ridge-ridge transform faults is sensitive to the history of recent changes in the direction of plate motion, and the pattern of the fracture zones that they leave on the flanks of the mid-oceanic ridges provides a record of these changes. Fracture zones are bands of rough topography that extend down ridge flanks from the ends of transform faults. Their name is inherited from an early false interpretation that they are belts of strike-slip faulting across the flanks of mid-oceanic ridges. A foundation of the theories of sea-floor spreading and plate tectonics was the recognition that fracture zones are not active fault zones, merely the seams between crust that differs in age becuase it has spread different distances from laterally offset spreading axes. The lateral offsets occur at transform faults, so fracture zones are their inactive continuations. The azimuth of a fracture zone is parallel to the direction of plate motion at the time that the crust on its younger side spread off the risecrest, with fracture zone bends and kinks marking changes in direction. Mapping the trends of fracture zones is a principal method of investigating the past movements of lithospheric plates. *See* MARINE GEOLOGY.

Geology of continental transforms. Transform faults within the continental lithosphere fracture crust that is much thicker and less homogeneous than oceanic crust. Perhaps as a result, the fault zones tend to be less straight, with many local deviations from azimuths parallel to relative plate motion. Bends in the fault zones add components of extension or compression to the dominantly strike-slip motion, resulting in along-strike alternations of collapsed extensional basins and uplifted compressional ridges. A well-known compressional bend is the Big Bend of the San Andreas fault zone north of Los Angeles (**Fig. 3**), where oblique convergence of the Pacific and North American plates is raising the San Bernandino and San Gabriel mountains. Some of the sediment-filled basins formed along continental transform faults are important petroleum reservoirs, and secondary deformation on the margins of the fault zone commonly folds the sediment layers to form trapping structures for oil fields. *See* BASIN; PETROLEUM GEOLOGY.

Most continental transform fault systems have several belts of faulting, with complex spatial patterns of overlapping and splaying fault zones, and complex geologic histories, involving constant shifting of the share of the total interplate displacement among several zones, accompanied by the birth of new fault zones and the abandonment of others. In southern California, for example, motion on the San Andreas transform fault system is now concentrated on three narrow fault zones (Fig. 3) which differ in age and are accompanied by a multitude of less active subparallel zones, some of which may become dominant traces in the near geologic future. Very detailed geologic studies are needed to unravel the histories of continental transform fault systems, which do not leave fracture zone traces like their oceanic counterparts. The total lateral displacement between the two sides of a fault zone, commonly amounting to tens or hundreds of miles, can be estimated by recognizing the two displaced halves of preexisting geologic features that were split and separated by fault motion. On a much shorter time scale, recent displacements can be monitored by offsets in human-made features such as fence lines.

Some of the less active fault zones in the region of the San Andreas and Gulf of California systems are senescent rather than nascent transform faults. Until 6 million years ago, transform faulting was centered west of Baja California and west of the southern California coastline, and the inland shift of the Pacific–North America plate boundary has caused the almost complete cessation of faulting on the offshore San Benito and San Clemente fault zones.

Shearing continental margins. The now-inactive San Benito fault zone was representative of an

important class of transform faults that extend along continental margins, at or near the boundary between oceanic and continental lithosphere. A still-active example is the Queen Charlotte fault zone off the British Columbia coast (Fig. 1). Continental margins shaped by lateral shearing (transform faulting) have very steep continental slopes, but often with steps on the slope called marginal plateaus, crustal blocks that have subsided between the shifting fault zones of a transform fault system. The shifting is commonly away from the oceanic/continental boundary into adjacent weaker continental lithosphere; indeed, the San Andreas fault system can be considered a marginal shear zone that has shifted unusually far inland.

Most shearing margins of western North America were formerly, with an earlier arrangement of a lithospheric plates, convergent (subduction zone) margins. Shearing margins with active transform faults also play a role during the birth of ocean basins by continental rifting. Initial rifting, as in the split of North America from Africa about 200 million years ago, is commonly on laterally displaced fractures that develop into spreading axes linked by transform faults. These transform faults become part of the oceanic/continental boundary once continental separation has proceeded far enough for sea-floor spreading to occur, and eventually become ridge-ridge transforms once a risecrest develops in the new ocean basin. Many of the ridge-ridge transform faults on the Mid-Atlantic Ridge are inherited from fault zones that once formed shearing parts of the continental margin. Shearing margins occur on the boundaries of the very small, young ocean basins that have opened by the splitting of Baja California from mainland Mexico (for example, Guaymas Basin, with the Guaymas transform fault on its northeast side; Fig. 3). *See* CONTINENTAL MARGIN.

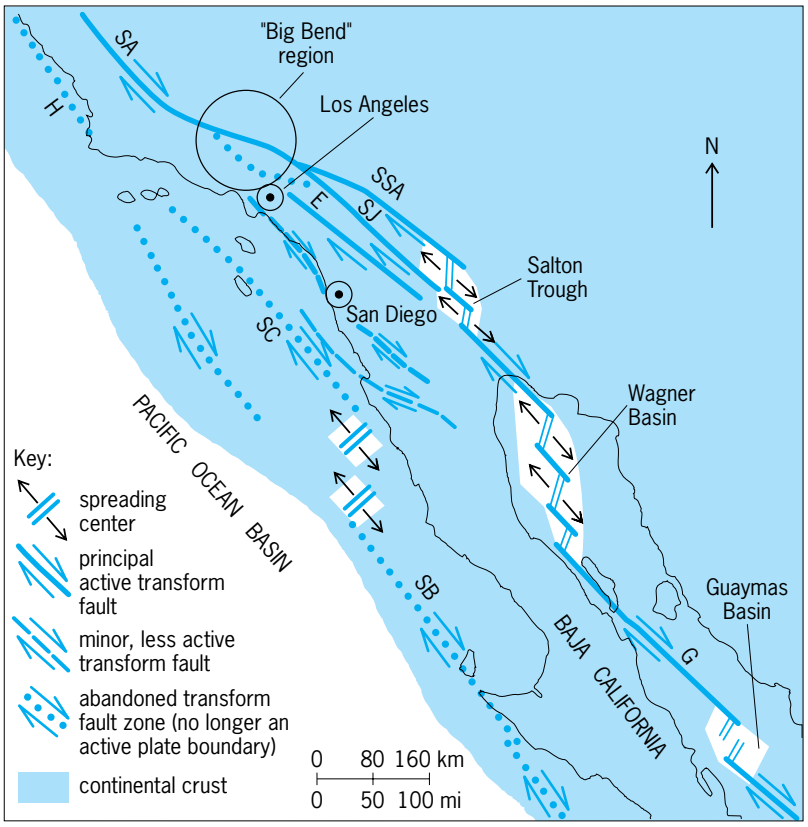


Fig. 3. Map showing parts of the San Andreas and Gulf of California fault systems. Many minor but still earthquake-generating strike-slip fault zones have been omitted. The abandoned offshore fault zones, relics of a time when the plate boundary was closer to the continental margin, still have a low level of residual earthquake activity. The individual fault zones are San Andreas (SA), Hosgri (H), South San Andreas (SSA), San Jacinto (SJ), Elsinore (E), San Clemente (SC), San Benito (SB), and Guaymas (G).

Earthquakes. Along a few strike-slip fault zones, lithospheric plates slide quietly and almost continuously past each other by the process called aseismic creep. Much more often, frictional resistance to the sliding in the brittle crust causes the accumulation of shear stresses that are episodically or periodically relieved by sudden shifts of crustal blocks, creating earthquakes. The largest lateral shifts (slips) of the ground surface along major continental transform faults have been associated with some of the largest earthquakes on record; in 1906 the Pacific plate alongside 270 mi (450 km) of the San Andreas Fault suddenly moved an average of 15 ft (4.5 m) northwest relative to the North American plate on the other side, and the resulting magnitude-8.2 earthquake destroyed much of San Francisco. The average slip in this single event was equivalent to about 150–250 years of Pacific–North American plate motion.

The maximum size of earthquake that a transform fault can generate is limited by the length of the fault, though generally, even in large earthquakes like the one in San Francisco in 1906, a fault does not fail along its entire length. The frequency of earthquakes is controlled by the average speed of relative plate motion across a transform fault plate boundary and, for fault systems with multiple overlapping fault zones, by the share of this motion that is carried by any individual fault. However, many local geologic and tectonic factors intrude to complicate estimates of how frequently a particular transform fault will produce earthquakes of any specified size or destructive power, and it is still more difficult to predict the exact timing of such an event. Extrapolation of the past record is probably the best method of estimating future magnitudes and frequencies, becuase many transform faults do seem to have a characteristic size of large earthquakes, and a consistent ratio of small to large events. In most cases, however, especially for remote oceanic transform faults, the record is too short and too incomplete to be of much practical use. *See* EARTHQUAKE; FAULT AND FAULT STRUCTURES; SEISMOLOGY.

Peter Lonsdale

Bibliography. W. G. Ernst (ed.), *The Geotectonic Development of California*, 1981; P. J. Fox and D. G. Gallo, The geology of North Atlantic transform plate boundaries and their aseismic extensions, *The Geology of North America*, vol. M, pp. 157–172, 1986; P. J. Fox and D. G. Gallo, Transforms of the eastern central Pacific, *The Geology of North America*, vol. N, pp. 111–124, 1989; J. T. Wilson, A new class of faults and their bearing on continental drift, *Nature*, 207:343–347, 1965.

Transformation (bacteria)

The addition of deoxyribonucleic acid (DNA) to living cells, thereby changing their genetic composition and properties. Such transformation was first reported by Fred Griffith in 1928 as occurring in pneumonia bacteria (pneumococci, or *Streptococcus pneumoniae*) growing in animals injected in the laboratory. Other investigators produced the same phenomenon in bacteria of the same species growing in the test tube. The process is looked upon as the transfer of a transforming principle, or substance, from donor bacteria to suitable recipient bacteria. The recipient bacteria are usually closely related to the donor strain. It is realized that the process may occur in natural conditions, for example, in a host animal infected with two parasitic strains, and indeed it might play a part in the rapid evolution of pathogenic bacteria.

There are several species of bacteria in which transformation has been achieved in the laboratory, including pneumococci, meningococci, *Haemophilus influenzae*, and *Bacillus subtilis*. There are several other species in which similar phenomena have been observed, and other strains of the same and other species in which it has not as yet been possible to achieve transformations.

Role of DNA. The most far-reaching clarification attained through laboratory investigation was the announcement by Oswald T. Avery and coworkers in 1944 that the purified agent (or transforming principle) inducing capsule-type transformation of pneumococcus had the properties of a pure DNA. It was later shown that other specific hereditary traits of these and other bacteria could be transformed by a similar process. In this way it is increasingly clear that the structures of deoxyribonucleic acids of different origins bear and transmit the imprint of their specific genetic origins. Supplementing this evidence, there have been signs that cell nuclei in general all contain DNA in constant, characteristic amounts, and that radiations and chemicals which can affect nucleic acids also provoke mutations in living organisms. A confirmation of the role of DNA was the evidence that the active components of certain viruses are the DNAs they carry. *See* ANIMAL VIRUS; DEOXYRIBONUCLEIC ACID (DNA); GENETICS; MUTATION; NUCLEIC ACID.

Nature of transformation. That bacterial transformation is true genetic transmission on a small scale, rather than controlled mutation, is demonstrated by the following characteristics: (1) A specific trait is introduced, coming always from donors bearing the trait. (2) The trait is transferred by determinant, gene-like material far less complex than whole cells or nuclei, and this material, DNA, is known to be present in gene-carrying chromosomes. (3) The trait is inherited by the progeny of the changed bacteria. (4) The progeny produce, when they grow, increased amounts of DNA carrying the specific property. (5) The traits are transferred as units exactly in the patterns in which they appear or in which they are induced by mutation. (6) The DNA transmits the full potentialities of the donor strain, whether these are in an expressed or in a latent state. (7) The traits are often attributable to the presence of a specific gene-determined enzyme protein. (8) Certain groups of determinants may occur "linked" within DNA molecules, just as genes may be linked, and if so, heat denaturation, radiation, or enzyme action will inactivate or separate them just to the extent that they can damage or break apart the DNA molecules. (9) Linked determinants, while transforming a new cell, may become exchanged (recombined) between themselves and their unmarked or unselective alternate forms in such a way that they bring about genetic variation, and in a pattern indicating the existence of larger organized genetic units. *See* BACTERIAL GENETICS; GENE.

Nonbacterial transformation. Through the application of a number of procedures prior to adding the DNA, transformation was extended first to many different bacterial species and then to eukaryotic cells. These procedures included adding high concentrations of calcium or dextran sulfate, or providing a large electric shock (electroporation). Today almost any cell type can be transformed. In some cases, tissues can be injected directly with naked DNA and transformed. However, unlike with bacteria, the naked DNA adds almost anywhere in the genome rather than recombining with its indigenous homolog. However, with special highly selective procedures, homologous recombination can be obtained. By treating embryonic stem cells and adding them to embryos that then go to term, specific and nonspecific transgenic animals can be obtained (for example, mice). *See* GENETIC ENGINEERING.

When the source of the DNA is some entity capable of independent replication, such as a virus or plasmid, the phenomenon is called transfection. If foreign DNA is then inserted into these entities, the result is recombinant DNA that can lead to transduction. *See* MOLECULAR BIOLOGY; TRANSDUCTION (BACTERIA).

Rollin D. Hotchkiss; Norton D. Zinder

Bibliography. O. T. Avery, C. M. MacLeod, and M. McCarty, *J. Exper. Med.*, 79:137–158, 1944.

Transformer

An electrical component used to transfer electric energy from one alternating-current (ac) circuit to another by magnetic coupling. Essentially it consists of two or more multiturn coils of insulated conducting material, so arranged that any magnetic flux linking one coil will link the others also. This configuration creates mutual inductances between the coils. The mutual magnetic field acts to transfer energy from one input coil (or primary winding) to the other coils, which are then referred to as secondary windings. Under steady-state conditions, only one winding can serve as a primary. *See* COUPLED CIRCUITS; INDUCTANCE.

The transformer accomplishes one or more of the following effects between two circuits: (1) an induced voltage of different magnitude, (2) an induced

current of different magnitude, (3) a difference in phase angle, (4) a difference in impedance level, and (5) a difference in voltage insulation level, either between the two circuits or to ground.

In an electric power system, transformers are used to perform a wide range of functions. Pole-type distribution transformers supply relatively small amounts of power to residences. Power transformers are used at generating stations to step up the generated voltage to high levels for transmission. The transmission voltages are then stepped down by transformers at the substations for local distribution. Instrument transformers are used to enable accurate measurents of voltages and currents. In other applications, audio- and video-frequency transformers must function over a broad band of frequencies. Radio-frequency transformers transfer energy in narrow frequency bands from one circuit to another. *See* INSTRUMENT TRANSFORMER.

Transformers are often classified according to the frequency for which they are designed. Power transformers are for power-frequency circuits, audio transformers for audio-frequency circuits, and so forth. Of course, many of the basic principles of operation apply to all.

Power Transformers

Systems for the transmission and distribution of alternating-current electrical energy would be impractical without transformers. Transformers are capable of stepping alternating voltages up or down with very little loss of power in the process. They thus allow engineers to design different parts of a power system to operate at different voltage levels. Electric generators may then be designed for voltages which make possible the most economical use of materials. These generators are connected through transformers to transmission lines operating at the much higher voltages required for high transmission efficiency. Transformers again step the voltage down to levels which permit the energy to be used safely by the ultimate consumer. Large ac generators, called alternators, are designed to operate at 20 to 25 kV. Power transmission voltages in the United States range from 115 to 750 kV, and utilization voltages range from 115/230 V for home use to 13,200 V in large industrial equipment. *See* ALTERNATING-CURRENT GENERATOR; ELECTRIC DISTRIBUTION SYSTEMS; ELECTRIC POWER SYSTEMS; TRANSMISSION LINES.

Power transformers, as a class, may be defined as those designed to operate at power-system frequencies: 60 Hz in the United States and Canada, and 50 Hz in much of the rest of the world. The largest power transformers connect generators to the power grid. Since a generator, together with its driving turbine and prime energy source, is called a generating unit, such transformers are called unit transformers. The classification "distribution transformers" refers to those supplying power to the ultimate consumers. They are designed for lower power and output-voltage ratings than the other transformers in the system.

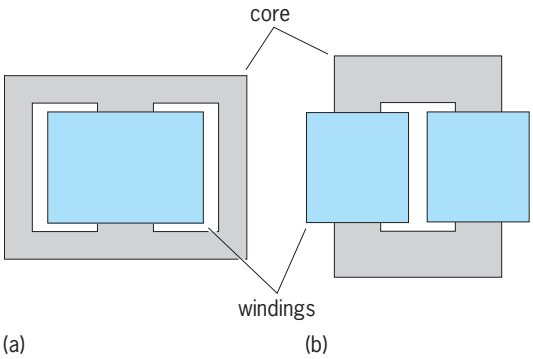


Fig. 1. Location of windings in single-phase cores. (*a*) Shell form. (*b*) Core form.

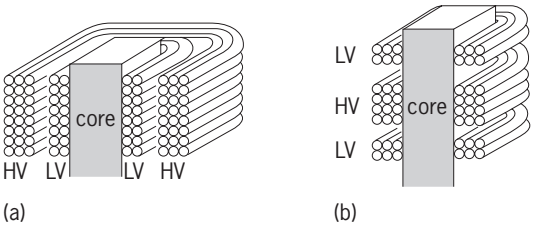


Fig. 2. Winding arrangements. (*a*) Concentric. (*b*) Interleaved.

Configuration. Any magnetic flux which arises from a current in one winding of a transformer but does not link (is not mutual to) the other windings is called the leakage flux of that winding. General-purpose power transformers are designed to maximize the mutual component of the flux of each winding by providing an iron-alloy core which links all windings. This core acts as a high-permeance path for the mutual magnetic flux. Practically speaking, all of the mutual flux flows in this core.

Typical configurations for single-phase transformers are shown in **Fig. 1**. The arrangement in Fig. 1*a* is called a shell-form transformer, while that in Fig. 1*b* is called a core-form transformer. Each of the rectangles labeled "windings" in this figure represents at least two coils. The coils may be concentric, or interleaved (**Fig. 2**). In the shell form, all of the windings are on the center leg. In the core form, half of the turns of the primary and half of those of the secondary are on each leg. The two halves of a given winding may be connected in series or in parallel. In the concentric arrangement, the low-voltage winding is closer to the core, so that less insulating material is required in the construction of the transformer.

Principles of operation. Transformers operate on the basis of two fundamental physical laws: Faraday's voltage law and Ampère's law. Faraday's law states that the voltage induced in a winding by a magnetic flux linking that winding is proportional to the number of turns and the time rate of change of the flux; that is, Eq. (1) holds, where e_i is the voltage induced

$$e_i = N_i \frac{d\phi_i}{dt} \quad \text{volts} \tag{1}$$

in a coil of N_i turns which is threaded by a flux of ϕ_i

webers changing at a rate of $d\phi_i/dt$ webers per second. The ratio of the voltages induced in two windings of a transformer by the core flux is, then, given by Eq. (2). In other words, the voltages induced in the

$$\frac{e_1}{e_2} = \frac{N_1}{N_2}\frac{d\phi_{\text{core}}/dt}{d\phi_{\text{core}}/dt} = \frac{N_1}{N_2} \tag{2}$$

windings are proportional to the numbers of turns in the windings. This is the basic law of the transformer. A high-voltage winding will have many turns, and a low-voltage winding only a few. The N_1/N_2 ratio is usually called the turns ratio or transformation ratio and is designated by the symbol a, so that Eq. (3)

$$\frac{e_1}{e_2} = a \tag{3}$$

holds. *See* FARADAY'S LAW OF INDUCTION.

Since the flux must change to induce a voltage, steady-state voltages can be obtained only by a cyclically varying flux. This means that alternating voltages and fluxes are required for normal transformer operation, and that is the fundamental reason for ac operation of power systems. Devices operated on ac have lower losses when the voltages and fluxes are sinusoidal in form, and sinusoidal fluxes and terminal voltages will be assumed in this discussion. *See* ALTERNATING-CURRENT CIRCUIT THEORY.

Core excitation. **Figure 3** shows the elements of a two-winding, shell-form transformer. The center leg of the core carries the full mutual flux, and each of the outside legs carries half of it. Thus the cross-sectional area of each outside leg is half of that of the center leg.

The core is laminated; that is, it is made up of a stack of iron-alloy sheets. The laminations are insulated from each other by iron oxide or by some other coating. If this were not the case, the core would act like a closed, single-turn winding, and the voltage induced in it by the core flux would cause large circulating currents in the core, called eddy currents. These currents would cause excessive core heating, and result in a severe reduction in the efficiency of the transformer. There is some eddy-current loss even in laminated cores. *See* EDDY CURRENT.

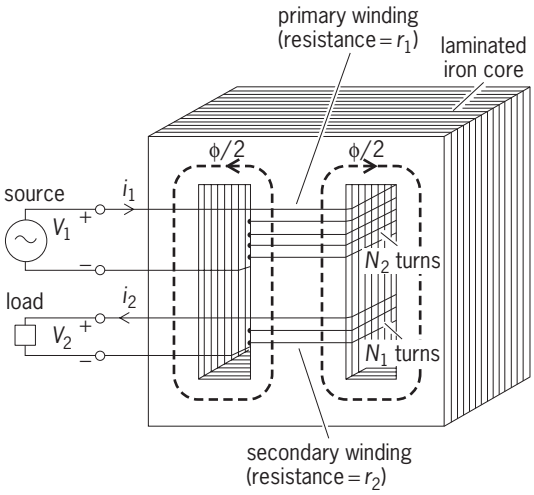


Fig. 3. Elements of a transformer.

The thickness chosen for the laminations is determined by the operating frequency. For 60 Hz, the thickness is usually 0.014 in. (0.36 mm). Lamination makes the core mechanically weak, so arrangements must be provided for its support. Cores for distribution transformers are often wound from iron-alloy strip, and the possibility of using magnetic glass in power transformer cores has been investigated.

The constant reorientation of the magnetic domains in the core iron as the flux alternates in direction results in hysteresis loss. The sum of the hysteresis and eddy-current losses is called the core loss. *See* CORE LOSS.

The concept of exciting current may be defined by considering the situation in which the secondary winding in Fig. 3 is open (the load is disconnected) and a sinusoidal voltage is applied to the terminals of the primary winding, given by Eq. (4). In this equation V_1 is the effective or root-mean-square (rms) voltage, the peak voltage is $2V_1$, f is the frequency in hertz, and $2\pi f$ is the angular frequency ω in radians per second. An alternating flux is set up in the core and a voltage e_1 is induced in the N_1 turns in the primary which is almost equal to v_1 and opposed to it. The small difference between v_1 and e_1 is due to the leakage flux and to the winding resistance. This difference is enough to allow a small current to flow, sufficient to set up the core flux and to supply the hysteresis and eddy-current losses of the core. This small current (1–5% of rated primary current) is called the exciting current, i_{ex}. Its root-mean-square phasor I_{ex} may be considered to be made up of two components, as in Eq. (5), where I_ϕ is the magnetizing component, which lags the applied voltage by 90°; and I_{cl} is the core-loss component, which is in phase with the voltage. Even though the applied voltage and core flux are sinusoidal, the nonlinearity of the iron causes the exciting current to be nonsinusoidal; it contains a large third-harmonic component. (The concept of the phasor I_{ex} is actually an approximation; it represents a sinusoid having the same root-mean-square value as the actual nonsinusoidal waveform.)

$$v_1 = \sqrt{2}\, V_1 \cos 2\pi f t \qquad \text{volts} \tag{4}$$

$$I_{\text{ex}} = I_\phi + I_{cl} \qquad \text{amperes} \tag{5}$$

The applied voltage v_1 and the voltage induced by the core flux are so nearly equal that Eq. (6) follows from Eqs. (1) and (4). Integrating Eq. (6) gives Eq. (7), where ϕ_K, the constant of integration, is a transient flux which decays after a few cycles, due to effects not included in Eq. (6). This transient is

$$\sqrt{2}\, V_1 \cos 2\pi f t = N_1 \frac{d\phi_{\text{core}}}{dt} \tag{6}$$

$$\phi_{\text{core}} = \frac{V_1}{N_1}\frac{\sqrt{2}}{2\pi f} \sin\, 2\pi f t + \phi_K \qquad \text{webers} \tag{7}$$

responsible for the phenomenon of current inrush. The peak value of the core flux is, in the steady state, given by Eq. (8), and is seen to be proportional to

$$\phi_{\max} = \frac{V_1}{N_1 f}\frac{\sqrt{2}}{2\pi} \tag{8}$$

the volts per turn of the primary winding. Since the core flux density (teslas, or webers/m^2) is limited by saturation, the designer's choice of volts per turn of the windings will determine the cross-sectional area of the core, or vice versa.

Transformation of the secondary load. When an electrical load is connected to the terminals of the secondary winding, current flows through that winding and the load as a result of the voltage e_2 induced in the secondary by the alternating core flux. In Fig. 3, source current flows into the positive terminal of the primary, but out of the positive terminal of the secondary. Application of the right-hand rule to the two windings shows that the load current in the secondary results in a magnetomotive-force (MMF) field, which is opposed to the magnetomotive-force field of the primary. Now, Ampère's law in circuital form states that the magnetomotive force applied to a closed path is simply equal to the net current enclosed by that path. In Fig. 3, Ampère's law may be applied to either of the two $\phi/2$ paths. In either case, the net current enclosed is $N_1 i_1 - N_2 i_2$, and the magnetomotive force applied to the core around either path is given by Eq. (9).

$$F_{\text{core}} = N_1 i_1 - N_2 i_2 \qquad \text{amperes} \tag{9}$$

See AMPÈRE'S LAW.

Now the core flux is also given by the core permeance P times the applied magnetomotive force, Eq. (10). Combining Eqs. (9) and (10) gives Eq. (11), and solving Eq. (11) for i_1 gives Eq. (12). The term

$$\frac{\phi_{\text{core}}}{2} = \frac{P}{2}F_{\text{core}} \qquad \text{webers} \tag{10}$$

$$\frac{\phi}{P} = N_1 i_1 - N_2 i_2 \qquad \text{amperes} \tag{11}$$

$$i_1 = \frac{N_2}{N_1} i_2 + \frac{\phi_{\text{core}}}{PN_1} \qquad \text{amperes} \tag{12}$$

ϕ_{core}/PN_1 is the instantaneous value of the exciting current of Eq. (5). The large permeance of the core makes this term very small. Equation (12) may be rewritten as Eq. (13). Thus, with a load on the secondary, the primary current is equal to the secondary current transformed by the inverse of the turns ratio, plus the current required to excite the core. *See* PERMEANCE.

$$i_1 = \frac{i_2}{a} + i_{\text{ex}} \qquad \text{amperes} \tag{13}$$

Ideal transformer. An ideal transformer would require no exciting current. In other words, its core would have infinite permeance, and the primary current would be given by Eq. (14). The windings of an ideal transformer would have zero resistance, and would be wound so as to have no leakage flux (a physical impossibility). Then the induced voltages would exactly equal the terminal voltages, and Eq. (15) follows from Eqs. (3) and (14). Thus in an ideal transformer, the instantaneous input apparent power equals the instantaneous output apparent power. There are no internal losses, and the ideal transformer has an efficiency of 100%.

$$i_1 = \frac{i_2}{a} \qquad \text{amperes (ideal)} \tag{14}$$

$$v_1 i_1 = v_2 i_2 \qquad \text{volt-amperes} \tag{15}$$

In terms of root-mean-square phasors, the voltages and currents are related by Eqs. (16). The complex load impedance, in ohms, is given by Eq. (17), but the impedance seen by the generator connected to the primary is given by Eq. (18). Thus impedances are transformed by the square of the turns ratio. Similarly, if a load were applied to the number 1 winding, the impedance appearing at the number 2 terminals would be given by Eq. (19). Most power transformers are so nearly ideal that, for many engineering purposes, Eqs. (16), (18), and (19) may be used to obtain satisfactory numerical values with real transformers. *See* ELECTRICAL IMPEDANCE.

$$V_1 = aV_2 \qquad I_1 = \frac{I_2}{a} \tag{16}$$

$$Z_{L2} = \frac{V_2}{I_2} \tag{17}$$

$$Z_{\text{in1}} = \frac{V_1}{I_1} = \frac{aV_2}{I_2/a} = a^2 Z_{L2} \tag{18}$$

$$Z_{\text{in2}} = \frac{Z_{L1}}{a^2} \tag{19}$$

Circuit model. When the detailed performance of a transformer must be accounted for in a power system calculation, a circuit which models the performance of the actual transformer is included in the system circuit (**Fig. 4***a*). To simplify the problem solution, the primary and secondary series impedances are nearly always combined into one equivalent impedance. This is accomplished through the impedance-transforming property of the ideal transformer. If the secondary winding resistance and leakage reactance are transferred to the primary by the factor a^2, and are then added to the primary series impedance, the result is called the equivalent impedance referred to the primary, given in Eq. (20)

$$\begin{aligned} Z_{\text{eq1}} &= R_{\text{eq1}} + jX_{\text{eq1}} \\ &= r_1 + a^2 r_2 + j(x_1 + a^2 x_2) \end{aligned} \tag{20}$$

and diagrammed in Fig. 4*b*, where r_1 and r_2 are the resistances of the primary and secondary windings, and x_1 and x_2 are the reactances due to leakage fluxes. When the primary impedance is transferred and combined with the secondary impedance, the equivalent impedance referred to the secondary

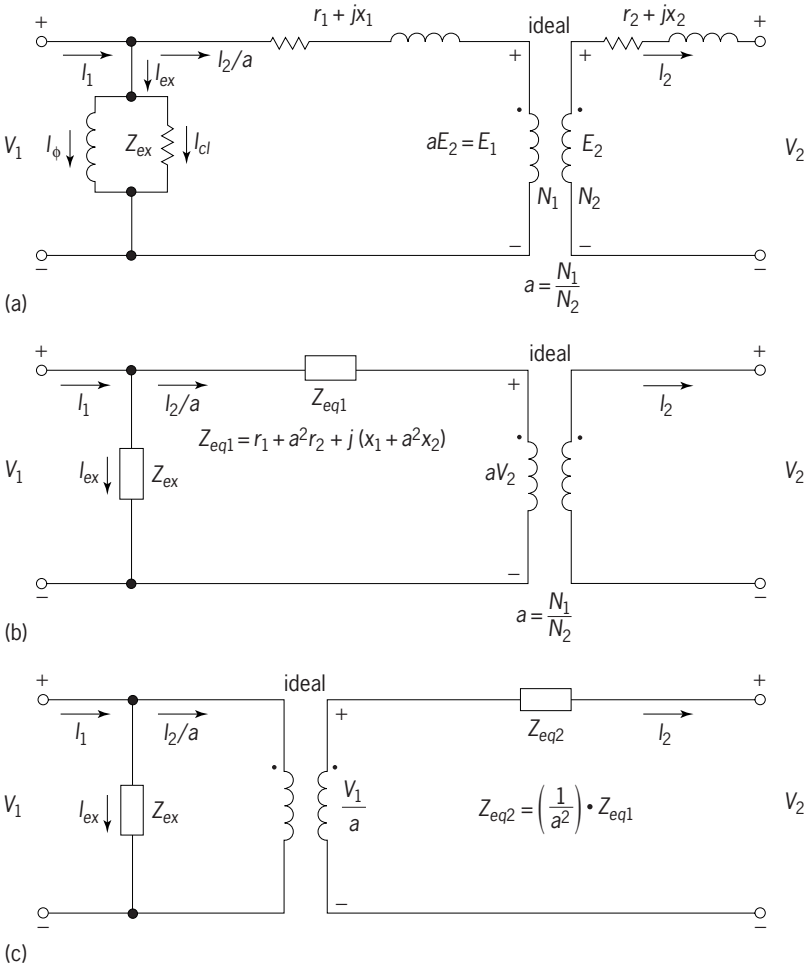


Fig. 4. Circuit models for performance calculations. (*a*) Transformer model showing separate primary and secondary impedances. r_1 and r_2 are resistances of primary and secondary windings, and x_1 and x_2 are reactances due to their leakage fluxes. (*b*) Secondary impedances transformed by ideal transformer and combined with primary impedance to form Z_{eq1}. (*c*) Transformer model in terms of the equivalent impedance referred to the secondary, Z_{eq2}.

results, as given in Eq. (21) and diagrammed in

$$Z_{eq2} = R_{eq2} + jX_{eq2} = \frac{r_1}{a^2} + r_2 + j\left(\frac{x_1}{a^2} + x_2\right)$$

$$= \frac{Z_{eq1}}{a^2} \tag{21}$$

Fig. 4*c*. The operator *j* in these expressions indicates that the voltages across the reactive elements of these impedances leads the voltages across the resistive elements by 90°. All voltages and currents in Fig. 4 are to be treated as ac phasor quantities. *See* REACTANCE.

The model makes it easy to calculate the primary voltage required to have rated secondary voltage at a given volt-ampere load and power factor. If the phase of the secondary voltage is set equal to 0°, as in Eq. (22), and if the power factor is cos θ, then the secondary current is given by Eq. (23), and from

$$V_2 = V_{2\text{rated}} \angle 0^\circ \tag{22}$$

$$I_2 = \frac{kVA \cdot 1000}{|V_{2\text{rated}}|} \angle\theta = |I_2| \angle\theta \tag{23}$$

Fig. 4*c*, the primary voltage is given by Eq. (24). The

$$V_1 = a[V_2 + |I_2|(\cos\theta \pm J \sin\theta) \cdot Z_{eq2}] \tag{24}$$

sign on the $j \sin\theta$ term is positive for leading and negative for lagging power factor, and Z_{eq2} is a complex number.

The model allows easy comparison between realistic and ideal transformers, as follows:

Realistic	*Ideal*
$V_1 = a(V_2 + I_2 Z_{eq2})$	$V_1 = aV_2$
$I_1 = \frac{I_2}{a} + I_{ex}$	$I_1 = \frac{I_2}{a}$

The ratio of rated primary to secondary voltages shown on the transformer nameplate gives the actual turns ratio. This means that, if the secondary voltage is at rated value, the primary voltage will differ a little from its rated value.

Per unit impedance. The equivalent series impedance of a transformer is nearly always expressed as a fraction of the load impedance which would draw rated current at the rated voltage. This full-load impedance for a given winding is called the base impedance of that winding, as in Eq. (25). The

$$Z_{base2} = \frac{V_{2\text{rated}}}{I_{2\text{rated}}} \tag{25}$$

"per unit" impedance of a transformer is given in percent value by Eq. (26). The *R*% is considerably

$$Z\% = \frac{Z_{eq2}}{Z_{base2}} \cdot 100\%$$

$$= R\% + jX\% \tag{26}$$

smaller than *X*%, so usually only the *X*% is shown on the transformer nameplate. Typical values of *R*% range from 0.3% in very large transformers to 1.0% in distribution transformers. It can be shown that *R*% is equal to the internal ohmic (I^2R) loss of the transformer at full load, expressed as a percent of the rated volt-amperes of the device.

The value of *X*% depends on the amount of leakage flux. High-voltage windings require more insulating material, and this results in more space between windings for leakage fluxes. Consequently a transformer with a very high-voltage winding will have a large value of *X*%. Typical values range from 1.5% in distribution transformers to 15% in transformers with, say, a 345-kV winding.

Transforming three-phase power. Three-phase power may be transformed from one voltage level to another by sets of three single-phase transformers or by three-phase transformers. Three-phase transformers have three primary/secondary sets of windings on a single core. The arrangements for core and shell-form three-phase transformers are shown in **Fig. 5**. **Figure 6** shows a core-form, three-phase transformer removed from its tank.

When the core-form, three-phase transformer is operating with balanced voltages, the alternating fluxes of the three phases are equal in magnitude and 120° out of phase with each other. They thus add to

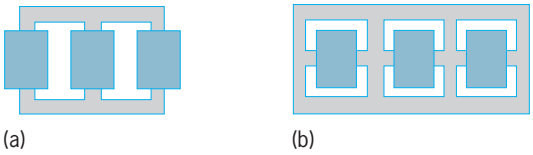


Fig. 5. Typical three-phase cores showing location of windings. (*a*) Core form. (*b*) Shell form.

zero at the center junctions of the core. When it is anticipated that the transformer will operate under unbalanced conditions, additional core legs are added at one or both ends of the core, to form "four-limbed" or "five-limbed" cores. These added legs carry the net flux resulting from the unbalance, which could otherwise flow in the steel transformer tank and cause overheating.

Three-phase transformers may be connected with their primary windings in wye or delta and their secondary windings in wye or delta, providing a great deal of flexibility. A wye-delta connection is illustrated in **Fig. 7**, having wye-connected windings with high-voltage line terminals a_1, b_1, and c_1, and delta-connected windings with low-voltage terminals a_2, b_2, and c_2. In wye-delta or delta-wye, there is always a 30° phase shift between the primary and secondary line voltages. It is standard practice in the United States to connect the windings in such a way that the voltages on low-voltage side (such as V_{a2b2}, the phasor voltage drop from line a_2 to line b_2 in Fig. 7) lag those on the high-voltage side (such as V_{a1b1}, the drop from line a_1 to line b_1) by 30°. There is some advantage in connecting the high-voltage windings in wye, because the neutral (with terminal N in Fig. 7) may be grounded, making the insulation requirements less severe. (In Fig. 7, V_{a1N}, the line a_1 to neutral phasor voltage on the high-voltage side, is the primary voltage of the transformer element whose secondary voltage is V_{a2b2}; thus V_{a2b2} is in phase with V_{a1N}.) It is desirable to have one set of windings in delta, because a delta tends to maintain balanced phase voltages and serves as a solution to certain third-harmonic problems associated with the wye connection. When a wye-wye connection is essential, a set of delta-connected "tertiary" windings is often provided, whose terminals may or may not be made accessible outside the transformer. Three-phase power transformation may also be accomplished by means of sets of two transformers, connected in T, open-delta, or open-wye-open-delta. These arrangements are usually used as distribution transformers, involving relatively small ratings, say, up to 500 kVA.

Efficiency. Efficiency is defined by Eq. (27). The

$$\eta = \frac{\text{output power}}{\text{input power}} \tag{27}$$

output power is equal to the input power, less the internal losses of the transformer. These losses include the ohmic (I^2R) loss in the windings, called copper loss, given by Eq. (28), and the core loss, called the

$$\begin{aligned}\text{Copper loss} &= |I_2|^2 R_{eq2} \text{ (single-phase)} \\ &= 3\,|I_2|^2 R_{eq2} \text{ (three-phase) watts}\end{aligned} \tag{28}$$

no load loss. The input power is thus the sum of the output power and the copper and core losses. Typical efficiency for a 20,000-kVA power transformer at full load is 99.4%, while that of a 5-kVA transformer is 94%.

Cooling. Transformer losses in the windings and core generate heat, which must be removed to prevent deterioration of the insulation and the magnetic properties of the core. Most power transformers are contained in a tank of oil. The oil is especially

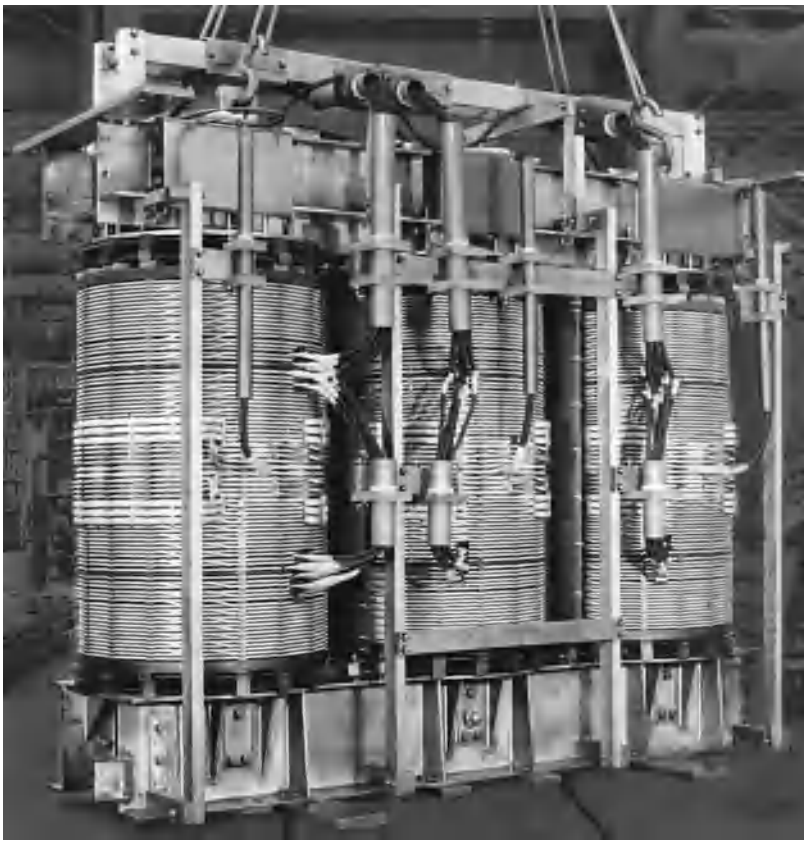


Fig. 6. Three-phase core and coils, rated at 50,000 kVA, 115,000 V.

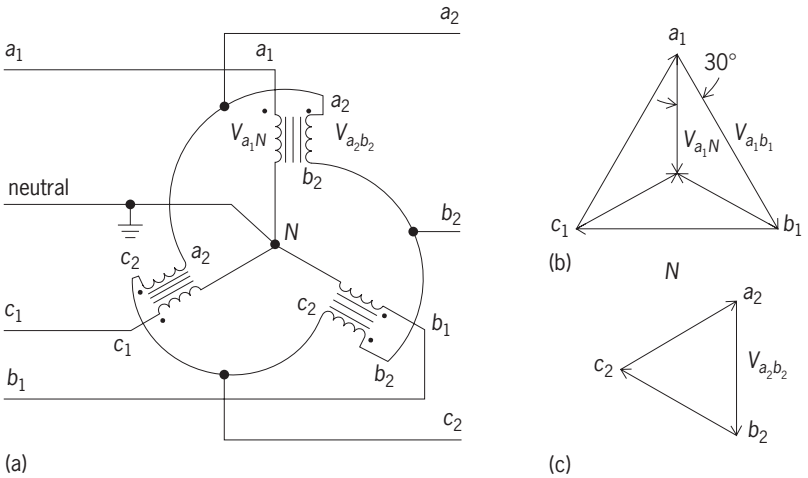


Fig. 7. Wye-delta connection of a three-phase transformer. (*a*) Wiring diagram. (*b*) Voltage-phasor diagram for the high-voltage (wye) side. (*c*) Voltage-phasor diagram for the low-voltage (delta) side.

formulated to provide good electrical insulation, and also serves to carry heat away from the core and windings by convection. In smaller transformers, the hot oil is cooled by flowing along the inside of the tank wall. Larger transformers have ribbed walls or are provided with fins or radiators to cool the oil. Very large transformers have pumps to circulate the oil through external radiators, which may be provided with fans. Transformers which are designed to operate in air are called "dry-type" transformers.

Bushings. Leads from the transformer windings are brought out of the tank through bushings inserted in openings in the tank wall or lid. When high voltages are involved, these bushings may be quite complex, consisting of concentric capacitors enclosed in a ceramic case filled with oil. The capacitors form a voltage divider which distributes the electric field between the transformer lead and the case in such a way as to minimize the probability of breakdown of the insulation. *See* CAPACITOR.

Figure 8 shows a cutaway 20-kVA, single-phase, shell-form, distribution transformer, designed for mounting on a utility pole. The tank is normally

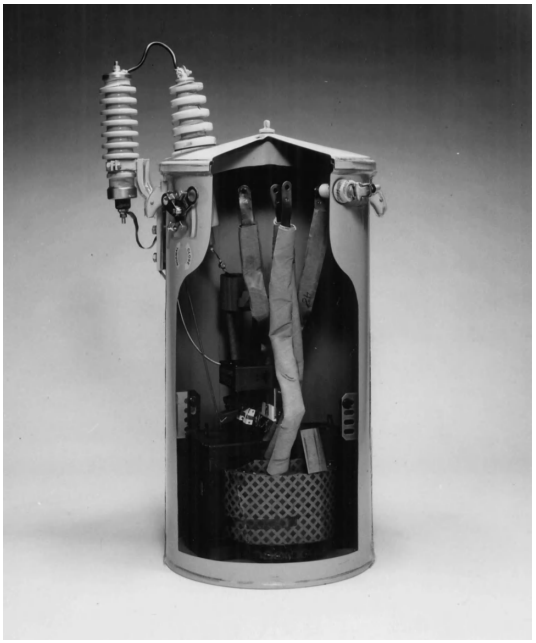


Fig. 8. Cutaway of single-phase, pole-type, completely self-protected (CSP) distribution transformer. (*Westinghouse Electric Corp.*)

three-quarters filled with insulating, cooling oil. The transformer has one high-voltage bushing; the other terminal of the high-voltage winding is connected to the steel tank. A spark gap connects the high-voltage terminal to a lightning arrester, which protects the transformer insulation from overvoltage. Immediately above the core, inside the tank, is an overcurrent circuit breaker. Such a transformer is said to be completely self-protected (CSP). *See* CIRCUIT BREAKER; SURGE ARRESTER.

George McPherson, Jr.

Audio- and Radio-Frequency Transformers

A second important application for transformers is in signal processing. Transformers are used to deliver a waveform of some kind from one circuit to a second (for example, the output of one transistor or integrated circuit to the input of another). Energy is delivered to the second circuit, and the manner of utilization of the energy is extremely important.

With signal processing, it is vital that the waveform of the signal be controlled. For this to be possible, the signal energy must be delivered to a load that is usually linear; often it should also be resistive. If the transformer is unloaded, the signal energy is used to induce magnetization in the transformer core, leading to harmonic and intermodulation distortion due to magnetic hysteresis, and frequency distortion due to the variation of the input reactance of the transformer primary. The resistive load must be placed on the secondary output to minimize the adverse effects of the nonlinear magnetizing current. In addition, the load impedance must be chosen so that it swamps out the magnetizing reactance effect and introduces the desired load impedance characteristics. It can be chosen to reduce the effects of variation of the impedance as a function of signal voltage developed on the secondary, since these variations can also introduce both harmonic and intermodulation distortion. As discussed below, the load may also be selected in a way that will limit the magnitude of the phase distortion introduced by a following amplifier and will assure stable operation. (This distortion often is called departure from the minimum-phase condition, and is a result of parasitic feedback paths in a circuit.) *See* DISTORTION (ELECTRONIC CIRCUITS).

Audio or video (broad-frequency-band) transformers are used to transfer complex signals containing energy at a large number of frequencies from one circuit to another. Radio-frequency (rf) and intermediate-frequency (i-f) transformers are used to transfer energy in narrow frequency bands from one circuit to another. Audio and video transformers are required to respond uniformly to signal voltages over a frequency range three to five or more decades wide (for example, from 10 to 100,000 Hz), and consequently must be designed so that very nearly all of the magnetic flux threading through one coil also passes through the other. These units are designed to have a coupling coefficient k, given in Eq. (29),

$$k = \frac{M}{\sqrt{L_1 L_2}} \tag{29}$$

nearly equal to 1. Here L_1 and L_2 are the primary and secondary inductances, respectively, and M is the mutual inductance (**Fig. 9**). The high coupling coefficient is obtained by the use of interleaved windings and a high-permeability iron core, which concentrates the flux. Typical values of k for highest-quality video transformers may be as high as 0.9998; that for power transformers need not be greater than 0.98.

All radio-frequency and intermediate-frequency

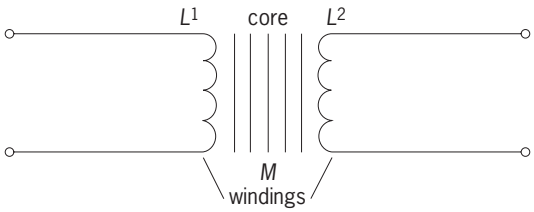


Fig. 9. Schematic of a transformer with symbols.

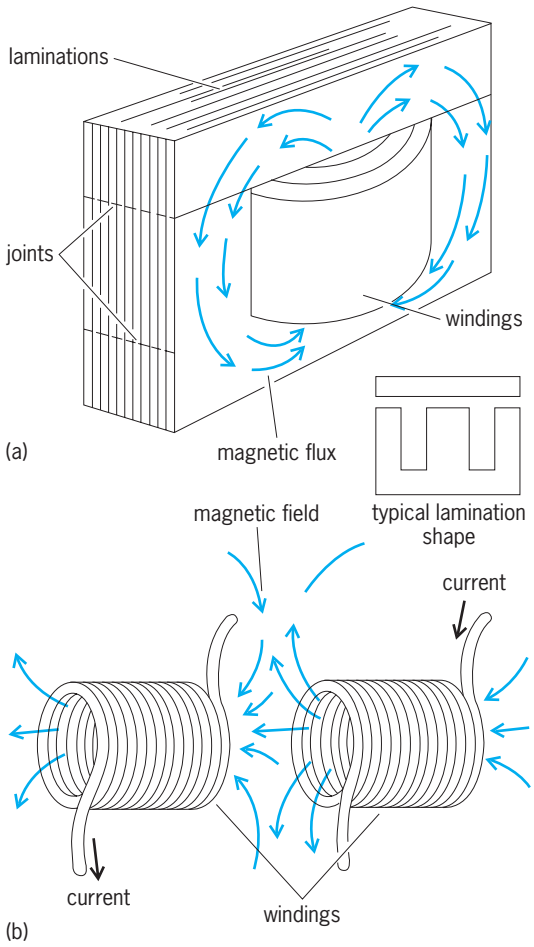


Fig. 10. Audio- and radio-frequency transformers. (*a*) Iron-core audio transformer. (*b*) Air-core radio-frequency transformer.

transformers are built from individual inductors whose magnetic fields are loosely coupled together, $k < 0.30$; one or more inductors are resonated with a capacitor to make efficient energy transfer possible near the resonant frequency. The structure of audio- and radio-frequency transformers is shown in **Fig. 10**. *See* RESONANCE (ALTERNATING-CURRENT CIRCUITS).

Audio and video transformers. Audio and video transformers have two resonances (caused by existing stray and circuit capacitances) just as many tuned transformers do. One resonance point is near the low-signal-frequency limit; the other is near the high limit. As the coefficient of coupling in a transformer is reduced appreciably below unity by removal of core material and separation of the windings, tuning capacitors are added to provide efficient transfer of energy. The two resonant frequencies combine to one when the coupling is reduced to the value known as critical coupling, then stay relatively fixed as the coupling is further reduced. It is possible for a single-tuned coupled circuit to be overcoupled, leading to a broad-tuned coupled circuit. This factor necessitates resistive loading on audio and video transformers. Overcoupling is also observed in the input coupled circuits commonly used in television receivers, and causes the susceptibility to spurious signals often noted in such receivers.

All transformers are devices for transferring energy from one circuit to another. The energy transferred is absorbed either in the circuits themselves or in an external load circuit. For this reason, proper termination is essential for achieving optimum behavior in circuits containing transformers.

Audio and video transformers have a minimum operating frequency at which the open-circuit reactance of the primary is approximately twice its effective loaded impedance. As with wide-band resistance-capacitance (*RC*) amplifiers, gain may be traded for bandwidth with transformer-coupled amplifiers. The reduction of the terminating resistance across the secondary of the transformer reduces the minimum operating frequency $f_{\min}$ and, in the presence of output capacitance, raises the maximum frequency $f_{\max}$. The approximate values of the minimum and maximum frequencies and the resonant frequency f_r are given by Eqs. (30), where

$$f_{\min} = \frac{R_c(N_1/N_2)^2}{\pi L_1}$$
$$f_{\max} = \frac{1}{2\pi R_c C_2} \tag{30}$$
$$f_r = \frac{1}{2\pi\sqrt{L_{22}C_2}}$$

$L_{22} = L_2 - (M^2/L_1) = L_2\,(1 - k^2)$. This L_{22} is the secondary inductance with the primary short-circuited; C_2 is the output capacitance, both external and internal, on the transformer; and R_c is the load resistance (**Fig. 11**). The resonant frequency f_r should be larger than $f_{\max}$ for best performance. *See* AMPLIFIER.

A transformer used to activate terminating circuitry is called an output transformer; one to activate an input circuit is an input transformer; and others are called interstage transformers.

Distortion. The distortion introduced into the amplified signal by a transformer is caused primarily by its hysteresis loss. This loss effect may be minimized by proper loading on the secondary. The load component of current then is large compared to the magnetizing current. In addition, a resistive load keeps

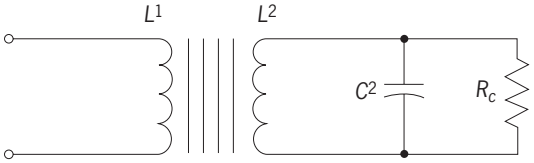


Fig. 11. Circuit of loaded transformer.

the amplification uniform as a function of frequency, and keeps the phase distortion to a minimum.

The magnetic core in an audio or video transformer is subject to two kinds of saturation, that due to applied direct current in the windings, and that due to excessively large signal currents. The direct current in the windings may make the hysteresis loop of the iron core nonsymmetrical, necessitating the use of a larger core having a built-in air gap. Both large signal amplitudes and low frequencies can cause signal saturation to occur in the core. *See* DISTORTION (ELECTRONIC CIRCUITS).

The development of ultracompact radios and other consumer electronics has caused critical problems for component designers. The problems are particularly difficult for units that employ bipolar transistors. The total volume of iron and copper severely influences the impedance and the coupling coefficient as well as the signal amplitude and the saturation characteristics of a transformer. The overall result is limitation of signal amplitude, frequency response, and power-handling capability.

RF and i-f transformers. These use two or more inductors, loosely coupled together, to limit the band of operating frequencies. Efficient transfer of energy is obtained by resonating one or more of the inductors. By using higher than critical coupling, a wider bandwidth than that from the individual tuned circuits is obtained, while the attenuation of side frequencies is as rapid as with the individual circuits isolated from one another.

The tuning of the primary, the secondary, or both may be accomplished either by the variation of the tuning capacitor or by an adjustable magnetic or conducting slug that varies the inductance of the inductor (**Fig. 12**).

Impedance. The operating impedance of a tuned circuit of a radio-frequency transformer is a function of its Q and its tuning capacitance. (Its Q measures the rate of dissipation of stored energy in a tuned circuit as it decays from a starting pulse.) In general, high-power circuits require a high capacitance for energy storage, and therefore have low values of impedance. In any application, the impedance level must be kept sufficiently small to prevent instability and oscillation. *See* Q (ELECTRICITY).

Transformer-coupled amplifiers. Control of voltage gain of transformer-coupled amplifiers is a crucial requirement for effective, stable, radio-frequency and intermediate-frequency amplifiers, both low-power and high-power types. Voltage gain is a function of tuned impedance and the transconductance of the active device, whether tube or transistor.

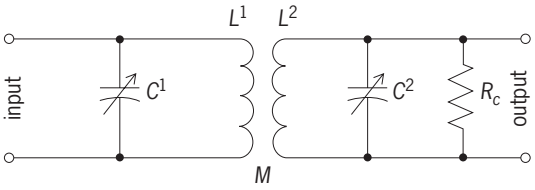


Fig. 12. Tuned radio-frequency transformer.

The tuned impedance Z_t of a tuned circuit depends on both the operating Q of the tuned circuit and its inductance-to-capacitance ratio. (Its operating frequency is a function of the product of inductance L and capacitance C.) Impedance of a capacitor is an inverse function of capacitance and frequency, according to Eq. (31).

$$Z_t \text{ (tuned)} = \frac{L}{CR} = \text{ function of } \frac{L}{C^{1/2}} \qquad (31)$$

The voltage gain takes the form of the product of the transconductance and the tuned impedance, $g_m Z_t$. The transconductance can be expressed in terms of Eq. (32), where kappa κ is the

$$g_m = \kappa \left(\frac{q}{kT}\right) I_c \qquad (32)$$

transconductance-per-unit-current efficiency, q is the electron charge, k is Boltzmann's constant, T is absolute temperature, and I_c is the instantaneous value of the output current for the device. The kappa for a bipolar transistor is approximately unity; for field-effect transistors and electron tubes, it normally is less than 0.02. The value of q/kT is approximately 39 siemens per ampere. A load impedance that gives an overall stage voltage gain of 10 or less in a circuit will usually operate in the proper mode and can be stabilized. Should the load impedance determined in this way prove to be too large, a radio-frequency transformer may be used to reduce the impedance level. The critical issue is that the input circuit to the amplifier look into a voltage source so that input loading will not degrade the operation of the previous amplifier. *See* CIRCUIT (ELECTRONICS); RADIO-FREQUENCY AMPLIFIER; SIGNAL PROCESSING.

Keats A. Pullen

Bibliography. W. M. Flanagan, *Handbook of Transformer Design and Applications*, 2d ed., 1993; I. M. Gottlieb, *Practical Transformer Handbook*, 1998; M. J. Heathcote, *J&P Transformer Book*, 12th ed., 1998; S. L. Herman and D. Singleton, *Delmar's Standard Guide to Transformers*, 1996; D. Horne (ed.), *Electrical Transformer Handbook*, vol. 2, 2005; R. W. Hurstled (ed.), *Electrical Transformer Handbook*, vol. 1, 2004; R. Lee, C. E. Carter, and L. Wilson, *Electronic Transformers and Circuits*, 3d ed., 1988; E. Lowdon, *Practical Transformer Design Handbook*, 2d ed., 1988; G. McPherson and R. D. Laramore, *An Introduction to Electrical Machines and Transformers*, 2d ed., 1990; A. J. Pansini, *Electrical Transformers and Power Equipment*, 3d ed., 1999; R. M. Del Vecchio et al., *Transformer Design Principles with Applications to Core-Form Power Transformers*, 2002.

Transfusion

The administration of blood or its components as a part of a medical treatment. There are certain fairly well-delineated indications for the use of some form of transfusion. Hemorrhage, severe burns, and certain forms of shock are perhaps the most important

conditions for which blood transfusion is utilized. Other disorders in which hemotherapy may be indicated include hemophilia, leukemia, certain anemias, and rare hereditary or familial disorders in which some portion of the blood is lacking or deficient. *See* HEMATOLOGIC DISORDERS.

Blood groups. In order for a recipient to accept a blood transfusion, the donor blood cells must be immunologically compatible with the recipient. That is, the recipient must recognize certain molecules (antigens) on donor blood cells as "self" and not foreign. The three main antigen systems on blood cells are the ABO, Rh, and HLA.

The ABO antigens are the most important for transfusion of red blood cells. A person may have A, B, both A and B, or neither A nor B antigens on their blood cells. Persons who do not have A antigen have anti-A antibody in their plasma, which would destroy red cells having A antigen; the same conditions apply to the B antigen. Therefore, red cells can be transfused from one person to another only if the recipient does not have antibodies against the antigens of the donor. Since O-type individuals have neither antigen, their blood can be given to people with types A, B, or AB blood.

The Rh antigens are the next most important factor in transfusion of red blood cells. Persons who lack one or more of the Rh antigens on their red cells (that is, they are Rh negative) do not have antibody against the Rh antigen. However, if an Rh negative person has been previously exposed to Rh by transfusion of Rh positive blood or if an Rh negative woman has been pregnant with an Rh positive infant, then that individual will develop antibodies against the Rh antigen. *See* RH INCOMPATIBILITY.

The human leukocyte antigen (HLA) complex is another factor to be considered in transfusions. The HLA antigens are found on platelets. If the donor HLA antigens are not the same as those in the recipients, the recipient may produce antibodies that destroy the donor platelets. *See* BLOOD GROUPS.

Blood collection and component preparation. Collection and preparation of blood follows certain general standards worldwide, but specifics may vary. The following discussion applies to the United States. Blood donors are screened to identify risks for the donor, such as any possible adverse effects of losing a unit of blood, as well as for the potential recipient. Blood donation is safe for healthy individuals. The blood is collected with sterile, disposable equipment so that there is no possibility of disease transmission to the donor. To determine whether the donor's blood is safe for recipients, donors are asked to complete a questionnaire about their past and present physical condition. In addition, after the blood donation the donor can notify the blood bank confidentially of any concerns regarding the safety of the blood for recipients. Donated blood is tested for blood groups, blood-group antibodies, and laboratory evidence of syphilis, hepatitis B, AIDS, human T-cell lymphotropic virus type I (HTLV-I, which is associated with adult T-cell leukemia), and hepatitis C. As a result, blood transfusions have become safer. However, persons who may have been exposed to AIDS should not donate blood, because in rare cases it has been found that blood may test negative for AIDS and yet still be capable of transmitting AIDS to recipients. This situation can arise because there is a period of time during which a recently infected individual has not yet made sufficient antibody to test positive.

Because of the concern that AIDS can be transmitted by blood transfusion, patients sometimes request donations from specific family members and friends. In general, such directed donations are statistically no safer than volunteer blood donation. More patients are donating their own blood (autologous blood) before elective surgery, for their own use during and after the surgery. Autologous blood is the safest blood for transfusion.

Generally, each donation consists of 1 pint (450 ml) of blood. The blood, which is collected in sterile plastic bags, can be separated into several components, such as red blood cells, plasma, and platelets (see **table**). Each component may be used as needed; often, these components are given to different recipients according to their specific needs. In addition, plasma from different donors can be pooled and made into plasma components such as Factor VIII or IX concentrate for hemophiliacs. *See* HEMOPHILIA.

Uses. Shortly before transfusion, the blood of the donor and recipient is tested once again to make sure that the blood groups are compatible. In an emergency, these tests are abbreviated, or type O red blood cells which can be transfused safely to any individual, are used. Transfusion of blood and blood components is essential to support many patients undergoing surgery, treatment for cancer, or organ transplantation, as well as premature infants. *See* TRANSPLANTATION BIOLOGY.

Most patients undergoing elective surgery do not need blood tranfusions. In some instances, the blood shed by the patient can be collected during surgery and given back to the patient. However, in emergency surgery, patients must rely on blood from volunteer donors. For the rare case in which a patient has lost a great deal of blood, massive transfusion may create special problems. Sufficient blood of the

Storage and indications for use of whole blood and common blood components

	Storage	Indications
Whole blood	35 days at 39°F (4°C)	Massive bleeding
Red blood cells	42 days at 39°F (4°C)	Inadequate red blood cells to deliver oxygen to tissues
Plasma	12 months at 0°F (−18°C)	Multiple coagulation deficiencies inadequate to stop or prevent bleeding
Platelets	5 days at room temperature	Inadequate platelets to stop or prevent bleeding

correct type may not be available. Also, during transfusion such a patient may develop abnormal bleeding, low body temperature, or shock and may require special blood components such as platelets and plasma to stop the bleeding.

Cancer patients frequently need blood transfusions. For example, since blood is produced in the bone marrow, patients with cancer of the bone marrow may fail to make enough of their own blood. Treatment with drugs and radiation often suppresses the production of blood in the bone marrow even further, so that transfusions must be given until the patient's bone marrow recovers. *See* CANCER (MEDICINE).

Premature infants often need blood transfusions, partly because blood drawn for laboratory tests constitutes a large percentage of their blood volume. They are, however, especially susceptible to certain diseases that are transmitted by transfusion, such as cytomegalovirus, and may need special blood.

Adverse effects. In addition to transmission of infections, adverse effects of blood transfusion are due to immune reactions between donor and recipient. Fever, the most frequent reaction, is caused by reaction of recipient antibody against donor white blood cells. Hives are due to allergic reactions to substances in donor plasma. Destruction of donor red cells (hemolytic reaction) occurs if the wrong type of blood is inadvertently given, or if recipient antibody is not detected prior to transfusion. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); BLOOD.

Pearl Toy

Bibliography. E. Rossi, T. Simon, and G. Moss (eds.), *Principles of Transfusion Medicine*, 2d ed., 1996; R. Walker (ed.), *Technical Manual*, 11th ed., 1993.

Transistor

A solid state device involved in amplifying small electrical signals and in processing of digital information. Transistors act as the key element in amplification, detection, and switching of electrical voltages and currents. They are the active electronic component in all electronic systems which convert battery power to signal power. Almost every type of transistor is produced in some form of semiconductor, often single-crystal materials, with silicon being the most prevalent. There are several different types of transistors, classified by how the internal mobile charges (electrons and holes) function. The main categories are bipolar junction transistors (BJTs) and field-effect transistors (FETs).

Single-crystal semiconductors, such as silicon from column 14 of the periodic table of chemical elements, can be produced with two different conduction species, majority and minority carriers. When made with, for example, 1 part per million of phosphorus (from column 15), the silicon is called *n*-type because it adds conduction electrons (negative charge) to form the majority carrier. When doped with boron (from column 13), it is called *p*-type because it has added positive mobile carriers called holes. For *n*-type doping, electrons are the majority carrier while holes become the minority carrier. For *p*-type doping holes are in larger numbers, hence are the majority carriers, while electrons are the minority carriers. All transistors are made up of regions of *n*-type and *p*-type semiconducting material. *See* SEMICONDUCTOR; SINGLE CRYSTAL.

The bipolar transistor has two conducting species, electrons and holes. Field-effect transistors can be called unipolar because their main conduction is by one carrier type, the majority carrier. Therefore, field-effect transistors are either *n*-channel (majority electrons) or *p*-channel (majority holes). For the bipolar transistor, there are two forms, n^+pn and p^+np, depending on which carrier is majority and which is the minority in a given region. As a result the bipolar transistor conducts by majority as well as by minority carriers. The n^+pn version is by far the most used as it has several distinct performance advantages, as does the *n*-channel for the field-effect transistors. (The n^+ indicates that the region is more heavily doped than the other two regions.)

Bipolar transistors. Bipolar transistors have additional categories: the homojunction for one type of semiconductor (all silicon), and heterojunction for more than one (particularly silicon and silicon-germanium, $Si/Si_{1-x}Ge_x/Si$). At present the silicon homojunction, usually called the BJT, is by far the most common. However, the highest performance (frequency and speed) is a result of the heterojunction bipolar transistor (HBT).

Bipolar transistors are manufactured in several different forms, each appropriate for a particular application. They are used at high frequencies, for switching circuits, in high-power applications, and under extreme environmental stress. The bipolar junction transistor may appear in discrete form as an individually encapsulated component, in monolithic form (made in and from a common material) in integrated circuits, or as a so-called chip in a thick-film or thin-film hybrid integrated circuit. In the *pn*-junction isolated integrated-circuit n^+pn bipolar transistor, an n^+ subcollector, or buried layer, serves as a low-resistance contact which is made on the top surface (**Fig. 1**). *See* INTEGRATED CIRCUITS; JUNCTION TRANSISTOR.

Basic operation. The n^+pn bipolar transistor (**Fig. 2***a*) has three differently doped regions and two junctions, the emitter-base junction and the collector-base junction, in a single crystal of silicon. It is

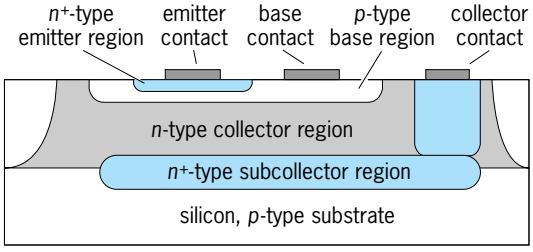


Fig. 1. Isolated n^+pn bipolar junction transistor for integrated-circuit operation.

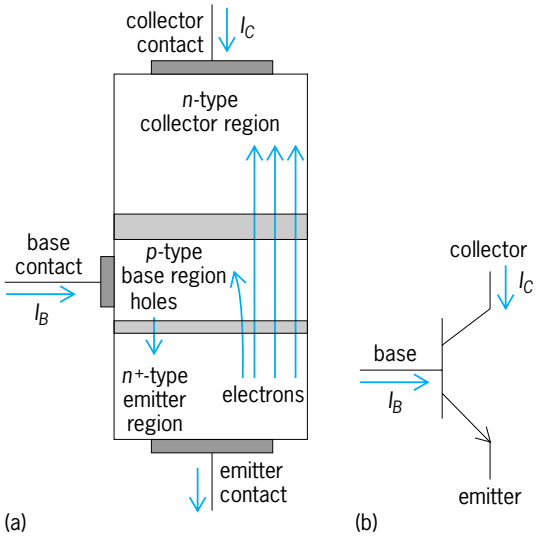


Fig. 2. Operation of an n^+pn bipolar junction transistor. (*a*) Conceptual cross section with carrier flows. (*b*) Circuit symbol.

possible to describe a large part of bipolar transistor operation by interpreting the device somewhat like a pair of back-to-back diodes. The emitter-base junction is usually forward-biased; that is, the voltage of the base with respect to the emitter, V_{BE}, is greater than 0. This voltage is small (less than 1 V), but the forward current across the junction is relatively large, just as in the forward-biased diode. The majority carriers in the emitter region (n^+-type) are electrons. These are injected across the emitter-base junction into the base region (p-type), which is quite thin, being on the order of 0.1–0.5 micrometer. With the base-collector junction reversed-biased (that is, the voltage of the collector with respect to the base, V_{CB}, is greater than 0), the holes that are present in the p-type base material do not cross it, as they are repelled by the electric field from the collector. However, the electrons that have been injected from the emitter into the very thin base region (where they become minority carriers) diffuse across the base-collector junction and are then collected under the influence of the positive collector potential (Fig. 2*a*). This electron current across the base-collector junction is almost as large as the electron current crossing from the emitter into the base region. In general, it runs from 99.5 to 99.99% of the emitter current, the small decrease being accounted for by electrons that are lost in the narrow base region. In terms of electron currents, for every 200 electrons flowing into the emitter from the external contact and crossing into the base region, perhaps one causes a current in the base lead, whereas 199 cross over into the collector and flow out into the external collector circuit. *See* DIFFUSION; JUNCTION DIODE; SEMICONDUCTOR DIODE.

The total current across any junction results from both hole and electron motion. In the emitter, electrons are injected from the n^+ region into the p region of the base, while holes are injected from the p-type base back into the emitter. These holes also cause current to flow into the base lead. The sum of these two currents crossing the junction determines the total emitter current. Even though holes and electrons flow in opposite directions, they add together as total current. However, in the emitter-base junction this hole current is usually several orders of magnitude less than the electron current, a consequence of much heavier doping in the emitter region compared to the base. The total base current is made up of two small components: the holes injected from the base to the emitter, and the electrons lost from the emitter due to recombination in the base. It is desirable for the base current to be as close to zero as possible, in order to obtain large current gains from base to collector.

Circuit symbol. The circuit symbol for an n^+pn bipolar transistor (Fig. 2*b*) has an arrowhead on the emitter lead whose direction indicates that electrons flowing in from the emitter to the base are the same as if positive charges were leaving that terminal. Hence, the arrowhead points out of the emitter.

Voltage-current characteristics. The input voltage-current characteristics of an n^+pn transistor (**Fig. 3***a*) are a family of curves of base current, I_B, versus base-emitter voltage, V_{BE}, for various values of collector-emitter voltage, V_{CE}. The output characteristics (Fig. 3*b*) are curves of collector current, I_C, versus collector-emitter voltage, V_{CE}, with the base current, I_B, as an independent parameter. These curves display two general characteristics. First, once the collector-emitter voltage is greater than 1 or 2 V, the collector current is relatively independent of this voltage; that is, each curve flattens out. Second, in this region the collector current is about 100 times the base current. The ratio of collector current to base current with the collector base voltage, V_{BE}, equal to zero is defined as the dc beta (β_{dc}) for the bipolar transistor.

The voltage-current characteristics emphasize the regions where the emitter-base junction is forward-biased (the base-emitter voltage, V_{BE}, is greater than 0) and the collector-base junction is reverse-biased (the collector-base voltage, V_{CB}, is greater than 0). These bias conditions define the active region for a bipolar transistor, which is mainly used for analog circuits. A quick glance at the input characteristics shows that the base current is greater than 0 under these conditions. When both junctions are reverse-biased, any currents that flow are several orders of magnitude smaller, and this is termed the cutoff region ("off" for digital circuits). The condition that both junctions are forward-biased defines the saturation region or the "on" region for digital circuits. Both the base-emitter voltage, V_{BE}, and the base-collector voltage, V_{BC}, are small positive values; and the collector-emitter voltage, V_{CE}, their difference, is also small and positive. The boundary between the active and saturation regions occurs where the collector-base voltage, V_{CB}, equals 0.

p^+np transistor. The p^+np transistor contains a narrow n-type base layer sandwiched between p-type emitter and collector regions. Forward bias on the emitter-base junction causes holes to be injected into

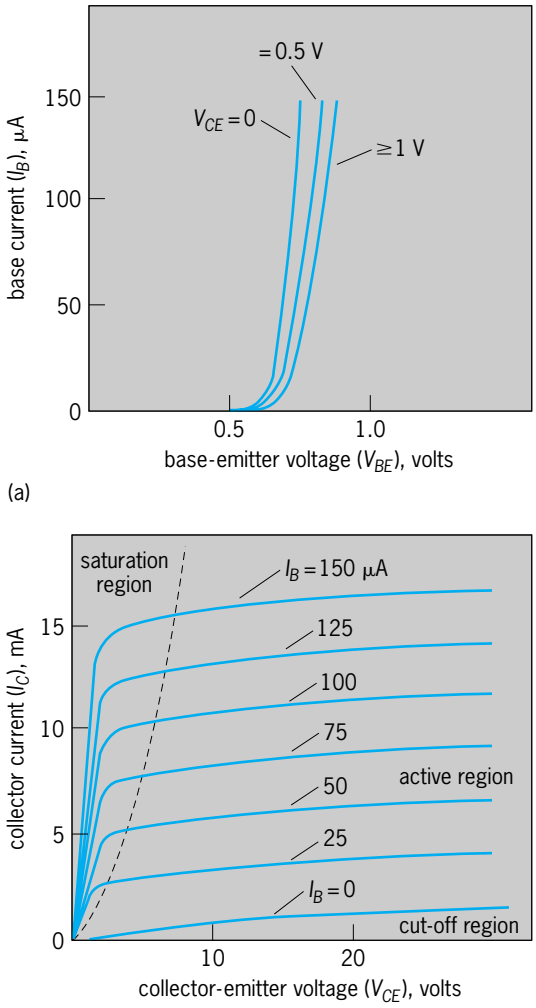


Fig. 3. Voltage-current characteristics of a typical low-power n^+pn bipolar transistor. (*a*) Input characteristics: base current (I_B) versus base-emitte voltage (V_{BE}) for various values of collector-emitter voltage (V_{CE}). When V_{CE} >1 V, the curves coincide. (*b*) Output characteristics: collector current (I_C) versus collector-emitter voltage (V_{CE}) for various values of base current (I_B). (*After W. H. Hayt, Jr., and G. W. Neudeck, Elecronic Circuit Analysis and Design, 2d ed., Wiley, reprint, 1995*)

the base region, most of which diffuse across the thin base and are collected by the reverse-biased base-collector junction. The total emitter current, I_E, is thus composed of the sum of this large hole current plus a much smaller electron current directed from the base toward the emitter.

Field-effect transistors. Majority-carrier field-effect transistors are classified as metal-oxide-semiconductor field-effect transistor (MOSFET), junction "gate" field-effect transistor (JFET), and metal "gate" on semiconductor field-effect transistor (MESFET) devices. MOSFETs are the most used in almost all computers and system applications. However, the MESFET has high-frequency applications in gallium arsenide (GaAs), and the silicon JFET has low-electrical noise performance for audio components and instruments. In general, the *n*-channel field-effect transistors are preferred because of larger electron mobilities, which translate into higher speed and frequency of operation.

MOSFETs. An *n*-channel MOSFET (**Fig. 4**) has a so-called source, which supplies electrons to the channel. These electrons travel through the channel and are removed by a drain electrode into the external circuit. A gate electrode is used to produce the channel or to remove the channel; hence it acts like a gate for the electrons, either providing a channel for them to flow from the source to the drain or blocking their flow (no channel). With a large enough voltage on the gate, the channel is formed, while at a low gate voltage it is not formed and blocks the electron flow to the drain. This type of MOSFET is called enhancement mode because the gate must have sufficiently large voltages to create a channel through which the electrons can flow. Another way of saying the same idea is that the device is normally "off" in an nonconducting state until the gate enhances the channel.

An *n*-channel MOSFET with a positive gate-source voltage, V_{GS}, and a small drain-source voltage, V_{DS}, has an electric field established across an insulating layer (Fig. 4). This field acts to repel positive carriers (holes) in the substrate and to attract negative carriers (electrons). As a result, a layer of substrate near the insulator becomes less *p*-type and its conductivity is reduced. As the gate-source voltage increases further, this surface region of the substrate eventually has more electrons than holes, and it inverts to *n*-type. Additional increases in gate voltage add more electrons to the channel and make it even more conductive. This *n*-channel (Fig. 4*a*) now conducts electrons from the n^+ source to the n^+ drain which has a positive voltage and attracts electrons. Between the *p*-type substrate and the *n*-type channel is a depletion (transition) region that serves to isolate the substrate from the channel, a process referred to as self-isolation. Since conduction is by electrons, the majority carrier, the MOSFET is a majority-carrier device.

The smallest value of the gate-source voltage, V_{GS}, that will produce a channel and a resultant value of drain current, I_D, greater than the few nanoamperes is called the threshold voltage, V_T, typically 0.2–2 V. V_T output voltage-current characteristics of the device are a family of curves of drain-current, I_D, versus drain-source voltage, V_{DS}, for several values of gate-source voltage, V_{GS} (**Fig. 5**). When the drain-source voltage is small (Fig. 5*a*), the device behaves as a voltage-controlled linear resistance. When the drain-source voltage becomes sufficiently large (Fig. 5*b*), the gate-to-drain voltage is less than the

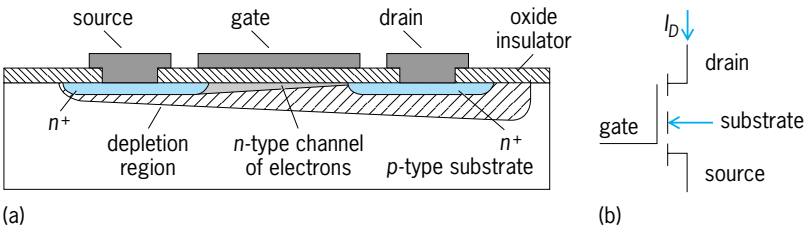


Fig. 4. An *n*-channel enhancement-mode metal-oxide-semiconductor field-effect transistor (MOSFET). (*a*) Cross section. (*b*) Standard circuit symbol.

threshold voltage, that is, Eq. (1) holds, and pinch-

$$V_{GD} = V_{GS} - V_{DS} \leq V_T \qquad (1)$$

off occurs at the drain end of the channel. Further increases in drain-source voltage do not lead to larger values of drain current, (that is, the current saturates), since the transistor is operating in the region beyond pinch-off. In this region of operation the MOSFET device behaves as a voltage-controlled current source.

The standard circuit symbol for the *n*-channel enhancement-mode MOSFET (Fig. 4*b*) shows the substrate as a separate connector. An arrow shows the direction from the *p* side (substrate) to the *n* side (channel) of the junction, while a segmented line indicates the enhancement mode; no channel is present until channel enhancement occurs at which point the gate-source voltage exceeds the threshold voltage.

The *p*-channel enhancement-mode MOSFET is the complement of the *n*-channel device. It has an *n*-type silicon substrate in which a *p*-type channel is induced (enhanced) by making the gate sufficiently negative that the gate-source voltage is less than the threshold voltage. The gate of a *p*-channel enhancement-mode MOSFET has an electric field between the gate and

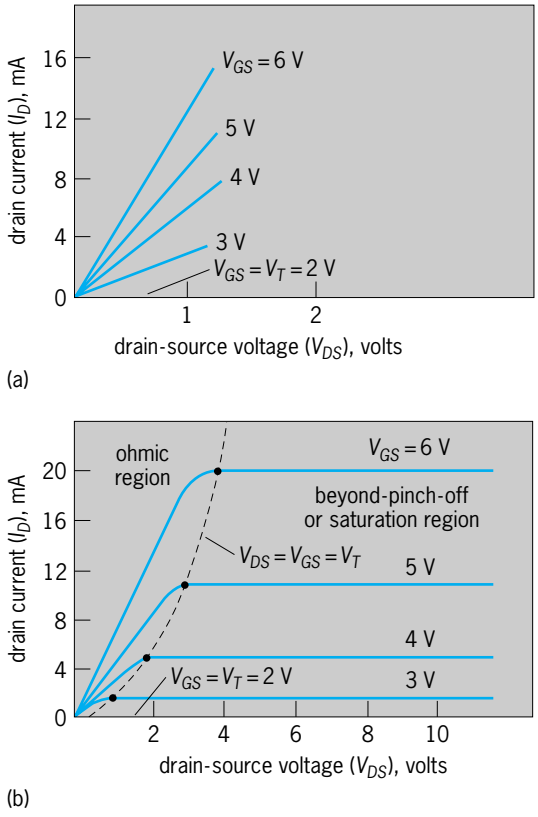


Fig. 5. Output characteristics of *n*-channel enhancement-mode MOSFET: drain current (I_D) versus drain-source voltage (V_{DS}) for various values of gate-source voltage (V_{GS}). (*a*) Small values of V_{DS}, where the device behaves as a voltage-controlled linear resistance. (*b*) Complete output characteristics. (*After W. H. Hayt, Jr., and G. W. Neudeck, Electronic Circuit Analysis and Design, 2d ed., reprint, Wiley, 1995*)

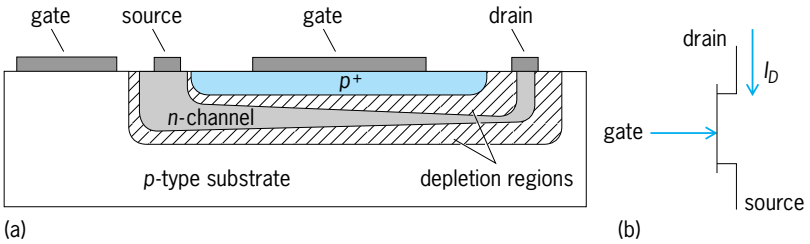


Fig. 6. An *n*-channel junction field-effect transistor (JFET). (*a*) Cross section. (*b*) Circuit symbol.

substrate which pushes out electrons, attracts holes, and eventually inverts the channel to *p* type. Now holes conduct between the p^+ source and drain electrodes.

JFETs. In the JFET (**Fig. 6***a*), a conducting majority-carrier *n* channel exists between the source and drain. When a negative voltage is applied to the p^+ gate, the depletion regions widen with reverse bias and begin to restrict the flow of electrons between the source and drain. At a large enough negative gate voltage (symbolized V_P), the channel pinches off. The standard circuit symbol (Fig. 6*b*) has a continuous bar since current flows with zero gate-source voltage, V_{GS}, at larger values of the drain source voltage, V_{DS}.

MESFETs. The MESFET is quite similar to the JFET in its mode of operation. A conduction channel is reduced and finally pinched off by a metal Schottky barrier placed directly on the semiconductor. Metal on gallium arsenide is extensively used for high-frequency communications because of the large mobility of electrons, good gain, and low noise characteristics. Its cross section is similar to that of the JFET (Fig. 6*a*), with a metal used as the gate. *See* SCHOTTKY BARRIER DIODE.

High-frequency transistors. High-frequency effects for the bipolar transistor are characterized by the emitter charging time (τ_e), the collector charging time (τ'_c), the minority-carrier transit time through the active base region (τ_b), and the base-collector depletion region transit time (τ_c). The emitter charging time equals the product of the emitter-base capacitance (proportional to the area of the emitter) and the thermal voltage divided by the dc current. The minority-carrier transit time through the active base region is approximately the square of the active width of the base region divided by twice the diffusion constant for the minority carriers that diffuse through the base. (The dependence on the active width indicates the need for a very thin base region.) The transit time through the collector-to-base depletion region equals the width or this region divided by twice a saturated velocity to which the carriers can accelerate. Thus, a short transit time requires a large saturated velocity or a small width, which means a small value of the collector-to-base voltage. The final term for the collector is its charging time, approximately the product of the collector contact resistance and the collector-base capacitance. A short charging time thus requires a small value of the former and a small collector area to reduce the latter.

Figures of merit. A figure of merit for the advanced bipolar transistor is the frequency, f_T, at which the short-circuit, current-signal gain is unity. This frequency equals the inverse of the sum of the four times discussed above, τ_e, τ_b, τ_c, and τ'_c. A large value of f_T indicates that the intrinsic device is fast.

A more circuit-oriented figure of merit is $f_{\max}$, the maximum frequency that gain can still be achieved in a circuit, given by Eq. (2). Here the external base

$$f_{\max} \cong \sqrt{\frac{f_T}{8\pi R_b C_{cb}}} \tag{2}$$

resistance, R_b, is important, as well as the base-collector area needed to reduce the collector base capacitance, C_{cb}.

Most very high speed logic circuits belong to the emitter-coupled logic (ECL) family of circuits or the current-mode logic (CML) family. The figure of merit for this type of circuit is given by Eq. (3), where R_c is

$$\tau_{cs} = 1.7\sqrt{\frac{(R_c + 2R_b)(3C_{cb} + C_{cs})}{2\pi f_{T\max}}} \tag{3}$$

the collector resistance, C_{cs} is the collector-substrate capacitance of the integrated bipolar transistor, and $f_{T\max}$ is the peak value of f_T when the collector current is varied. Again, this expression indicates the need for thin base regions, small emitter and collector areas, and low values of resistances contacting the device. *See* LOGIC CIRCUITS.

Structural improvements. The function of the subcollector in the integrated-circuit bipolar transistor (Fig. 1) is to reduce the collector resistance. Typical values of the current gain, that is, the dc beta, range from 80 to 300, and f_T ranges from 5 to 45 GHz with values of $f_{\max}$ up to 450 GHz. In an emitter-coupled logic circuit the transistor has a gate delay of as low as 15 picoseconds. An improvement to this structure is to reduce the sidewall components of capacitance with the local-oxidation-of-silicon (LOCOS) structure. In addition, a polysilicon-contacted emitter can be added to improve the dc beta, and the external base resistance can be reduced by increased base doping, somewhat similar to what is done in heterojunction bipolar transistors.

Single, self-aligned transistor. The single self-aligned bipolar transistor (SST) reduces the emitter area to 0.35×5 μm and has f_T values up to 20 GHz and a dc beta of 180. The use of a pedestal collector and double self-alignment improves the value of $f_{\max}$ and the emitter charging time by reducing the area and the external parasitic resistances. In all these cases the fabrication methods strive to reduce the area, hence the capacitances.

Heterojunction bipolar transistor. The heterojunction bipolar transitor is made from two different types of semiconductor material. The most promising is the silicon-germanium type. It is produced by epitaxially growing a narrow band-gap base region of heavily doped p-type $Si_{1-x}Ge_x$ on an n-type silicon collector and then capping it with an n^+ type silicon emitter. The silicon-germanium compound suppresses the base-injected holes (Fig. 2), and at the same time this allows the base to be doped very heavily to reduce the external base resistance. By grading the germanium content and the doping, f_T, values up to 32 GHz and $f_{\max}$ of 120 GHz have been achieved, with good values of beta. The circuit delay is about 20 ps. Other heterojunction bipolar transistors of interest include those using the compound semiconductors GaAlAs/GaAs, InGaP/GaAs, and InGaAs/InP. These devices have achieved f_T values of 37 GHz and $f_{\max}$ of 90 GHz with powers of 1–5 W. *See* SEMICONDUCTOR.

High-frequency field-effect transistors. The inability of the MOSFET to conduct large currents into capacitive loads has limited its use in extremely high-frequency circuits. However, because of its low power consumption it can be integrated into very dense circuits. The first requirement is that the channel length be small (approximately 1 μm), as it controls how fast the majority carrier can traverse between the source and drain. The carrier mobility must also be large; hence, electrons are preferred, as their mobility is typically two to three times larger than holes. A second requirement is for low values of source and drain resistance. In circuit applications a small value of capacitance between the gate and drain is necessary to reduce the total effective capacitance that is multiplied by the circuit voltage gain. Self-aligned gates, polysilicon, and channel lengths of less than 0.15 μm are used. Typical performance characteristics for a 0.5-μm gate length are an f_T of 10 GHz and an $f_{\max}$ of 15 GHz. In complementary metal-oxide-semiconductor (CMOS) circuits with gate lengths of 0.15 μm, gate delays as low as 21 ps per stage are possible.

The more advanced techniques use silicon-on-insulator (SOI) technology to further reduce the external parasitic capacitances around the source and drain. Other device structures include the high electron mobility transistor (HEMT), silicon-germanium MOSFETs, and combinations of bipolar transistors and MOSFETs (BiCMOS). Each technology has its particular advantages. The HEMT is produced from compound semiconductors and can yield an f_T of 300 GHz with gate delays of 25 ps. *See* MICROWAVE SOLID-STATE DEVICES.

Gerold W. Neudeck

Models. Whether the transistor is used in the design of small analog circuits or very large scale integrated circuits, its behavior has to be adequately understood by the designer. Analysis of the circuit is a prerequisite to its fabrication, thus pointing to the need for models. The higher levels of integration as well as of the cost of fabrication have increased the need for more accurate models and also their complexity. Circuit simulation programs have become rather commonplace and generally available for use on personal computers. The usefulness of such computer-aided design programs is directly influenced by the accuracy of the transistor models and their adequacy for the design application. In general the models can be categorized as large-signal (nonlinear) models used for dc or transient analysis, and

small-signal (linear) models used for ac or frequency-domain analysis.

Most large-signal models are represented by systems of equations relating currents and charges to terminal voltages. Different equations are typically used for different combinations of terminal voltages or regions of operation.

In many analog circuits, the signals are small enough that the nonlinear models can be replaced by linearized equivalent circuit models. Linear circuits are much less complicated to analyze than nonlinear ones. The hybrid-π configuration can be used for linear modeling of field-effect transistors of bipolar junction transistors. *See* AMPLIFIER; CIRCUIT (ELECTRONICS); ELECTRICAL MODEL.

Michael Artaki; Robert M. Fox

Bibliography. W. H. Hayt, Jr., and G. W. Neudeck, *Electronic Circuit Analysis and Design*, 2d ed., 1984, reprint 1995; Institute of Electrical and Electronics Engineers, *1994 International Electron Devices Meeting Technical Digest*, San Francisco, California, December 11–14 1994; G. Massobrio and P. Antognetti, *Semiconductor Device Modeling with SPICE*, 2d ed., 1998; G. W Neudeck, *The Bipolar Transistor*, 2d ed., 1989; G. W. Neudeck, *The P-N Junction Diode*, 2d ed., 1989.

Transit (astronomy)

The apparent passage of a celestial body across the apparent disk of a larger body, such as a planet across its parent star or of a satellite across its parent planet; also, the apparent passage of a celestial object or reference point across an adopted line of reference in a celestial coordinate system. Classically, the observed data were instants of internal and external tangency of the disks (contacts) at ingress and egress of the smaller body. In the modern era, data may also include the differential brightness of the two disks and the duration of any change of brightness.

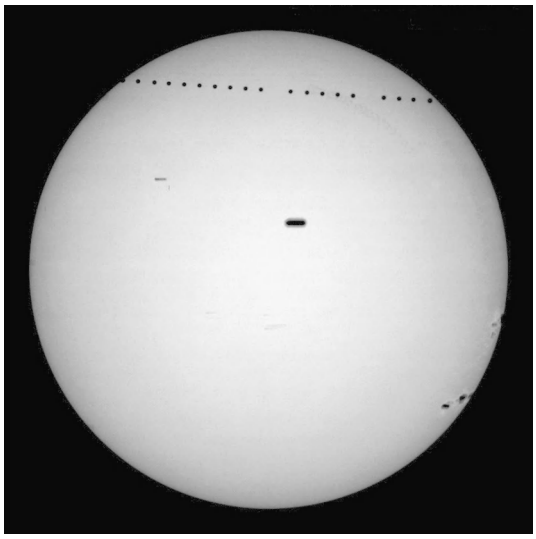
Transits of Mercury and Venus. Mercury and Venus are the only planets whose orbits lie between the Earth and the Sun and thus can be seen from Earth to cross the disk of the Sun. The conditions are that the planet is in inferior conjunction at the same time that it passes one of the two nodes of its orbit, thus putting it essentially in a straight line between the Earth and the Sun. The first transit predicted and observed was that of Mercury on November 7, 1631. Venus transited exactly 1 month later. From then through the early twentieth century, transits of Mercury were observed for the purpose of getting precise positions of the planet to improve knowledge of its orbit, and transits of Venus to determine the solar parallax. The inherent errors of the observations, caused primarily by Earth's atmosphere, made analysis so difficult that there was great disagreement among astronomers as to their value.

Before these difficulties were known, Edmond Halley proposed the observation of Venus transits for determination of the solar parallax, one of the foremost objectives of astronomy. [The solar parallax is the angle, p, subtended by the Earth's equatorial radius, r, at the mean distance of the Sun from the Earth, a; $p = r/a$ when p is measured in radians. Measurement of the solar parallax determines the astronomical unit (AU) since 1 AU $= a = r/p$.] Dozens of international expeditions were sent to observe the four transits of Venus in the eighteenth and nineteenth centuries. These strained the bounds of contemporary methods of observation and analysis, and the value of the results was controversial. In hindsight, the value of the solar parallax derived from them is closer to the modern value than that obtained by any other method of the time. *See* ASTRONOMICAL UNIT; PARALLAX (ASTRONOMY).

In a century, there are 13 or 14 transits of Mercury. In the current era, the Earth crosses the line of nodes of Mercury's orbit each year on May 8 or 9 and November 10 or 11. If Mercury is in inferior conjunction within 3 days of the May crossing or 5 days of the November crossing, a transit will occur. November transits occur twice as often as May transits. The last two occurred on November 15, 1999, and May 7, 2003 (see **illus.**). The next one occurs on November 8–9, 2006.

The size, shape, and orientation of the orbit of Venus causes transits to be very rare—only 81 between the years −2000 and +4000. At present, inferior conjunction must occur within 2 days of June 7 or December 9 for a transit to occur. Transits usually occur in pairs separated by 8 years, with 105.5 or 121.5 years between pairs. The last one occurred on December 6, 1882; the next pair occurs on June 8, 2004, and June 6, 2012. *See* MERCURY (PLANET); PLANET; VENUS.

Transits of Jupiter's satellites. Transits of the galilean satellites of Jupiter occur at each of their inferior conjunctions with the exception of satellite IV, which occasionally passes clear of the planet's disk.



Transit of Mercury, as seen by the GONG instrument at Udaipur, India, on May 7, 2003, Mercury is at the top of the image. The dark spot near the middle is a sunspot, smeared from the compositing process. (*Photo courtesy of Cliff Turner, National Solar Observatory, Global Oscillation Network Group, National Science Foundation*)

They are difficult to observe and are used mainly to estimate the albedo (reflectivity) of the satellites relative to that of Jupiter. As each satellite passes in front of the planet, it casts its shadow on the planet's disk and causes the phenomenon of shadow-transit. *See* JUPITER; SATELLITE (ASTRONOMY).

Meridian transits. Until the close of the twentieth century, passages of stars and other celestial bodies across the local meridian were observed extensively for determining precise coordinates of the stars and planets, accurate time, or the position of the observer. The instrument commonly used is often called a transit circle. This type of observation has been almost completely superseded by interferometric methods from Earth's surface and orbiting satellites, and by other astrometric observations from spacecraft. *See* ASTRONOMICAL COORDINATE SYSTEMS; ASTRONOMICAL TRANSIT INSTRUMENT.

Transits of extrasolar planets. The search for planets around solarlike stars other than the Sun yielded the first positive results in 1995. More than 100 planets have been found since, and the number is rising rapidly. They have been found by techniques using astrometry and radial velocity. The most recently developed technique is to detect photometrically the minute decrease in brightness of a star as an orbiting planet crosses or transits its face. This can occur only if the planet's orbital plane lies nearly edge-on to the Earth. The photometric method can be used for fainter and more distant stars than the other techniques.

As a tool for verification of discovery by other methods, the first planetary companion detected this way was reported in 1999, orbiting the star HD 209458 in Pegasus. Even though that companion was larger than Jupiter, the technique is considered the most mature for detecting Earth-class extrasolar planets, that is, those that are 0.5–10 times the size of Earth. In December 2002 the first planet was initially discovered photometrically, orbiting the star OGLE-TR-56 in the next spiral arm of the Milky Way Galaxy. Within 6 months, two more planets were announced within the same group of candidate stars.

The observed data are the period of recurrence of the transit (frequency), the duration of the transit (expected to be 2–16 hours), and the decrease in brightness of the star, which may be as little as 1 part in 12,000. These values must be consistent for at least three occurrences to be considered confirmation of a single planet in transit. Otherwise there might be more than one, or a physical variability in the star occurring faster than the duration of the transit.

The star's spectral type is used to estimate its mass, diameter, and brightness. The frequency of transit gives the orbital period and, through Kepler's third law, the semimajor axis of the orbit. The fractional change in brightness gives the ratio of cross-sectional areas, or diameters, of the planet and star. From all this, the planet's temperature and density may be deduced. *See* KEPLER'S LAWS; SPECTRAL TYPE; STAR.

From statistical considerations of the orientation of orbits, 1% of solarlike stars with planets would show transits of inner planets. Very high precision photometry with sensitive charge-coupled-device (CCD) detectors and regular monitoring of hundreds of thousands of faint stars, using relatively small telescopes, makes Earth-based searching practical. At least three space missions planned for launch in 2005 or later [Kepler, FRESIP (Frequency of Earth-Sized Inner Planets), and COROT (Convection, Rotation, and Planetary Transits)] are proposed for continuous and simultaneous photometric monitoring of 5000–160,000 stars for up to 8 years. They are expected to find 50–480 or more Earth-class planets, and a smaller number of giant planets in transit. *See* CHARGE-COUPLED DEVICES.

Alan D. Fiala

Bibliography. Better size for the transiting exoplanet, *Sky Telesc.*, 101(1):29, January 2001; S. Dreizler et al., OGLE-TR-3: A possible new transiting planet, *Astron. Astrophys.*, 402:791–799, 2003; Extrasolar planet seen transiting its star, *Sky Telesc.*, 99(2):16–17, February 2000; J. Meeus, *Transits*, Willmann-Bell, Richmond, 1989; A new way to find planets, *Sky Telesc.*, 105(4):9, April 2003; P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomical Almanac*, University Science Books, Mill Valley, CA, 1992.

Transition elements

In broad definition, the elements of atomic numbers 21–31, 39–49, and 71–81, inclusive. The symbols of these elements, along with their atomic numbers and valence-shell electronic configurations, are given in **Fig. 1**. The elements are arranged in the order in which they appear in the long, or Bohr, form of the periodic table.

A more restricted classification of the transition elements, preferred by many chemists, is indicated by the heavy border drawn about the central portion of the table. All of the elements in this section of the table have one or more electrons present in an unfilled *d* subshell in at least one well-known oxidation state.

Chemical properties. In their compounds, the transition elements tend to exhibit multiple valency, the maximum valence increasing from 3+ at the beginning of a series (Sc, Y, Lu) to 8+ at the fifth member (Mn, Re). For the elements in any vertical column, the highest oxidation state is usually observed in the element at the bottom of the column. Thus the highest oxidation state of iron is 6+, whereas osmium attains an oxidation state of 8+.

One of the most characteristic features of the transition elements is the ease with which most of them form stable complex ions. Features which contribute to this ability are favorably high charge-to-radius ratios and the availability of unfilled *d* orbitals which may be used in bonding. Examples of such complexes include a very stable cyanide complex of aurous gold $Au(CN)_2^-$, of commercial importance for the recovery of the metal from low-grade ores; a similar complex $Ag(CN)_2^-$, useful in obtaining bright, firmly adherent deposits of silver by electroplating; and numerous ammonia complexes, of which the

21 Sc $3d4s^2$	22 Ti $3d^24s^2$	23 V $3d^34s^2$	24 Cr $3d^54s$	25 Mn $3d^54s^2$	26 Fe $3d^64s^2$	27 Co $3d^74s^2$	28 Ni $3d^84s^2$	29 Cu $3d^{10}4s$	30 Zn $3d^{10}4s^2$	31 Ga $3d^{10}4s^24p$
39 Y $4d5s^2$	40 Zr $4d^25s^2$	41 Nb $4d^45s$	42 Mo $4d^55s$	43 Tc $4d^65s$	44 Ru $4d^75s$	45 Rh $4d^85s$	46 Pd $4d^{10}$	47 Ag $4d^{10}5s$	48 Cd $4d^{10}5s^2$	49 In $4d^{10}5s^25p$
71 Lu $5d6s^2$	72 Hf $5d^26s^2$	73 Ta $5d^36s^2$	74 W $5d^46s^2$	75 Re $5d^56s^2$	76 Os $5d^66s^2$	77 Ir $5d^9$	78 Pt $5d^96s$	79 Au $5d^{10}6s$	80 Hg $5d^{10}6s^2$	81 Tl $5d^{10}6s^26p$

Fig. 1. Portion of periodic table showing transition elements, their atomic number, symbol, and electron configurations in valence shells; heavy border indicates the more restricted classification.

deep blue $Cu(NH_3)_4^{2+}$ is a representative example, widely used in colorimetric analyses for copper. Vitamin B_{12} is an example of a cobalt(III) complex that is important in nutrition; and hemin, the red pigment of blood, is an important iron(II) complex.

Most of the ions and compounds of the transition metals are colored, and many of them are paramagnetic; that is, when they are placed in a magnetic field, the magnetic flux within the compound is higher than that of the surrounding field. Both color and paramagnetism are related to the presence of unpaired electrons in the *d* subshell. Excitation of these relatively loosely bound electrons to higher energy states accounts for the absorption of light in the visible region of the spectrum, while the magnetic field associated with the electron spin is responsible for the magnetic behavior of the compounds. Study of the magnetic behavior of compounds and complex ions of the transition elements has contributed much to an understanding of chemical bonding in these elements, since utilization of the *d* electrons in bonding involves electron-pair formation, with consequent cancellation of the magnetic moments and altered magnetic properties. Because of their ability to accept electrons in unoccupied *d* orbitals, transition elements and their compounds frequently exhibit catalytic properties.

Many of the most important catalysts, such as nickel used in hydrogenation, are transition elements.

Broadly speaking, the properties of the transition elements are intermediate between those of the so-called representative elements, in which the subshells are completely occupied by electrons (alkali metals, halogen elements), and those of the inner or *f* transition elements, in which the subshell orbitals play a much less significant role in influencing chemical properties (rare-earth elements, actinide elements). *See* ATOMIC STRUCTURE AND SPECTRA; CATALYSIS; COORDINATION CHEMISTRY; COORDINATION COMPLEXES; MAGNETOCHEMISTRY.

The metals and their uses. All the transition elements are metals and, in general, are characterized by high densities, high melting points, and low vapor pressures. Included, for example, are tungsten and tantalum, melting at 3370 and 3030°C (6098 and 5486°F), respectively. At room temperature, the vapor pressure of tungsten is so low that it compares to only one gaseous atom in a volume of space equal to that of the known sidereal universe. In general, those properties related to strong cohesiveness or binding between the atoms in the metallic state, such as high density, extreme hardness, and high melting point, reach a broad maximum in the neighborhood of the fourth member of each series. Within a given subgroup, these same properties tend to increase with increasing atomic weight. Facility in the formation of metallic bonds is demonstrated also by the existence of a wide variety of alloys between different transition metals. The variation in some representative properties of the transition elements as a function of atomic number is shown in the **Fig. 2**.

The transition elements include most metals of major economic importance, such as the relatively abundant iron and nickel, on one hand, and the rarer coinage metals, copper, silver, and gold, on the other. Also included is the rare and relatively

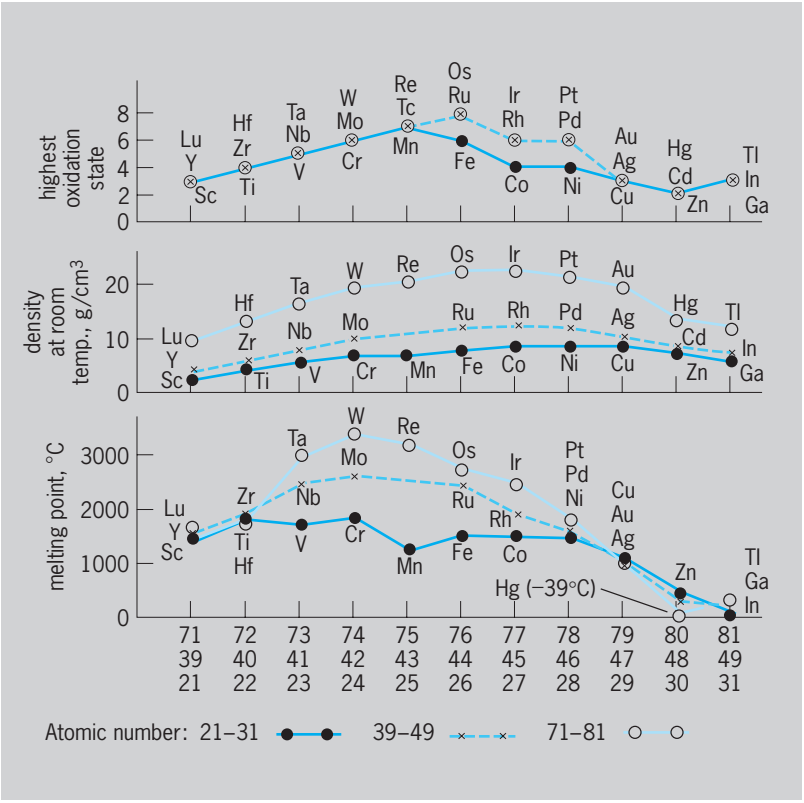


Fig. 2. Physical properties of transition elements as function of their atomic number.

unfamiliar element technetium, which is not found naturally in the terrestrial environment but is available in small amounts as a product of nuclear fission. Burris B. Cunningham

Bibliography. F. Basolo et al. (eds.), *Transition Metal Chemistry*, vol. 1, 1973, vol. 2, 1977; F. Basolo and J. L. Burmeister (eds.), *On Being Well-Coordinated: A Half-Century of Research on Transition Metal Complexes*, 2003; R. H. Crabtree, *The Organometallic Chemistry of the Transition Metals*, 3d ed., 2000; K. H. Whitmire, *The Interface of Main Group and Transition Metal Cluster Chemistry,* 1988.

Transition point

The point at which a substance changes from one state of aggregation to another. This general definition would include the melting point (transition from solid to liquid), boiling point (liquid to gas), or sublimation point (solid to gas); but in practice the term transition point is usually restricted to the transition from one solid phase to another, that is, the temperature (for a fixed pressure, usually 1 atm or 100 kilopascals) at which a substance changes from one crystal structure to another.

Some typical examples of transition points are:

$$\beta\text{-Fe (body-centered cubic)} \xrightarrow[(1664.3^\circ F)]{\text{at 1180 K}} \gamma\text{-Fe (face-centered cubic)}$$

$$S_8 \text{ (rhombic)} \xrightarrow[(203.5^\circ F)]{\text{at 369 K}} S_8 \text{ (monoclinic)}$$

$$CCl_4 \text{ (monoclinic)} \xrightarrow[(-53.8^\circ F)]{\text{225.5 K}} CCl_4 \text{ (tetragonal)}$$

$$NH_4NO_3 \text{ (}\beta\text{-rhombic)} \xrightarrow[(90.2^\circ F)]{\text{at 305.3 K}} NH_4NO_3 \text{ (}\alpha\text{-rhombic)}$$

$$NH_4NO_3 \text{ (}\alpha\text{-rhombic)} \xrightarrow[(183.6^\circ F)]{\text{at 357.4 K}} NH_4NO_3 \text{ (trigonal)}$$

Another kind of transition point is the culmination of a gradual change (for example, the loss of ferromagnetism in iron or nickel) at the lambda point, or Curie point. This behavior is typical of second-order transitions. *See* BOILING POINT; MELTING POINT; PHASE EQUILIBRIUM; SUBLIMATION; TRIPLE POINT. Robert L. Scott

Transition radiation detectors

Detectors of energetic charged particles that make use of radiation emitted as the particle crosses boundaries between regions with different indices of refraction. An energetic charged particle moving through matter momentarily polarizes the material nearby. If the particle crosses a boundary where the index of refraction changes, the change in polarization gives rise to the emission of electromagnetic transition radiation. About one photon is emitted for every 100 boundaries crossed, for transitions between air and matter of ordinary density. Transition radiation is emitted even if the velocity of the particle is less than the light velocity of a given wavelength, in contrast to Cerenkov radiation. Consequently, this radiation can take place in the x-ray region of the spectrum where there is no Cerenkov radiation, because the index of refraction is less than one. *See* CERENKOV RADIATION; REFRACTION OF WAVES.

The radiation extends to frequencies greater than the plasma frequency by the factor γ = particle energy divided by particle mass. The production of x-rays requires γ equal to or greater than 1000. A threshold as high as this is difficult to achieve by other means. This fact has led to the application of this effect for the identification of high-energy particles. For example, electrons of about 10^9 eV will produce x-rays of a few kiloelectronvolts, while the threshold for pions is on the order of 10^{11} eV. The solid material should be of low atomic number, carbon or lighter, to minimize absorption of x-rays, which are emitted close to the particle direction. The material is often in the form of foils, which must be of the order of 0.0004 in. (0.01 mm) thick to avoid destructive interference of the radiation from the two surfaces, and similarly, the foil spacing is typically 0.004 in (0.1 mm). Random assemblies of fibers or foams are almost as effective as periodic arrays of foils. Effective electron detectors have been made with several hundred foils followed by a xenon proportional chamber for x-ray detection. *See* PARTICLE DETECTOR; PLASMA (PHYSICS); RELATIVISTIC MECHANICS. William J. Willis

Bibliography. X. Artru, G. B. Yodh, and G. Menessier, Practical theory of the multilayered transition radiation detector, *Phys. Rev. D,* 12:1289–1306, 1975; J. Cobb et al., Transition radiators for electron identification at the CERN ISR, *Nucl. Instrum. Meth.,* 140:413–427, 1977; G. M. Garibian, Transition radiation effects in particle energy losses, *Sov. Phys.—JETP,* 10:372–376, 1960.

Translucent medium

A medium which transmits rays of light so diffused that objects cannot be seen distinctly; that is, the medium is only partially transparent. Familiar examples are various forms of glass which admit considerable light but impede vision. Inasmuch as the term translucent seems to imply seeing, usage of the term is ordinarily limited to the visible region of the spectrum. M. G. Mellon

Transmission lines

A system of conductors suitable for conducting electric power or signals between two or more termini. Transmission lines take many forms in practice and have application in many disciplines. For example, they traverse the countryside, carrying telephone signals and electric power. The same transmission

lines, with similar functions, may be hidden above false ceilings in urban buildings. With the need to reliably and securely transmit ever larger amounts of data, the required frequency of operation has increased from the high-frequency microwave range to the still higher frequency of light. Optical fibers are installed in data-intensive buildings and form a nationwide network. Increasing demand also requires that transmission lines handle greater values of electric power.

Transmission lines can, in some cases, be analyzed by using a fairly simple model that consists of distributed linear electrical components. Models of this type, with some permutations, can also be used to describe wave propagation in integrated circuits and along nerve fibers in animals. The study of hollow metal waveguides or optical fibers is usually based upon an analysis starting from Maxwell's equations rather than employing transmission-line models. Fundamentals and definitions that can initially be obtained from a circuit model of a transmission line carry over to waveguides, where the analysis is more complicated.

Coaxial cables and strip lines. Two particular types of transmission lines for communication that have received considerable attention are the coaxial cable (**Fig. 1***a*) and the strip line (Fig. 1*b*). The coaxial cable is a flexible transmission line and typically is used to connect two electronic instruments together in the laboratory. In the coaxial cable, a dielectric separates a center conducting wire from a concentric conducting sleeve. The strip line is used in integrated circuits to connect, say, two transistor circuits together. A strip line also has a dielectric that separates the top conducting element from the base, which may be an electrical ground plane in the circuit. If the conductivity of the metal is sufficiently high and the dielectric is lossless, both of which are reasonable approximations in practice, then it is possible to assume that the time-varying electric and magnetic fields that are associated with the propagating wave are polarized in a plane that is transverse to the direction of the wave propagation. This is similar to plane-wave propagation in free space and is called the transverse electric and magnetic (TEM) mode of propagation. In that case, the electric field is polarized in a direction that is between the two conductors, and the magnetic field is in a direction that is perpendicular to the electric field and to the direction of propagation. Losses in the transmission

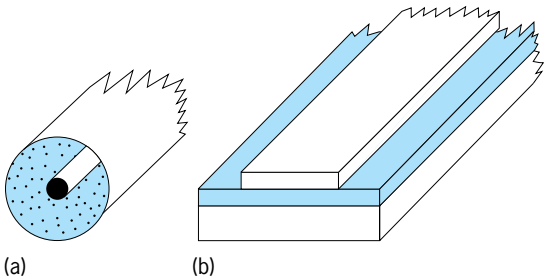


Fig. 1. Two common transmission lines for communication. (*a*) Coaxial cable. (*b*) Strip line.

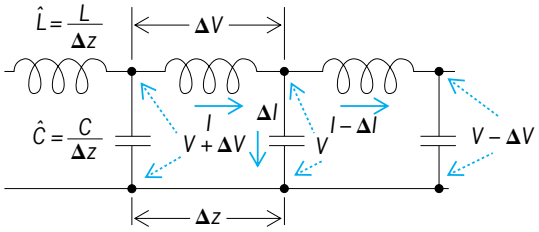


Fig. 2. Equivalent circuit that represents three sections of a lossless transmission line. Each section is of length Δz, and has an inductor (L) and capacitor (C). Voltages (V) and currents (I), and their differences (ΔV and ΔI) between successive sections are indicated.

line cause the signal to attenuate and become distorted as it propagates down the transmission line. *See* COAXIAL CABLE; INTEGRATED CIRCUITS.

Circuit model. These transmission lines can be most easily analyzed in terms of electrical circuit elements consisting of distributed linear inductors and capacitors. The values of these elements are in terms of the physical dimensions of the coaxial cable and the strip line, and the permittivity of the dielectric. Each of the elements is interpreted to be measured in terms of a unit length of the element. An equivalent circuit (**Fig. 2**) represents either the coaxial cable or the strip line as well as other transmission lines such as two parallel wires. *See* ALTERNATING-CURRENT CIRCUIT THEORY; CAPACITANCE; CAPACITOR; INDUCTANCE; INDUCTOR; PERMITTIVITY.

Losses are incorporated into the transmission-line model with the addition of a distributed resistance in series with the inductor and a distributed conductance in parallel with the capacitor. Additional distributed circuit elements can be incorporated into the model in order to describe additional effects. For example, the linear capacitors could be replaced with reverse-biased varactor diodes and the propagation of nonlinear solitons could be studied. *See* CONDUCTANCE; ELECTRICAL RESISTANCE; SOLITON; VARACTOR.

Telegraphist's equations. The circuit model (Fig. 2) can be analyzed by using basic laws from circuit theory in order to derive the wave equation for waves that propagate along the transmission line. The voltage difference ΔV between two adjacent nodes equals the voltage across the inductor. A current ΔI leaves a node and passes through the adjacent capacitor. Equations (1) and (2) describe these voltages

$$\Delta V(z,t) = \hat{L}\Delta z \frac{\partial I(z,t)}{\partial t} \Rightarrow \frac{\partial V(z,t)}{\partial z} = \hat{L}\frac{\partial I(z,t)}{\partial t} \tag{1}$$

$$\Delta I(z,t) = \hat{C}\Delta z \frac{\partial V(z,t)}{\partial t} \Rightarrow \frac{\partial I(z,t)}{\partial z} = \hat{C}\frac{\partial V(z,t)}{\partial t} \tag{2}$$

and currents, where Δz is the length of a section of the line, and $\hat{L}$ and $\hat{C}$ are the inductance and capacitance per unit length. In writing the partial differential equations that appear on the right of Eqs. (1) and (2), the limit of $\Delta z \to \partial z$ has been taken. The variables $V(z,t)$ and $I(z,t)$ depend on both

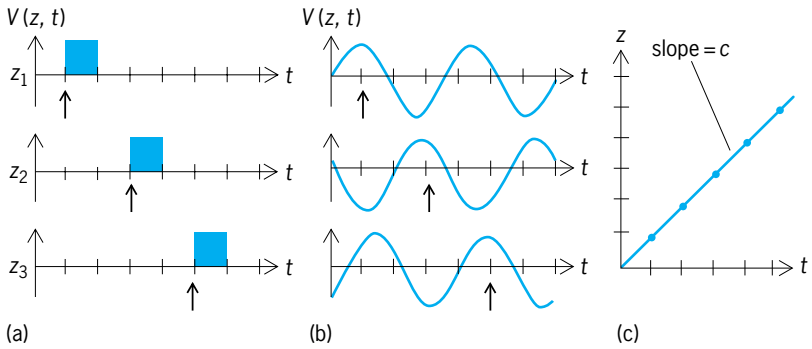


Fig. 3. Signal detection on an oscilloscope. (*a*) Oscilloscope pictures of voltage pulses at three locations, z_1, z_2, and z_3, on a transmission line, successively more distant from the signal generator. (*b*) Oscilloscope pictures of a periodic voltage sine wave at the same three locations on the transmission line. (*c*) Trajectory of propagating wave. The detecting probe position, z_j, is plotted as a function of the time of flight, t_j (indicated with arrows in parts *a* and *b*), of the wave. The slope of the line equals the velocity of propagation, *c*.

of the independent variables, position z and time t. This set of two linear coupled first-order partial differential equations is frequently called the telegraphist's equations.

Wave equation. Eliminating one of the dependent variables, say, the current $I(z,t)$, between Eqs. (1) and (2) results in a second-order partial differential equation (3), where $c = 1/\sqrt{\hat{L}\hat{C}}$. Equation (3) is a

$$\frac{\partial^2 V(z,t)}{\partial z^2} - \frac{1}{c^2}\frac{\partial^2 V(z,t)}{\partial t^2} = 0 \qquad (3)$$

standard form of the wave equation, and the parameter c is the velocity of propagation of the wave. This velocity equals the velocity of light in the dielectric that separates the two conductors. The most general solution for Eq. (3) is given by Eq. (4), where

$$V(z,t) = V_1(z - ct) + V_2(z + ct) \qquad (4)$$

$V_1(z - ct)$ is a wave that propagates to increasing values of the coordinate z, and $V_2(z + ct)$ is a wave that propagates to decreasing values of the coordinate z. This solution is predicated on the fact that time is increasing and that the functions V_1 and V_2 cannot be determined and sketched unless a numerical value appears within the parentheses. The actual shape of the propagating voltage wave which will be the same as the propagating current wave is determined by the excitation signal and the particular value of the number that appears within the parentheses. The signal could have the shape, for example, of a narrow pulse, of a step, or of a periodic sine wave. *See* WAVE EQUATION; WAVE MOTION.

Signal propagation. In a typical experimental setup, a signal generator is connected to one end of a transmission line. A trigger signal is available from the signal generator, and it is used to trigger an oscilloscope. The trigger signal causes the trace on the oscilloscope screen to have a known temporal relationship with the excited repetitive pulses from the signal generator or the phase of a periodic sine wave from the generator. The propagating signal is detected at various locations along the transmission line. Because of the finite velocity of propagation, c, there will be a nonzero time delay for the signal to propagate from one node to the next. *See* OSCILLOSCOPE.

If the signal generator excites a pulse, then a pulse will propagate on the transmission line (**Fig. 3***a*). As the distance from the signal generator increases, the response on the oscilloscope will appear later in time. If the signal generator excites a periodic sine wave signal, then the propagating wave will be a sine wave (Fig. 3*b*). From such experimental data, it is possible to obtain the velocity of propagation of the wave. This can be done by following the pulse or a point of constant phase of the periodic sine wave. By knowing the position and the time of flight, the trajectory of the wave's propagation can be given (Fig. 3*c*). The slope of the line is the velocity of propagation. In the case of the sinusoidal wave excitation, this velocity is called the phase velocity. *See* PHASE VELOCITY; WAVEFORM.

Equivalent signals for the current wave exist since the two components of the wave are coupled together via Eqs. (1) and (2). Both components of the wave have the same form; that is, a voltage pulse yields a current pulse, a voltage sine wave yields a current sine wave, and so forth. Once the voltage is determined, the current is also known. It is possible to launch a wave so that it will propagate in only one direction or in both directions.

Characteristic impedance. The ratio of the voltage $V(z,t)$ at a location z divided by the current $I(z,t)$ at the same location of the propagating wave is called the characteristic impedance of the transmission line. This parameter, $Z_c = \sqrt{\hat{L}\hat{C}}$, and it specifies several properties of the line. If the losses in the transmission line can be neglected, the characteristic impedance is a real number that corresponds to a resistance. The characteristic impedance of a transmission line depends upon the dielectric constant of the material inserted between the two conductors and the physical dimensions of the trasmission line. The transmission line is specified by its characteristic impedance, and the detailed equivalent circuit elements need not further be employed. A typical value for the characteristic impedance of a coaxial cable is $Z_c = 50\ \Omega$ and of two parallel wires, termed a twin lead, is $Z_c = 4$ or $8\ \Omega$. *See* ELECTRICAL IMPEDANCE.

Termination. A transmission line is terminated at one end with either a load impedance or another transmission line that may have a different characteristic impedance (**Fig. 4**). It is convenient to assume

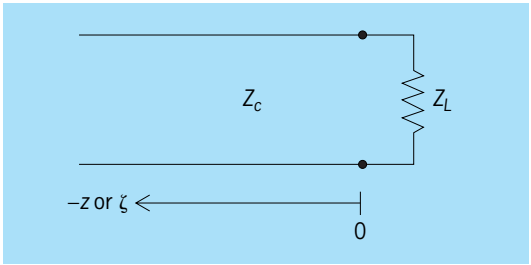


Fig. 4. Transmission line with characteristic impedance Z_c, terminated in a load impedance Z_L. A change of coordinate system (from z to ζ) is indicated.

that the excitation signal is located at $z = -\infty$, to define the location of the load impedance to be at $z = 0$, and to specify a time-harmonic wave excitation of the form $e^{+j\omega t}$. It is also common to redefine the coordinate system at this stage and assume that the coordinate $\zeta = -z$ increases to the left (Fig. 4). The voltage and current can be obtained from Eq. (1) as Eqs. (5) and (6), where $k = 2\pi/\lambda$ and λ are

$$V(\zeta,t) = Ae^{j(\omega t+k\zeta)} + Be^{j(\omega t-k\zeta)} \tag{5}$$

$$I(\zeta,t) = \frac{1}{Z_c}\left\{Ae^{j(\omega t+k\zeta)} - Be^{j(\omega t-k\zeta)}\right\} \tag{6}$$

the wave number and wavelength, respectively, A is the amplitude of the wave that propagates to increasing values of z (decreasing values of ζ), and B is the amplitude of the wave that propagates to decreasing values of z (increasing values of ζ). In this example, A would correspond to the amplitude of the wave launched from the signal generator and B would be the amplitude of the wave that is reflected by the impedance Z_L. The ratio of the voltage wave divided by the current wave, evaluated at $z = 0$, equals the load impedance Z_L. The choice of placing the load impedance at $z = \zeta = 0$ causes the exponential terms in Eqs. (5) and (6) to cancel, yielding Eq. (7).

$$Z_L = Z_c\frac{1+\dfrac{B}{A}}{1-\dfrac{B}{A}} \tag{7}$$

or

$$\mathscr{R} = \frac{B}{A} = \frac{Z_L - Z_c}{Z_L + Z_c}$$

The symbol $\mathscr{R}$ is called the reflection coefficient, and it can have values between -1 and $+1$. If $Z_L = Z_c$, then $\mathscr{R} = 0$; in that case, the transmission line is said to be matched and the reflected component is equal to zero. If $Z_L = 0$ (short circuit), then $\mathscr{R} = -1$; or if $Z_L = \infty$ (open circuit), then $\mathscr{R} = +1$. These values imply that a reflected pulse would have the opposite or the same polarity as the incident pulse. *See* REFLECTION AND TRANSMISSION COEFFICIENTS.

Junctions. If two transmission lines with different characteristic impedances are connected together, and if a voltage signal is incident upon the junction from the first line, a portion of this signal will be transmitted into the second line and a portion will be reflected back into the first line at the junction. The reflection coefficient is found from Eq. (7), where the load impedance is interpreted to be the characteristic impedance of the second line.

Standing waves. The sum of the incident and the reflected waves for the time-harmonic voltage can be written as Eq. (8), where the time dependence has

$$V(\zeta,t) = Ae^{j\omega t}[e^{jk\zeta} + \mathscr{R}\,e^{-jk\zeta}] \tag{8}$$

been separated from the other two terms. The sum of the two components within the square braces is called a standing wave. As noted above, if $Z_L = 0$ (short circuit), $\mathscr{R} = -1$, or if $Z_L = \infty$ (open circuit), $\mathscr{R} = +1$; in these two cases the voltage standing wave assumes a sinusoidal or cosinusoidal spatial variation. In general, the amplitude of the standing wave oscillates in time but is confined within an envelope (**Fig. 5**). The ratio of the maximum voltage to the minimum voltage of the standing wave, called the voltage standing-wave ratio (VSWR), is an easily measured parameter, given by Eq. (9).

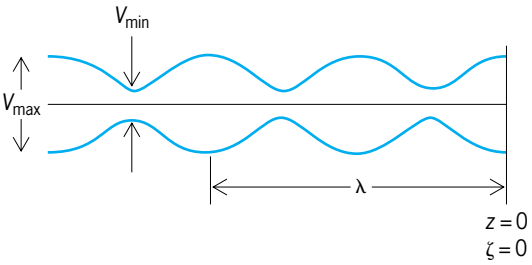


Fig. 5. Envelope of the voltage standing wave consisting of a wave propagating from $z = -\infty$ and being reflected at $z = 0$. The wavelength, λ, and the maximum and minimum voltages, V_{max} and V_{min}, are indicated.

$$\text{VSWR} = \frac{V_{\max}}{V_{\min}} = \frac{1+|\mathscr{R}|}{1-|\mathscr{R}|} \tag{9}$$

In practice, it is desirable to have the VSWR be as close to unity as possible. This implies that $\mathscr{R} \approx 0$ and the load impedance should be matched to the characteristic impedance of the transmission line, that is, $Z_L \approx Z_c$. Equation (9) can be solved for the reflection coefficient $|\mathscr{R}|$ in terms of the VSWR. Laboratory measurements of a standing voltage wave at several locations in space produce data from which the VSWR, hence from Eq. (9), $|\mathscr{R}|$ can be determined. The wavelength can be measured, and from the known value of the frequency, the phase velocity can be computed. The value of the wavelength and the location of the first minima yield the value of the normalized load impedance (Z_L/Z_c) by using Eq. (7). This allows the value of a complex load impedance to be computed from Eq. (7) since Z_c is known in practice. *See* WAVELENGTH MEASUREMENT.

Impedance matching. There are several techniques that can be employed to match transmission lines. These include the addition of circuit elements or sections of shorted transmission lines at critical locations in the main transmission line. Hence, the terminating end which consists of the actual load impedance and the additional elements will be matched to a microwave transmission line over a significant length of the line. The Smith chart is frequently used to provide graphical visualization of matching techniques. *See* IMPEDANCE MATCHING.

Fault location. Monitoring voltage pulse propagation, both the incident and the reflected pulses, can yield similar results. In addition, the measurement of the time of flight of a signal as it propagates along the transmission line and is reflected from a discontinuity can be used to find the location of a fault in a transmission line if the velocity of propagation is previously known. The technique is called time-domain reflectometry. This is particularly useful for finding faults in buried cables.

Attenuation. If the losses due to the distributed series resistance or shunt conductance terms are included in the model, it is found that the voltage and current signals attenuate as they propagate. This attenuation is exponential. In addition, the characteristic impedance becomes complex. *See* ATTENUATION (ELECTRICITY).

Waveguides. Up to this point, transmission lines have been described that could be modeled with distributed circuit elements. The derivation of the telegraphist's equations, (1) and (2), and the resulting wave equation, (3), directly followed. There are, however, several cases where it is better to directly start from Maxwell's equations in order to ascertain the characteristics of the electromagnetic waves that propagate in a certain direction. These waves are constrained or guided to propagate in a certain direction in hollow metal tubes called waveguides or a long dielectric slabs or optical fibers called dielectric waveguides. Transmission-line models for these structures do exist; hence all of the material that has been described already is applicable here.

The electromagnetic fields are determined from Maxwell's equations. In a vacuum, the wave equation (10) can be derived from these equations, where

$$\nabla^2\mathbf{E} - \frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} = 0 \tag{10}$$

$\mathbf{E}$ is the electric field, c is the speed of light, and ∇^2 is the laplacian operator. If the wave propagates in a waveguide (**Fig. 6**), in the direction of the z axis of a coordinate system, as given in Eq. (11), then Eq. (10) can be rewritten as Eq. (12). This is a vector wave

$$\mathbf{E}(x,y,z,t) = \mathbf{E}(x,y)e^{j(\omega t-\beta z)} \tag{11}$$

$$\nabla_t^2\mathbf{E} + \left(\frac{\omega^2}{c^2} - \beta^2\right)\mathbf{E} = 0 \tag{12}$$

equation in that the polarization of the electric field $\mathbf{E}$ is at this stage arbitrary. The notation ∇_t^2 indicates differentiation with respect to coordinates that are in the transverse plane, perpendicular to the direction of propagation. The wave equation (10) is similar to Eq. (3), and the description of waves with reference to transmission lines is also applicable here. *See* CALCULUS OF VECTORS; LAPLACIAN; MAXWELL'S EQUATIONS.

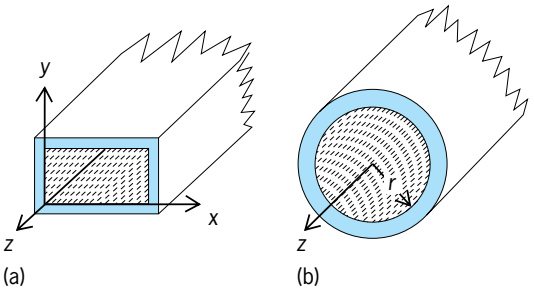


Fig. 6. Two waveguides, with conventional coordinates. (*a*) Rectangular waveguide. (*b*) Cylindrical waveguide.

Modes of propagation. Associated with the solution of Eq. (12), are boundary conditions that must be applied. Because of the high conductivity of the walls, the applicable boundary condition is that the component of the electric field that is tangent to the wall at the wall must be equal to zero. If the electric field in the transverse plane for a rectangular waveguide (Fig. 6*a*) is polarized in the y direction and is independent of the y coordinate, the solution of Eq. (12) is one in which the field depends sinusoidally on the x coordinate. Moreover, the field has an integral number of half-periods across the waveguide and has nodes at both walls. In this particular case, there will be no component of electric field in the direction of propagation, and this field configuration is called a transverse electric (TE) mode. Two integers m and n are further used to specify the mode of propagation within the waveguide. (In the case just discussed, $m = 0$, and n is the number of half-periods of the field across the waveguide.) The lowest-order mode is the $\mathrm{TE}_{nm} = \mathrm{TE}_{10}$ mode, with $n = 1$ and $m = 0$. Another family of transverse magnetic (TM) modes also exists. For a cylindrical waveguide (Fig. 6*b*), the radial distribution of the field will be in terms of Bessel functions. *See* BESSEL FUNCTIONS.

Dispersion curve. If the solution discussed above is substituted into Eq. (12), an algebraic relation, Eq. (13) is obtained that must be satisfied, where

$$\beta^2 = \frac{\omega^2}{c^2} - \left(\frac{n\pi}{a}\right)^2 \tag{13}$$

n is the number of half-periods of the field across the waveguide and a is the broad dimension of the waveguide. In order that the wave be able to propagate, the propagation constant β must be a real number. This implies that the frequency, $f = \omega/(2\pi)$, must be higher than a certain value $f_c = (nc/2a)$. This frequency, f_c, is called the cutoff frequency. When the frequency is reduced to this value, the wavelength of the wave equals $2a$ and the wave just bounces back and forth in the transverse direction with no energy propagating down the waveguide. There is a nonlinear relation between the propagation constant, β, and the angular frequency, ω. This relationship between β and ω is called a dispersion curve (**Fig. 7**). The propagation constant β asymptotically approaches a straight line as the angular frequency, ω, increases; the slope of the line equals the velocity of light, c.

Phase and group velocities. If two waves simultaneously propagate in the same waveguide and their frequencies are slightly above the cutoff frequency, there will be constructive and destructive interference between them. In the case that the excited signal is a pulse, the signal may become distorted as it propagates since individual frequency components of the pulse propagate with different velocities. This is called dispersion. A point of constant phase of an individual sine wave will propagate with the phase velocity, and the envelope of the wave will propagate with a different velocity that is called the group velocity. *See* DISTORTION (ELECTRONIC CIRCUITS).

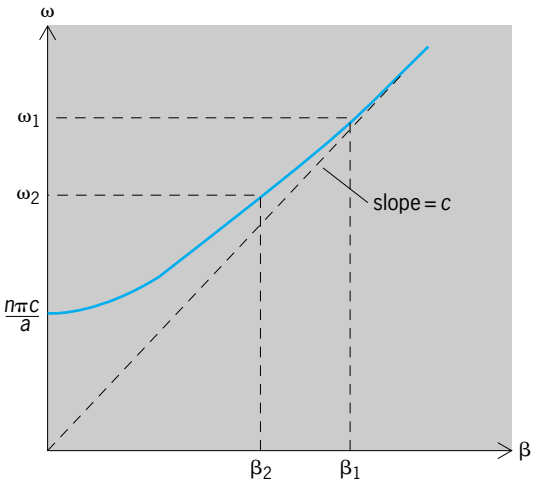


Fig. 7. Dispersion curve for a rectangular waveguide of width *a*, showing relationship between the propagation constant, β, and the angular frequency, ω. Here, *n* specifies the mode of propagation, and *c* is the speed of light.

The derivation of the group velocity can be carried out in a straightforward manner by considering two waves with slightly different frequencies, $\omega_1 \approx \omega_0 + \Delta\omega$ and $\omega_2 \approx \omega_0 - \Delta\omega$ (Fig. 7), propagating in the waveguide, where the intermediate frequency ω_0 is slightly above the cutoff frequency of a particular mode in the waveguide. There will be a unique propagation constant $\beta_1 = \beta_0 + \Delta\beta$ and $\beta_2 = \beta_0 - \Delta\beta$ associated with each frequency. The total electric field in the waveguide is just the linear superposition of the two electric field components in the waveguide. An individual wave will propagate with the phase velocity $v_\phi = \omega/\beta$. The amplitude modulation will propagate with a velocity that is called the group velocity $v_g = \Delta\omega/\Delta\beta$, which becomes $v_g = \partial\omega/\partial\beta$ in the limit that the frequency difference, $\Delta\omega$, approaches zero. In nondispersive media such as a vacuum, the two velocities are the same. They are, however, dramatically different in a waveguide. *See* GROUP VELOCITY.

Attenuation reduction. There are several techniques to match and reduce attenuation in hollow metal waveguides. Screws are frequently inserted in the center of a waveguide, or thin metal strips are placed across it. Losses are reduced by plating the inside of the waveguide with silver or gold, whose conductivity may be higher than the host waveguide.

Dielectric waveguides. Electromagnetic waves can also be guided at an interface between two dielectrics. The amplitude of the fields will exponentially decay to zero in the region external to the dielectric. The dispersion relation for this structure is determined by requiring that the tangential components of the electric and the magnetic fields be continuous at the interface. In certain materials, the inhomogeneous electric field may locally modify the dielectric and permit nonlinear waves called solitons to propagate in the dielectric. *See* ELECTROMAGNETIC WAVE TRANSMISSION; OPTICAL FIBERS; WAVEGUIDE.

Karl E. Lonngren

Power transmission lines. Electric power generating stations and distribution substations are connected by a network of power transmission lines, mostly overhead lines. Power transmitted is generally in the form of three-phase alternating current (ac) at 60 or 50 Hz. In a few instances, where a clear technical or economic advantage exists, direct-current (dc) systems may be used. As the distances over which the power must be transmitted become great and as the amount of power transmitted increases, the power lost in the transmission lines becomes an important component of the production cost of electricity, and it becomes advantageous to increase the transmission voltage. This basic consideration has led to electric power networks which use higher voltages for long-distance bulk power transfers, with several layers of underlying regional networks at progressively lower voltages which extend over shorter distances. The most common transmission voltages in use are 765, 500, 400, 220 kV, and so forth. Voltages below 69 kV are termed subtransmission or distribution voltages, and at these and lower voltages the networks may have fewer alternative supply paths (loops) or may be entirely radial in structure. *See* ALTERNATING CURRENT; DIRECT CURRENT; DIRECT-CURRENT TRANSMISSION; ELECTRIC DISTRIBUTION SYSTEMS.

Performance during normal conditions. The three phases of power transmission lines generally carry balanced voltages and currents; that is, the voltages (and currents) of the three phases are of equal magnitude and differ in phase angles by 120°. This balance is achieved by balancing the loads among the three phases at the distribution level, and by making sure that the impedances of the transmission lines, generators, and transformers are equal in all three phases. In the case of overhead transmission lines, phase impedance balancing can be achieved by transposing the phases, so that the relative position of each of the phases on the tower between the two ends of the transmission line averages out to be identical. Transposition of phases requires special towers which can accommodate the additional mechanical stresses, and consequently is often omitted at the highest voltage levels in the interest of economy.

Under balanced conditions, it is sufficient to analyze a single-phase circuit and assume that the performance of the other two phases is identical. For the purposes of engineering analysis of a power system in normal steady state, the equivalent circuit of each phase of a power transmission line is a π circuit (**Fig. 8**). The resistance and reactance (at the operational frequency), and the capacitance of the entire transmission line can be estimated from corresponding values per mile and then multiplying by the length of the line. This is the nominal-π representation, and is sufficiently accurate for lines of up to 100 mi (160 km) in length. For longer lines, an exact-π model, which uses the exact transmission-line equations in its derivation, is needed. For lines of several hundred miles, the difference between the nominal and exact π circuits is of the order of 5–10%. The resistance is quite small compared to the reactance, and for short lines the capacitance may be omitted. Transmission-line equivalent circuits

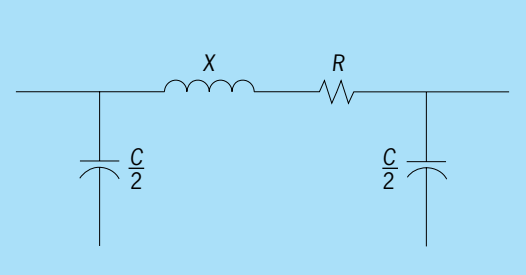


Fig. 8. Equivalent circuit of a power transmission line. The single-phase circuit is used to represent balanced three-phase lines. Here, *R* and *X* are the resistance and reactance (at the operational frequency), and *C* is the capacitance of the entire transmission line.

(Fig. 8) are used in calculation of power flows, voltage drops, power losses in transmission networks, electromechanical oscillations of connected generators and loads, and so forth. *See* REACTANCE.

The power-handling capability (transmission capacity) of a transmission line depends on many factors. The thermal capacity of the transmission-line conductors places a fundamental limit on power transfer capability. In many transmission lines, the power transfer capability is determined by a different phenomenon: transient stability. A power system must recover its normal operation after a fault is cleared by the protection system, and new power flow patterns are established in the remaining network. The synchronous generators supplying the network may go unstable if the transmission line which suffered the fault was carrying a large amount of power, or if the fault persisted for too long a duration. If the power carried by the transmission line is below a certain limit, the fault and its clearing would not lead to instability. The power transfer limit determined in this manner is known as the transient stability limit. Making reasonable assumptions about a power system, the transient stability limit of transmission lines can be shown to be a fairly well defined parameter.

Faults and protection systems. Faults on transmission lines are caused by short circuits between the phases, or between phases and the Earth. Faults may be caused by overvoltages caused by lightning, by the transients set up during swiching operations, or by equipment failure. The faulted line must be removed from service promptly, so that the remaining system can resume service normally. The fault removal is accomplished by the protection system installed at the two ends of each transmission line. It consists of circuit breakers, relays, and current and voltage transformers (transducers). Upon the occurrence of a fault, the relays, acting upon the changed conditions reflected in the outputs of the current and voltage transformers, make a determination if the fault is in the zone of protection for which the protection system is responsible. If the fault is within the zone, the relays will trip the circuit breakers, which will isolate the transmission line at its terminals from the rest of the network. At the highest operating voltages, the entire protection action may be completed in about 30–50 milliseconds after the occurrence of the fault. In most cases, the faults on the transmission lines are temporary, and the protection system will automatically attempt to reenergize the transmission line in order to test if the fault has disappeared. About 80% of the faults lead to successful reclosing, and are never perceived by the end user of electricity. *See* CIRCUIT BREAKER; ELECTRIC PROTECTIVE DEVICES; INSTRUMENT TRANSFORMER; RELAY.

Compensation. There are three types of compensation systems currently used in ac transmission systems: series capacitors, shunt reactors, and shunt capacitors. The power transfer capability limit imposed by transient stability can be relaxed if the series reactance X of the transmission line can be reduced. The obvious method for reducing the reactance is to put a capacitor in series with the transmission line. If the reactance of the capacitor (at the operating frequency) is X_c, the net reactance of the transmission line is reduced from X to $(X - X_c)$. The improvement in the power transfer capability of a transmission line is (approximately) inversely proportional to the reactance of the line. The capacitors used in this fashion are known as series compensation, and may be of a fixed size, or be switched in sections as needed, or could be made variable by using thyristors as control elements. *See* SEMICONDUCTOR RECTIFIER; STATIC VAR COMPENSATOR.

On long transmission lines, the capacitors in the equivalent circuit of the π-section representation lead to a voltage rise at the receiving end if the load on the transmission lines is dropped, or is reduced, for example at night. The voltage rise at light loads or no loads presents a serious operational problem to the connected equipment. This voltage rise can be reduced by connecting shunt reactors across the transmission line; these compensate the shunt capacitors of the π circuit. And finally, under heavy loading conditions, the drop in voltgage at the receiving end due to the line impedance may also be unacceptable. This can be corrected by using shunt capacitors across the transmission lines. *See* REACTOR (ELECTRICITY).

Environment. Power transmission lines are the most visible components of the electric power infrastructure. They impact the environment in several ways. The most obvious is the visual impact. Designs have been introduced to minimize the visual impact of transmission towers, and they are sited to avoid high ground or prominent ridges.

Transmission-line conductors have high electric field gradients at their surfaces, and under certain conditions will ionize and break down the surrounding air, causing what is generally referred to as corona discharge. Corona is a function of moisture and pollution in the air, as well as the smoothness of the conductor surface. It contributes to radio and television interference, and should be minimal in a well-designed and -maintained line. (Corona is often visible at night as a glow surrounding the power-line conductors.) *See* CORONA DISCHARGE.

Power-line currents may interfere with nearby telephone circuits, especially if there is

unbalance in the currents in the three phases. If significant ground currents exist under normal (or fault) conditions, they may flow through buried metallic objects in the ground, such as control cables and gas pipe lines. If ground currents persist, they may lead to galvanic corrosion of the metal structures. *See* COMMUNICATIONS SYSTEMS PROTECTION; CORROSION; INDUCTIVE COORDINATION.

Electric and magnetic fields exist in the neighborhood of power transmission lines. Although there has been considerable discussion of effects of power-frequency magnetic fields on human and animal health, no significant evidence for these effects has been found after a number of careful studies. *See* ELECTRIC POWER SYSTEMS; ELECTRIC POWER TRANSMISSION.

Arun G. Phadke

Bibliography. D. K. Cheng, *Field and Wave Electromagnetics*, 2d ed., 1989; R. E. Collin, *Foundations for Microwave Engineering*, 2d ed., 2000; A. Hirose and K. E. Lonngren, *Introduction to Wave Phenomena*, 1991, reprint 2003; S. Liao, *Microwave Devices and Circuits*, 3d ed., 1996; R. L. Liboff and G. C. Dalman, *Transmission Lines, Waveguides, and Smith Charts*, 1985; W. D. Stevenson, Jr., *Elements of Power System Analysis*, 4th ed., 1982.

Transmutation

The nuclear change of one element into another, either naturally, in radioactive elements, or artificially, by bombarding an element with electrons, deuterons, or alpha particles in particle accelerators or with neutrons in atomic piles.

Natural transmutation was first explained by Marie Curie about 1900 as the result of the decay of radioactive elements into others of lower atomic weight. Ernest Rutherford produced the first artificial transmutation (nitrogen into oxygen and hydrogen) in 1919. Artificial transmutation is the method of origin of the heavier, artificial transuranium elements, and also of hundreds of radioactive isotopes of most of the chemical elements in the periodic table. Practically all of these elements also have been artificially transmuted into neighboring elements under experimental conditions. *See* NUCLEAR REACTION; PERIODIC TABLE.

Frank H. Rockett

Transonic flight

In aerodynamics, flight of a vehicle at speeds near the speed of sound. When the velocity of an airplane approaches the speed of sound, roughly 660 mi/h (1060 km/h) at 35,000 ft (11 km) altitude, the flight characteristics become radically different from those at subsonic speeds. The drag increases greatly, the lift at a given attitude decreases, the moments acting on the airplane change abruptly, and the vehicle may shake or buffet. Such phenomena usually persist for flight velocities somewhat above the speed of sound. These flight characteristics, as well as the speeds at which they occur, are usually referred to as transonic. For configurations designed for subsonic flight these changes may occur at velocities of 70–110% of the speed of sound (Mach numbers of 0.7–1.1); for airplanes intended for transonic or supersonic flight they may be present only at Mach numbers of 0.95–1.05. *See* FLIGHT CHARACTERISTICS; MACH NUMBER.

Shock waves. The transonic flight characteristics result from the development of shock waves about the airplane. Because of the accelerations of airflow over the various surfaces, the local velocities become supersonic while the airplane itself is still subsonic. (The flight speed at which such local supersonic flows first occur is called the critical speed.) Shock waves are associated with deceleration of these local supersonic flows to subsonic flight velocities (**Fig. 1**). Such shock waves cause abrupt streamwise increases of pressure on the airplane surfaces. These gradients may cause a reversal and separation of the flow in the boundary layer on the wing surface in roughly the same manner as do similar pressure changes at subcritical speeds. When the wing carries lift, the shock-induced separation is particularly strong on the upper surfaces. *See* AERODYNAMIC FORCE; AERODYNAMIC WAVE DRAG; BOUNDARY-LAYER FLOW; SCHLIEREN PHOTOGRAPHY; SHOCK WAVE.

As for boundary-layer separation at lower speeds, the flow breakdown in this case leads to increases of drag, losses of lift, and changes of aerodynamic moments. The unsteady nature of the separated flow results in an irregular change of the aerodynamic forces acting on the airplane with resultant buffeting and shaking. As the Mach number is increased, the shock waves move aft so that at Mach numbers

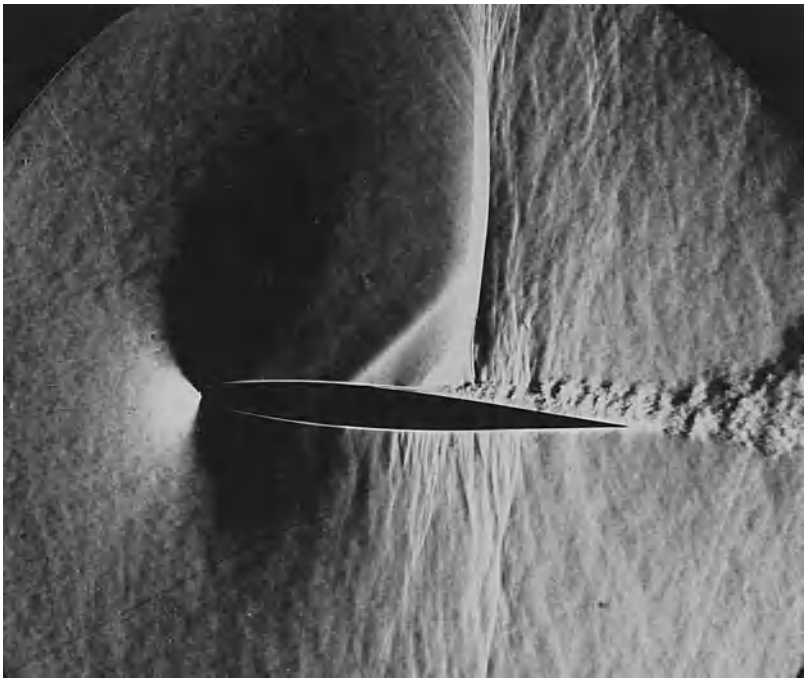


Fig. 1. Schlieren photograph of flow about airfoil section at low transonic speeds. Shock wave is the nearly vertical line near midchord.

of about 1.0 or greater they reach the trailing edges of the surfaces. With the shocks in these positions, the associated pressure gradients have relatively little effect on the boundary layer, and the shock-induced separation is greatly reduced.

Effects on flight characteristics. When the speed is increased to the higher transonic range, at and just above the speed of sound, the energy losses in the shock waves about an airplane may become large. As a result, the drag may increase to many times the subsonic value. At these speeds shock waves, in addition to those present near the aft parts of the surfaces, form ahead of the components. The various waves extend outward, interact, and merge to form two shock waves at a distance from the airplane. These two waves are relatively strong and extensive; they may extend outward for miles from the airplane. *See* SONIC BOOM.

As the speed is increased through the transonic range, the changes of the distribution of load on the wing, resulting first from boundary-layer separation and then from rearward movement of the shock wave, cause a marked rearward shift of the center of lift. This shift causes a nose-downward moment on the airplane that must be corrected by an increase of the negative lift on the usual tail to maintain trim. Also, the effectiveness of the usual flap-type elevator and aileron control surfaces used on subsonic airplanes decreases greatly at transonic speeds. Deflections of such flaps provide differences in the pressures on upper and lower parts of the main surface ahead of flap, as well as on the flap itself. At transonic speeds, the effect of the flap on the pressures on the main surface is greatly reduced because of the presence of local supersonic flows on this surface. In addition, the hinge moments required to deflect the control may be greatly increased at transonic speeds. *See* AILERON; ELEVATOR (AIRCRAFT); FLIGHT CONTROLS.

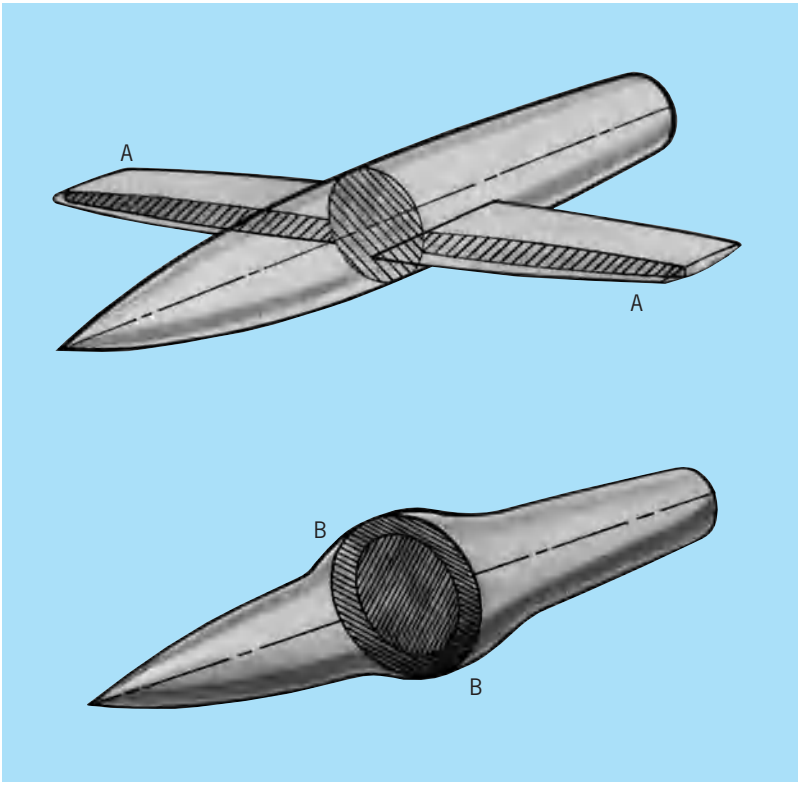


Fig. 2. Area rule comparison. The various normal cross-sectional areas for the body of revolution, such as at BB, are the same as those for the wing-fuselage combination at the corresponding longitudinal station, such as at AA. Therefore, the shock wave and resulting drag near the speed of sound are approximately the same for the two configurations.

Corrective means. Various means may be used to delay and reduce the adverse transonic characteristics.

Sweepback. The most effective means for improving the overall transonic characteristics is to mount the wings slanting backward. The action of such sweep may be understood by considering the airflow over a very long swept surface. Only the component of airflow normal to swept elements of this panel is effective in determining the nature of flow over the surface. Thus on such a swept surface, the onset of a shock wave, with the associated separation, is delayed until the reduced component of local velocity normal to the swept elements becomes supersonic. The use of sweep also greatly reduces the magnitude of the changes in the aerodynamic characteristics, once they occur. *See* WING.

The transonic characteristics are progressively delayed and reduced by increasing the sweep to relatively high values, but excessive sweep leads to a number of aerodynamic problems. Most significantly, a highly swept wing may have an abrupt nose-up moment at the higher lifts. This phenomenon, known as pitch-up, may result in excessive aerodynamic loads of stall. Pitch-up results from an initial separation of the boundary layer on the outboard part of the swept wing with an associated loss of lift for this region. Because this portion of a swept wing is aft of the center of gravity, the loss of lift causes a nose-up moment. Increasing the wing sweep also reduces the lift available for takeoff and landing.

Because of these limitations, most transport-type airplanes designed to fly at transonic speeds incorporate only moderate wing sweep. Usually the obliqueness of the midchord element for such airplanes is about 30°. Transonic and supersonic military airplanes may incorporate as much as 45° of midchord sweep.

According to simple sweep theory, sweepforward is as effective as sweepback in delaying the onset of adverse effects. However, a sweptforward wing would be structurally divergent; that is, a bending of the wing due to air loads would result in increases of the angle of attack of the outboard sections of the wing with resulting increases of the loads on these sections, which, of course, would then result in further bending.

Reduced thickness ratio. Substantial improvements in the adverse transonic characteristics are provided by reducing the thickness-to-chord ratios for the wing and tail surfaces. Such changes reduce the acceleration of the flow over the surfaces with a resulting delay in the onset of local supersonic flows and the associated shock wave. Also, the severity of the adverse longitudinal pressure gradients on the wing

surface is lessened so that boundary-layer separation is reduced. However, reductions of thickness are limited by the considerable increases in the weight of a structurally sound wing. Most high-speed transport wings have mean thickness-to-chord ratios of about 10%, whereas military airplanes may have thickness ratios as low as 3%.

Reduced aspect ratio. Reductions in the aspect ratio provide delays and reductions of the transonic changes similar to those provided by reductions in thickness ratio, although the magnitude of the effect is usually considerably less. More important, lower aspect ratios result in improvements of the wing structural characteristics, which allow the use of smaller thickness-to-chord ratios. However, the use of reduced aspect ratios increases the subcritical drag due to lift. Most high-speed transport wings have aspect ratios of about 7; transonic and supersonic military airplanes may have aspect ratios as low as 2.

Supercritical wing. The onset of the adverse transonic characteristics can be delayed substantially beyond the critical by the use of special speed streamwise section shapes (airfoils) for the wing. These shapes are called supercritical, and wings incorporating them are referred to as supercritical wings. *See* SUPERCRITICAL WING.

Area rule. For a nonlifting condition, the forms of the shock waves and consequently the drag are primarily a function of the longitudinal development of cross-sectional area, in section normal to the airstream, for the complete airplane. According to this relationship, called the area rule, the shock wave and the resulting drag near the speed of sound are approximately the same for two configurations that have the same total normal cross-sectional areas at corresponding longitudinal stations, even though their shapes may be very different (**Fig. 2**).

On the basis of the area rule, the transonic drag increment is reduced by shaping and arranging the airplane components so that area development for the airplane more nearly approaches the shape with the lowest drag (**Fig. 3**). The magnitude of the drag associated with such a shape is greatly reduced by increasing the overall length and reducing the maximum cross-sectional area. However, for practical reasons, the lengths and cross-sectional areas of airplanes must be limited to values corresponding to a body of revolution with a ratio of length to diameter of about 9.

The longitudinal developments of area for conventional subsonic airplanes differ greatly from the ideal shape, and as a result, the maximum transonic drag for such airplanes may be as much as 10 times subsonic drag values. The various wing features discussed above usually result in airplane area developments more nearly approaching the most satisfactory shape (Fig. 3), reducing the transonic drag to approximately three times the subsonic level.

The area developments for some transonic and supersonic airplanes have been made to approach the form for lowest drag by special shaping of the fuselage, reducing the transonic drag to as low as twice

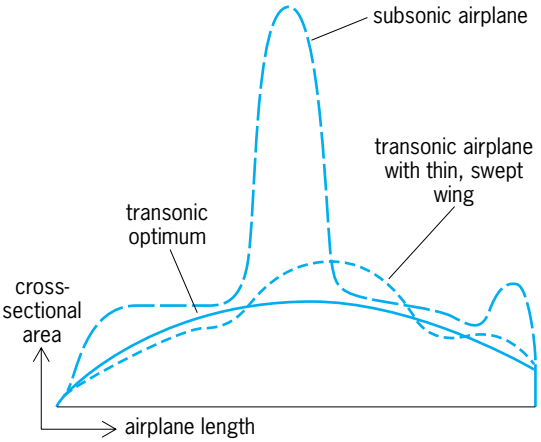


Fig. 3. Longitudinal area developments for various types of airplanes, designed to reduce drag.

the subsonic values. Such a shaping has been provided through the subtraction of fuselage volume in the region of the wing and tail, as well as by the addition of volume ahead of and behind these surfaces. The area development may also be improved by specially locating external bodies such as engine nacelles. *See* AIRPLANE; SUBSONIC FLIGHT; SUPERSONIC FLIGHT.

Richard T. Whitcomb

Bibliography. J. D. Anderson, *Fundamentals of Aerodynamics,* 3d ed., 2001; J. D. Cole and L. P. Cook, *Transonic Aerodynamics,* 1986; T. H. Moulden, *Fundamentals of Transonic Flow,* 1984, reprint 1991; H. J. Ramm, *Fluid Dynamics for the Study of Transonic Flow,* 1990.

Transplantation biology

The science of transferring a graft from one part of the body to another or from one individual, the donor, to another, the recipient. The graft may consist of an organ, tissue, or cells. If donor and recipient are the same individual, the graft is autologous. If donor and recipient are genetically identical (monozygotic), it is syngeneic. If donor and recipient are any other same-species individuals, the graft is allogeneic. If the donor and recipient are of different species, it is called xenogeneic.

Use in experimental biology. The transplantation of cells, tissues, or organs is not only a medical therapy but also an important research tool. For example, a defining criterion of the hormone-secreting, or endocrine, function of a tissue is the demonstration of its ability to reverse the effects of its removal after transplantation to an anatomically abnormal, or heterotopic, site. In embryology, grafting can reveal the interactions of cells of different types that lead to organogenesis. The manner in which genes determine coat coloration has been elucidated by studies involving the grafting of melanocytes (pigment-forming cells) of one genetic constitution into hair follicles of a different one. Primary chimeras, produced by combining cells from different embryos at a very early stage of development and allowing them

to develop in the uteri of surrogate mothers, are useful in the study of problems in embryology, developmental genetics, and immunology. Age chimeras, individuals bearing a functional organ or tissue graft that is chronologically much younger or older than its host, are useful in the study of aging. Studies on such individuals, usually mice, indicate that tissues have a finite life span. In cancer research, transplantation is widely employed to propagate malignant tissues in appropriate hosts for biochemical, immunological, and therapeutic investigations. In its extreme form, "transplantation" of a somatic cell nucleus from an appropriate adult donor to an enucleated recipient oocyte, which is then implanted in a surrogate mother, has been used to generate cloned mice, sheep, and pigs. *See* CHIMERA; MOSAICISM.

Clinical applications. There is remarkable evidence that successful transplantation surgery was performed by ancient Hindu *vaidya* some 2000 years ago. These Ayurvedic physicians reconstructed noses using pedicle flap autografts from the patient's own forehead. Saints Cosmas and Damian were said to have transplanted a human leg in Roman times, but transplantation progressed little until the twentieth century due to the problem of graft rejection. Large-scale use of one type of graft, blood transfusion, was developed during World War II as methods for cross-matching donor and recipient for ABO antigens were perfected. In the second half of the twentieth century, as understanding of the biology of transplantation antigens and the immunology of rejection advanced, the grafting of replacement organs and tissues to cure disease became a clinical reality. In the United States, nearly 90,000 kidney grafts had been performed by the turn of the century. Heart, lung, liver, and pancreas grafts have also become commonplace. Successful transplantation of islets of Langerhans for the treatment of diabetes mellitus has also been achieved and may replace pancreas grafts for this purpose. Hematopoietic stem cell transplantation is used in the treatment of aplastic anemia, leukemia, and immunodeficiency diseases, and its use in the treatment of autoimmune diseases is under investigation. These stem cells can be obtained from bone marrow, peripheral blood, or umbilical cord blood.

In theory, virtually any tissue or organ can be transplanted. The principal technical problems have been defined and, in general, overcome. Remaining major problems concern the safety of methods used to prevent graft rejection and the procurement of adequate numbers of donor organs.

Procurement, storage, and types of grafts. Living volunteers can donate one of a pair of organs, such as a kidney, only one of which is necessary for normal life. Volunteer donors may also be employed for large unpaired organs such as small bowel, liver, or pancreas, segments of which can be removed without impairment of function. Living donors can also provide tissues capable of regeneration; these include blood, bone marrow, and the superficial layers of the skin. In the case of a vital, unpaired organ, such as the heart, the use of cadaver donors is obligatory.

In practice, with the exception of blood and bone marrow, the great majority of transplanted organs are cadaveric in origin, a necessity that presents difficult logistic problems. Except for skin, corneas, and possibly lungs, most organs deteriorate rapidly after death. Clinical transplantation depends on the use of "brain-dead" donors whose vital functions are maintained with the aid of a respirator until organs can be removed. To minimize autolysis, organ grafts must be removed from donors as quickly as possible, generally within 30 min of termination of cardiorespiratory support. Recovered organs must be chilled immediately by immersion in, or perfusion with, an ice-cold physiological salt solution. Under these conditions, large organ grafts have a storage life of about 24–48 h. A few tissues, among them blood, bone marrow, embryos, spermatozoa, skin, and corneas, can be stored for months if frozen to very low temperatures. When impregnated with cryoprotective agents such as glycerol or dimethyl sulfoxide and then maintained at the temperature of liquid nitrogen (−196°C or −321°F), these cryopreserved tissues retain their viability on subsequent rewarming. The prospects for similar long-term preservation of larger organ grafts are still poor.

Certain types of grafts need not be viable, even at the time of transplantation. They may be stored frozen, in a freeze-dried or lyophilized state, or in chemical preservatives for long periods. Lyophilized bone allografts, for example, are used in orthopedic surgery to provide an inert scaffolding within which regeneration of new bone takes place through the activity of cells of the recipient. Such grafts are described as allostatic (as opposed to allovital). Xenogeneic heart valves from pigs that have been chemically treated and contain no living cells are routinely used to replace diseased human valves. *See* PROSTHESIS.

Autografts are used for an increasing number of purposes. Skin autografts are important in the treatment of full-thickness skin loss due to extensive burns or other injuries. Provided that the grafts comprise only the superficial levels of the skin, the donor sites reepithelialize spontaneously within a week or two. Autografts of bone and cartilage are used in orthopedic procedures, and a relatively unimportant blood vessel, nerve, or tendon can be sacrificed to repair or replace a more important one elsewhere in the body. The saphenous vein of the ankle is frequently transplanted to the heart to bypass coronary arteries obstructed by atherosclerosis. Autologous hematopoietic stem cell transplantation is sometimes used to restore blood cells to cancer patients who receive forms of chemotherapy that are lethal to their bone marrow. Islets of Langerhans can be removed from the pancreas of a person undergoing removal of a diseased exocrine pancreas; the recovered autologous islets can then be transplanted to the donor to prevent the diabetic state that would otherwise ensue.

Healing of grafts. In the case of large organs, the main blood vessels must be coupled at operation with those of the host to conserve their viability and

enable them to function. However, thin sheets of skin and small tissues, such as the parathyroid gland and islets of Langerhans, may be transplanted heterotopically with no attempt to preserve or restore the blood supply. Juxtaposed graft and host tissue soon become knit together through the activity of fibroblasts and other cells, and revascularization takes place naturally within 2–3 days as a consequence of penetration of graft tissue by regenerating vessels. This process involves some reutilization of the original vascular channels of the graft. *See* REGENERATIVE BIOLOGY.

Transplantation immunology. The most serious problem restricting the use of allografts is immunological. Because cells in the donor graft express on their surface a number of genetically determined transplantation antigens that are not present in the recipient, allografts provoke a defensive reaction analogous to that evoked by pathogenic microorganisms. As a consequence, after a transient initial period of apparent well-being, graft function progressively deteriorates and the donor tissue is eventually destroyed. The host response, known as allograft rejection, involves a large number of immunological agents, including cytotoxic antibodies and effector lymphocytes of various types. The modus operandi of the destructive process varies as a function of the type of graft involved as well as the degree of disparity between donor and recipient transplantation antigens. For example, the hyperacute rejection of kidneys that can occur within hours following transplantation is mediated by antibodies, whereas acute rejection is a lymphocyte-mediated process. In the absence of prophylactic treatment, rejection of most allografts is usually complete within about 2 weeks.

In the case of grafts exchanged between identical human twins, or between members of the same highly inbred or syngeneic strain of experimental animals, rejection does not occur because donor and recipient are genetically identical, and consequently no antigenic differences are involved. An exception to this rule is observed when the organ being replaced in the recipient was destroyed by an autoimmune disease process. For example, islets of Langerhans transplanted from a normal twin into a monozygotic sibling suffering from type 1 autoimmune diabetes mellitus are destroyed by recurrence of autoimmunity, an acquired immunological difference between the twins. *See* AUTOIMMUNITY; DIABETES.

There are a few special exemptions from rejection that apply to certain sites in the body or to certain types of graft. For example, the use of corneal allografts in restoring sight to individuals with corneal opacification succeeds because of the absence of blood vessels in the host tissue. Allogeneic islets of Langerhans transplanted into the testis survive in the absence of any antirejection therapy through a mechanism that is not well understood. Mammalian fetuses are natural, highly successful allografts in the maternal uterus in the sense that their tissues express foreign antigens inherited from the father. The reason for their success is still not completely understood. Contributory factors include (1) the absence of continuity between maternal and fetal blood circulations in the placenta; (2) failure of the layer of fetal tissue in direct contact with maternal tissue in the placenta (the trophoblast) to express transplantation antigens in an effective manner; (3) the local operation of nonspecific immunosuppressive factors, including hormones and suppressor cells; and (4) the absence of certain co-stimulatory signals required for lymphocyte activation. *See* PLACENTATION; PREGNANCY.

Overcoming graft rejection. Successful transplantation of allografts such as kidneys and hearts currently requires suppressing the recipient's immune response to the graft without seriously impairing the immunological defense against infection. Treatment of individuals with so-called immunosuppressive drugs and other agents prevents allograft rejection for prolonged periods, if not indefinitely. Under cover of nonspecific immunosuppression, the recipient's immune system appears to undergo an adaptation to the presence of the graft, allowing the dosage of the drugs to be reduced. However, in almost all successfully transplanted individuals, drug therapy at some low dose is required indefinitely.

Unfortunately, long-term treatment with drugs that nonspecifically suppress or disrupt the immune system does interfere with normal immune function and over time can predispose to the development of infection or cancer. The ideal solution would be to abrogate specifically the body's capacity to react against the foreign antigens of the graft, leaving its reactivity to other antigens completely unimpaired. Such a solution is likely to be based on the phenomenon of immunological tolerance—the survival of foreign (allogeneic or xenogeneic) tissue in normal recipients in the absence of immunosuppression. Analyses of the mechanisms of this phenomenon in transplanted laboratory animals suggest that it is possible to induce a state of graft tolerance through a variety of different strategies. A clinically applicable solution is not yet available, but recent advances suggest that "tolerance-based" transplantation may soon become a reality. *See* ACQUIRED IMMUNOLOGICAL TOLERANCE; IMMUNOSUPPRESSION.

Immunosuppressive agents. Whole-body irradiation was one of the first forms of immunosuppression employed in transplantation, but its use is now confined largely to stem cell transplantation. Most graft recipients receive immunosuppressive drugs, which suppress the proliferation of cells involved in the immune response. One of the first of these drugs was azathioprine, introduced in 1962. It is an antimetabolite that interferes with protein synthesis. Steroids, frequently administered in high doses to prevent or treat acute rejection, have many actions, including depression of nucleic acid and protein synthesis, suppression of host inflammatory response, and killing of lymphocytes.

Newer drugs target effector lymphocytes responsible for graft rejection. Cyclosporin, tacrolimus, sirolimus, and mycophenolate mofetil have dramatically improved results with human renal transplantation. Each interferes with the proliferation of

activated T lymphocytes. Monoclonal antibodies directed against surface antigens present on activated T cells, for example the interleukin-2 receptor or the CD3 antigen, now play an increasingly important role in the initial prevention of rejection and in the treatment of later rejection crises. These monoclonal reagents lead to the destruction or inactivation of T cells to which they bind. *See* MONOCLONAL ANTIBODIES.

Tissue typing. An individual's response against an allograft is directed against a large number of cell-surface transplantation antigens controlled by allelic genes at many different loci. However, in all species, one of these loci, the major histocompatibility complex (MHC), transcends all the other histocompatibility loci (the minor loci) in terms of its genetic complexity and the strength of the antigenic response it controls. In humans, the MHC, known as the HLA (human leukocyte antigen) complex, is on the sixth chromosome; its principal loci are designated A, B, C, DR, and DQ. The allelic products of the HLA genes can be detected by serology, polymerase chain reaction technology, or microcytotoxicity assays. The ABO red cell antigens are also important because they are expressed on all tissues. *See* BLOOD GROUPS; HISTOCOMPATIBILITY.

In kidney transplants between closely related family members, the degree of HLA antigen matching can be determined very precisely, and there is a very good correlation between the number of shared HLA antigens and the survival of the graft. With grafts from unrelated donors, HLA matching is more difficult and can delay transplantation, but it may be beneficial. The estimated 10-year rate of kidney graft survival was recently reported to be 52% for HLA-matched transplants, as compared with 37% for HLA-mismatched transplants. Donors and recipients are usually matched for both alleles at the HLA-A, -B, and -DR loci. HLA matching is not as clearly beneficial in the case of most other solid organ grafts, and no attempt is made to HLA-match heart, lung, liver, and pancreas grafts. With few exceptions, however, most donors and recipients are matched for the expression of ABO blood group antigens.

The early assumption that blood transfusions should not be given to transplant recipients for fear of sensitizing them has been disproved. Prior transfusion of these patients actually increased cadaver kidney survival rates when older immunosuppressive regimens were in use, but is no longer needed since the advent of newer drugs. The combination of donor-specific transfusion (a transfusion provided by the individual who is also the organ donor) with reagents that prevent signals required for T-cell activation is under investigation as a method of tolerance-based transplantation.

Bone marrow transplantation presents a unique problem in its requirements for HLA matching and for immunosuppression in advance of grafting. In addition to the possibility of rejection of the graft by the recipient, by virtue of immunologically competent cells still present in the recipient, bone marrow grafts can react against the transplantation antigens of their hosts. These are known as graft-versus-host reactions, and they can be fatal. *See* IMMUNOLOGY.

Xenotransplantation. As success rates for allograft transplantation to cure disease have increased, the supply of donor organs has been increasingly inadequate. Cadaveric donor organs are sufficient for approximately 10% of total current need. In certain special cases, for example skin and pancreatic islets, it may become possible to genetically engineer a suitable replacement tissue. Another possible source of replacement tissue is human fetal material, but ethical and legal barriers to its use may be insurmountable. A final alternative is the use of xenografts. Pigs, for example, could supply an unlimited number of organs similar in size and function to those required by humans. The immunological events that lead to the rejection of xenografts are, however, different from and less well understood than those responsible for allograft rejection. The small number of xenografts attempted to date have failed. In particular, xenografts are susceptible to hyperacute rejection by humans. This is due to the presence of certain glycoproteins in blood vessels of many species that are recognized by antibodies present in the blood of all humans. The potential use of xenografts in humans raises special ethical and psychological issues. It also raises the specter of introducing new diseases into the human population. The transfer of pig endogenous retroviruses to cultured human cells has been demonstrated. The specter of xenozoonotic epidemics has become a concern.

John P. Mordes; Rupert E. Billingham; Dale L. Greiner; Aldo A. Rossini

Bibliography. D. K. Cooper and R. P. Lanza, *Xeno: The Promise of Transplanting Animal Organs into Humans*, Oxford University Press, 2000; E. Fuchs and J. A. Segre, Stem cells: A new lease on life, *Cell*, 100:143–155, 2000; H. Gudmundsdottir and L. A. Turka, Transplantation tolerance: Mechanisms and strategies?, *Semin. Nephrol.*, 20:209–216, 2000; N. S. Hakim (ed.), *Introduction to Organ Transplantation*, World Scientific Publishing, River Edge, NJ, 1998; F. Locatelli, D. Rondelli, and G. R. Burgio, Tolerance and hematopoietic stem cell transplantation 50 years after Burnet's theory, *Exp. Hematol.*, 28:479–489, 2000; A. M. Marmont, New horizons in the treatment of autoimmune diseases: Immunoablation and stem cell transplantation, *Annu. Rev. Med.*, 51:115–134, 2000; J. L. Platt and T. Nagayasu, Current status of xenotransplantation, *Clin. Exp. Pharmacol. Physiol.*, 26:1026–1032, 1999; A. A. Rossini, D. L. Greiner, and J. P. Mordes, Induction of immunological tolerance for transplantation, *Physiol. Rev.*, 79:99–141, 1999.

Transport processes

The processes whereby mass, energy, or momentum are transported from one region of a material to another under the influence of composition, temperature, or velocity gradients. If a sample of a material in which the chemical composition, the temperature,

or the velocity vary from point to point is isolated from its surroundings, the transport processes act so as eventually to render these quantities uniform throughout the material. The nonuniform state required to generate these transport processes causes them to be known also as nonequilibrium processes. Associated with gradients of composition, temperature, and velocity in a material are the transport processes of diffusion, thermal conduction, and viscosity, respectively. For a large class of materials, the laws that govern the transport processes are quite simple.

Diffusion. **Figure 1** shows a sample of a material which is composed of two chemical species. The sample is stationary and has a uniform temperature throughout, but a composition difference is maintained across its two ends, and in this steady state the two species continuously migrate down their concentration gradients. The composition of the material may be expressed by means of the molar concentration of one species, c_1 (moles/m^3); then it is found that the number of moles of this species which cross unit area of the sample perpendicular to the z direction in unit time, known as the flux of mass (J_1), is given by Eq. (1), which is Fick's law of diffusion.

$$J_1 = -D\frac{dc_1}{dz} \tag{1}$$

The constant of proportionality, D, between the mass flux and the concentration gradient, which depends upon the nature of the material, its temperature, pressure, and composition, is known as the diffusion coefficient.

The phenomenon of diffusion occurs widely in nature, and it is frequently important in technological applications. For example, the transpiration of the leaves of plants, in which they absorb carbon dioxide from the atmosphere and give off water vapor, is controlled by a diffusion process. The rates of many chemical reactions in fluids that are promoted by catalysts may similarly be controlled by the diffusion of reactants to the active catalyst sites. *See* DIFFUSION.

Thermal conduction. In **Fig. 2**, a sample of a material is subjected to a steady temperature difference between two faces perpendicular to the z direction. Under these conditions, energy is continually transported from the hotter face to the colder, and the

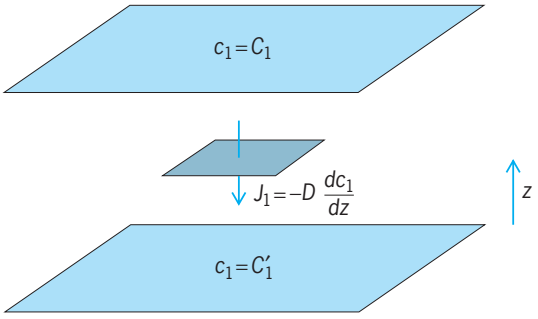


Fig. 1. Diffusion in a sample of material composed of two chemical species, c_1 and C'_1 represent the molar concentrations of one of the species in the two planes bounding the material.

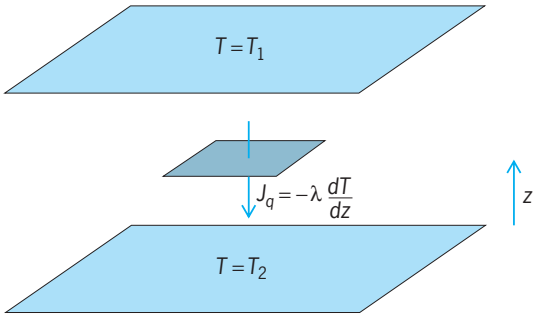


Fig. 2. Thermal conductivity in a sample of material subjected to a steady temperature difference.

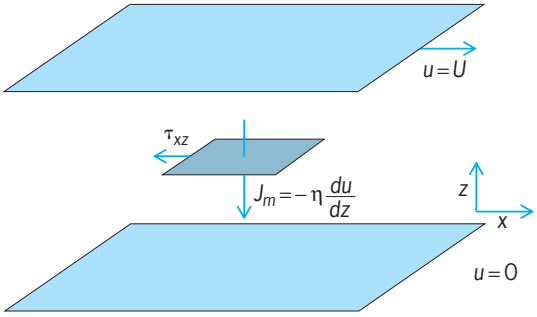


Fig. 3. Viscosity in a fluid whose upper surface is in contact with a solid boundary.

energy flux J_q in the z direction (the energy crossing unit area in unit time) is given by Fourier's law as Eq. (2). The constant of proportionality, λ, between

$$J_q = -\lambda\frac{dT}{dz} \tag{2}$$

the flux and the temperature gradient, dT/dz, is the thermal conductivity coefficient, which again depends upon the material as well as its temperature, pressure, and composition. *See* THERMAL CONDUCTION IN SOLIDS.

Viscosity. The phenomenon of viscosity is associated with the gradient of velocity in a material. Since it is difficult to maintain velocity gradients in solids, the phenomenon is readily observed only in fluids. Because velocity is a vector quantity, **Fig. 3** shows a fluid whose upper surface is in contact with a solid boundary which moves with a steady velocity U in the x direction only; the lower surface of the fluid is held stationary. As a result, various layers of the fluid in the z direction move with different x-direction velocities u. Associated with the motion in the x direction the fluid possesses a momentum, and the x-direction momentum is transported down the velocity or momentum gradient. The flux of x momentum in the z direction, J_m, is equivalent to a tangential shear stress τ_{xz} acting in the negative x direction on each layer of the fluid. This means that a tangential force must be applied to the upper plate to keep it in steady motion. Again the flux is proportional to the imposed gradient and is given by Eq. (3), which is Newton's law of viscosity. The pro-

$$J_m = \tau_{xz} = -\eta\frac{du}{dz} \tag{3}$$

portionality constant η is the viscosity coefficient for the material, and it too depends on the thermodynamic state of the material. The phenomenon of viscosity is revealed whenever a fluid flows near a solid boundary, and it is therefore of significance in almost every aspect of engineering. *See* VISCOSITY.

Thermal processes. Other, more subtle transport processes can occur. For example, in a mixture of two chemical species, the imposition of a temperature gradient leads not only to energy transport but also to a mass transport that causes a partial separation of the mixture. This phenomenon is known as thermal diffusion. Conversely, when diffusion takes place in an initially isothermal mixture as a result of a composition gradient, small temperature gradients can be observed in the material arising from an energy transport. This is the diffusion thermoeffect.

Transport coefficients. The coefficients D, λ, and η are known collectively as transport coefficients. The measurement of these coefficients for materials in solid, liquid, and gaseous phases has been the object of a considerable research effort for many years. The measurements can be carried out only rarely by directly implementing the situations envisaged in Figs. 1–3. This is because it is difficult to achieve the one-dimensional gradients of the quantities required when the sample is of a finite size. The exceptions to this are the measurement of thermal conductivity and diffusion in solids, where simple methods have proved effective. In fluids, the diffusion coefficient has most often been determined in a time-dependent experiment in which an initial concentration gradient in a mixture is allowed to decay in a closed vessel of known geometry. The approach to equilibrium, which is governed by the diffusion coefficient, is observed with a suitable concentration monitor.

The coefficient of thermal conductivity in fluids is also most accurately determined in a transient experiment. The fluid surrounds a thin, vertical wire which is suddenly heated by an electric current; the rate of the temperature rise of the wire is observed, and the thermal conductivity of the fluid is deduced from it.

The viscosity coefficient of gases and liquids is generally determined by one of two techniques. In the first, the fluid flows through a capillary tube of known geometry, and the pressure difference across the ends of the tube necessary to maintain a given rate of flow is determined. This pressure difference is then proportional to the viscosity coefficient of the fluid. In the second method, the damping of the torsional oscillations of a thin, solid, horizontal, circular disk is observed when the disk is suspended in the fluid. The measurement of the logarithmic decrement of the oscillations serves to determine the viscosity coefficient. *See* QUASIELASTIC LIGHT SCATTERING.

The results of measurements of the transport coefficients of materials are of importance since there are few technological activities that do not involve one or more of the transport processes. However, because the transport coefficients derive their values from the properties and behavior of the atoms and molecules that make up the material, they are also of more fundamental significance. In the particular case of gases at low density (near atmospheric pressure), the kinetic theory of gases has provided an almost complete description of their transport coefficients. In such gases the molecules are in continual random motion, undergoing collisions with each other and with the walls of a containing vessel. Thus, in a gas subject to, for example, a velocity gradient, the molecules that have an additional component of momentum, because they are in regions of high velocity in the gas, transport that momentum down the velocity gradient. Naturally, the rate of transport of momentum is influenced by the number of molecules transporting the momentum and their speed, as well as the likelihood and effect of collisions with other molecules. For the simplest molecular model, in which the molecules are seen as rigid elastic spheres of diameter σ, there is no force between molecules except when they touch. Thus, the dynamics and effects of the collisions are straightforward, and the transport coefficients for viscosity and thermal conductivity are given by Eqs. (4) and (5), in which m is the mass of a molecule, k is Boltzmann's constant, T the absolute temperature, and c_v the specific heat of the gas. Real molecules exert forces on each other over a distance, which are expressed in terms of an intermolecular pair potential such that the forces between molecules are repulsive at short range and attractive at long range. This means that the dynamics of molecular encounters are more complicated and the effects of the collisions on the transport of momentum and energy are altered. Within the kinetic theory of gases, these effects can be included; for systems of monatomic gases, the only modification to the formulas given in Eqs. (4) and (5) is effectively to introduce a temperature-dependent diameter $\sigma(T)$. The effective diameter is smaller at higher temperatures because the atoms approach each other more closely as their translational kinetic energy increases. For polyatomic molecules that possess internal energy, this single idea remains valid for the viscosity; but for the thermal conductivity, it is necessary to consider the transport of internal energy down the temperature gradient as well, and the corresponding theoretical result is more complicated. When the transport coefficients are evaluated for real molecules, the viscosity and thermal conductivity are found to increase with temperature approximately linearly, whereas the diffusion coefficient varies more nearly as the square of the temperature.

$$\eta = \frac{5}{16}\frac{(\pi mkT)^{1/2}}{\pi\sigma^2} \tag{4}$$

$$\lambda = \frac{25}{32}\frac{(\pi mkT)^{1/2}}{\pi\sigma^2}\frac{c_v}{m} \tag{5}$$

At low densities the viscosity and thermal conductivity of a gas are independent of pressure, while the diffusion coefficient is inversely proportional to

it. As the density of the material is increased toward that of a liquid and finally to that of a solid, significant changes in the molecular mechanism of the transport processes occur. The transport by free molecular motion becomes a smaller contribution as the volume available for such motion decreases. In addition, the attractive forces between molecules, which become increasingly significant, tend to inhibit molecular motion. Thus, on the one hand, the diffusion coefficient in condensed phases, which is still determined by molecular motion, is very much smaller (about 10^4 times) than that in low-density gases. On the other hand, the viscosity of liquids is very much greater than that in gases because the attractive forces between molecules make the relative motion of various layers in the fluid much more difficult to achieve. Because increasing the temperature of a liquid increases the average separation of the molecules as well as their energy, the diffusion coefficients for liquids increase rapidly with temperature and, for the same reasons, the viscosity decreases. In the solid the molecules acquire almost fixed positions, and the diffusion coefficient consequently becomes even smaller, whereas the viscosity is practically infinite. W. A. Wakeham

Bibliography. J. Kestin and W. A. Wakeham, *Transport Properties of Fluids: Thermal Conductivity, Viscosity and Diffusion Coefficients,* 1988; G. C. Maitland et al., *Intermolecular Forces: Their Origin and Determination,* 1987.

Transportation engineering

That branch of engineering related to the safe and efficient movements of people and goods. The primary modes of travel considered in transportation engineering include roadway, rail, air, water, and pipeline, along with nonmotorized pedestrian and bicycle travel. Special categories include urban and intermodal transportation.

Highway transportation. Engineering for highway transportation involves planning, construction, and operation of highway systems, urban streets, roads, and bridges, as well as parking facilities. Important aspects of highway engineering include (1) overall planning of routes, financing, environmental impact evaluation, and value engineering to compare alternatives; (2) traffic engineering, which plans for the volumes of traffic to be handled, the methods to accommodate these flows, the lighting and signing of highways, and general layout; (3) pavement and roadway engineering, which involves setting of alignments, planning the cuts and fills to construct the roadway, designing the base course and pavement, and selecting the drainage system; and (4) bridge engineering, which involves the design of highway bridges, retaining walls, tunnels, and other structures. *See* HIGHWAY ENGINEERING; TRAFFIC-CONTROL SYSTEMS; VALUE ENGINEERING.

Highway transportation engineers face the challenge of moving increasing volumes of traffic over existing routes while improving safety records. Therefore, a major initiative was launched in the United States to develop technology for a nationwide intelligent vehicle highway system (IVHS). In urban areas, such systems are being designed to feature increased use of vehicle detection systems, video cameras, variable message signs, and electronic toll collection with automatic vehicle identification. In more remote areas, sensors on highways are being designed to detect unsafe conditions such as slippery roads, relay the information via radio waves to a central computer, and transmit a warning via satellite to individual vehicles. Eventually, new cars will be instrumented so that the driver can communicate with municipalities and receive directions or other information. Posted speed limits will be changed automatically according to traffic flow.

In modern highway construction, particularly in urban areas, major emphasis is placed on developing regional transportation plans utilizing significant public input. Highways must meet the transportation needs of the public, adhere to strict environmental regulations, and satisfy concerns over traffic and esthetics, in order to gain wide support. Air pollution, noise problems, and destruction of wetlands are some concerns that must be carefully addressed.

Highway pavements must be maintained with surfaces of acceptable riding quality. This places severe demands on paving materials because of heavy traffic volumes, repetitive truck axle loads, weather conditions that may vary from severe freezing to hot summers, and the use of chemicals to aid in snow and ice removal. As a result of a national Strategic Highway Research Program (SHRP) in the United States, an improved asphalt paving system known as Superpave has been developed and has been implemented to improve pavement life substantially. *See* PAVEMENT.

Bridge engineers are now using new or improved materials to develop more cost-effective, longerlasting structures. A high-performance steel (HPS-70W) that provides improved toughness, weldability, and strength has been used to build several bridges. Also, a high-performance concrete (HPC) that provides higher strength and improved workability is being used. Bridge rehabilitation is a major area for infrastructure innovation. Estimates show that approximately one-third of the bridges in the United States are in urgent need of repair. Deterioration of the bridge deck often leads to deterioration of the underlying superstructure, and methods of constructing more durable decks have been developed. This includes the use of epoxy-coated steel reinforcing bars and high-performance concrete with microsilica for deck construction. Precast elements are sometimes effective in deck replacement, because disruption to traffic can be minimized. Bridges are susceptible to the forces generated by seismic events, as evidenced by the destruction resulting from the Loma Prieta and Northridge earthquakes in California. Bridge engineers have developed methods for retrofitting existing structures to resist earthquakes, as well as designing improved details for new construction. To increase load capacity and reliability, posttensioned cables can be added to beams and girders for longer

spans. For shorter spans of about 50 ft (15 m) or less, bridges can be replaced with large culverts constructed from prefabricated steel or aluminum components, or from precast concrete units. Other techniques include the use of lead-rubber seismic isolation devices to replace traditional bearings at supports, and constructing steel or fiberglass casings around concrete bridge columns to increase strength and ductility. *See* BRIDGE; EARTHQUAKE; STRUCTURAL MATERIALS.

Rail transportation. Engineering for railway transportation involves planning, construction, and operation of terminals, switchyards, loading/unloading facilities, trackage, bridges, tunnels, and traffic-control systems for freight and passenger service. For freight operations, there is an emphasis on developing more efficient systems for loading, unloading, shifting cars, and operating trains. Facilities include large marshaling yards where electronic equipment is used to control the movement of railroad cars. Also, there is a trend to developing more automated systems on trackage whereby signals and switches are set automatically by electronic devices. To accommodate transportation of containers, tunnels on older lines are being enlarged to provide for double-stack container cars.

Efforts for more efficient operations have begun to pay dividends. Reversing the declines of earlier years, rail freight volume has been steadily increasing since 1987. New facilities are being built to handle the increased volume, and tracks are being added at points of congestion. Although the total trackage in the United States today is only 113,000 mi (182,000 km), major railroads haul significant tonnage.

In Europe and Japan, rail passenger service has long been an important part of the overall transportation system. In the United States, passenger service is small compared to that of earlier years, but there is a renewed interest in rail passenger service in certain congested urban corridors. In Japan and Europe, high-speed systems have been developed. For example, the French have developed an Aerotrain that is propelled by a fan jet engine and floats on a cushion of low-pressure air on a guideway at 267 mi/h (430 km/h). A more conventional high-speed system introduced in 1994, the Eurostar, operates at service speeds of up to 190 mi/h (300 km/h). One development that has received limited attention in Germany, Japan, and the United States is a magnetic levitation system in which cars on a frictionless magnetic suspension are propelled along guideways by linear induction motors. A proposed line in Germany would be capable of speeds near 300 mi/h (480 km/h). *See* MAGNETIC LEVITATION.

Most high-speed rail systems require new alignments to accommodate the increased speed. However, in many areas it is difficult and expensive to construct new alignments. Thus, engineers are challenged to develop high-speed rail passenger cars that can operate on existing trackage. One approach has been to develop suspension systems that tilt the coaches on curves to enhance stability. In the mid-1990s the fastest trains in the United States were running about 100 mi/h (160 km/h). In 1998 the Transportation Research Board initiated a program to develop technologies to upgrade rail systems to higher speeds.

A major accomplishment in intercontinental rail transportation was the completion in late 1993 of the Channel Tunnel. Culminating nearly 8 years of design and construction, the twin tunnels facilitate a trip from London to Paris in 3 h. The tunnel was designed to handle a full range of services, including conventional and intermodal freight operations, motor car and truck carrying services, and high-passenger trains. *See* RAILROAD ENGINEERING; TUNNEL.

Air transportation. Engineering for air transportation encompasses the planning, design, and construction of terminals, runways, and navigation aids to provide for passenger and freight service. High-capacity, long-range, fuel-efficient aircraft, such as the 440-seat Boeing 777 with a range of 7200 mi (12,000 km) are desirable. Wider use of composites and the substitution of electronic controls for mechanical devices reduce weight to improve fuel economy. Smaller planes are more efficient for shorter runs. *See* AIR TRANSPORTATION; COMPOSITE MATERIAL.

Many airport terminals for smaller population centers are designed with little uncommitted space but with provisions for incremental expansion as traffic grows. In many larger population centers, state-of-the-art facilities and major improvements have been constructed, such as the terminals at Denver and Washington, D.C. (both National and Dulles). As air transportation continues to grow, increased emphasis is placed on terminal designs that facilitate the rapid movement of passengers and baggage. Extensive conveyor belt systems with electronic identifications have been installed in the newer facilities to speed movement of baggage from planes to common central areas. At the larger airports, moving sidewalks are commonly used to move passengers through widely dispersed gate areas. Also, computer-controlled shuttle cars without operators (people movers) are used to move people rapidly from central areas to gate areas. In some cases, rapid transit rail facilities or dedicated busways are available to carry passengers directly from the airport to central business districts. As airways continue to concentrate flights in large hub operations to increase efficiency, engineers and planners focus on developing improved facilities to speed the transfer of passengers and baggage. *See* AIRPORT ENGINEERING.

Air freight cargo facilities are increasingly being automated. An example is the cargo facility at New York's John F. Kennedy Airport, which features a computerized control system, a network of automated distribution conveyors, transfer vehicles, and stacking cranes. A large rack storage building with robotic cranes provides temporary storage and speeds loading and unloading of roll boxes and lightweight containers that hold groups of packages.

Runways must be of sufficient size to safely accommodate landings and takeoffs. The design of

pavements for runways and taxiways involves many of the same considerations as for highways, except that runways must be designed to reliably withstand the heavier wheel loads imposed by the aircraft. Navigation aids and other instrumentation, such as for the detection of destructive wind shears or approaching aircraft, are being developed and deployed to enhance safety. *See* AIR NAVIGATION; AIR-TRAFFIC CONTROL.

Water transportation. Engineering for water transportation entails the design and construction of a vast array of facilities such as canals, locks and dams, and port facilities. The transportation system ranges from shipping by barge and tugboat on inland waterways to shipping by oceangoing vessels. Although there is some transportation of passengers, such as on ferries and cruise ships, water transportation is largely devoted to freight.

One important inland waterway in North America is the St. Lawrence Seaway, which connects natural waterways, canals, and the Great Lakes. It permits shipping from the Atlantic Ocean to Duluth, Minnesota. Another is the Ohio-Missouri-Mississippi River System, which extends from Pittsburgh, Kansas City, and Minneapolis to the Gulf of Mexico. Many of the locks and dams in the latter system were constructed many years ago. Engineers are challenged to design and construct larger locks to serve larger barges. Improved navigation systems are needed to enhance safety. *See* CANAL; DAM; RIVER ENGINEERING.

Special facilities are required for loading and unloading oceangoing vessels. For example, supertankers require special offshore mooring where the oil can be rapidly pumped to shore through pipelines. Also, unit trains of 100 or more railcars, carrying bulk products such as grain or coal, require rapid unloading facilities for efficient operations. *See* HARBORS AND PORTS.

Pipeline transportation. Pipeline engineering embraces the design and construction of pipelines, pumping stations, and storage facilities. Pipelines are used to transport liquids such as water, gas, and petroleum products over great distances. Also, products such as pulverized coal and iron ore can be transported in a water slurry.

Water pipelines, probably the most common, can run from minimal sizes to diameters of 20 ft (6 m) or more for large penstocks. Pipelines are often underground, but may run aboveground, particularly in lightly populated areas. Either submerged lines or bridges are required at stream crossings. One aboveground line is the 798-mi (1280-km) 48-in.-diameter (1200-mm) Trans-Alaska pipeline, which was constructed under difficult conditions to transport oil from near Prudhoe Bay on the Arctic Ocean to the city of Valdez on the Gulf of Alaska.

Consideration must be given to route selection, determining the appropriate diameter and thickness of pipe, installation (trench construction, backfilling, and compaction), and durability. Installation under highways, rivers, and other difficult areas is sometimes done using microtunneling techniques to avoid disturbing the surface. With this method, the pipe is advanced underground by jacking from one side, using a laser-controlled guidance system to maintain pipe alignment. Design of pumping facilities requires study of power requirements for different types of material moved, standby facilities, and related considerations. *See* PIPELINE; STORAGE TANK.

Urban transportation. Engineering for urban transportation concerns the design and construction of light rail systems, subways, and people-movers, as well as facilities for traditional bus systems. To enhance public acceptance of new and expanded systems, increased use is being made of computer-aided design (CAD) to visualize alternatives for stations and facilities. Also, animated video systems are used for interactive visualization of plans. *See* COMPUTER-AIDED ENGINEERING.

As congestion in urban areas grows, increased emphasis is being placed on the construction and expansion of subways in major cities. Tunnel-boring machines and segmental liners are often used, along with various techniques to mobilize the strength of the surrounding soil or rock. Light-rail cars are used interchangeably aboveground on controlled right-of-ways or underground through tunnels. *See* SUBWAY ENGINEERING.

In some cities, high-occupancy vehicle lanes have been used to encourage car pooling. Typically these special lanes are reserved during rush hours for vehicles with three or more passengers.

Intermodal transportation. Intermodal transportation, often referred to as containerization, entails the use of special containers to ship goods by truck, rail, or ocean vessel. Engineers must design and construct intermodal facilities for efficient operations. The containers are fabricated from steel or aluminum, and they are designed to withstand the forces from handling. The ships are constructed with a cellular grid of compartments for containers below deck, and they can accommodate one or two layers on deck as well. Advantages include savings in labor costs, less pilferage, and lower insurance costs.

Seaports have special facilities for handling the containers, which arrive by rail or truck. For example, a modernization project in Boston, Massachusetts, provides a 13-acre (5.2-hectare) yard for container storage with 1950 ft (594 m) of continuous berth face for ships. Containers are handled by two special 30-ton-capacity (27-metric ton) gantry cranes that travel on rails with a 96-ft (29-m) gage, and provide an outreach of 115 ft (35 m). Such facilities, which are found throughout the major port cities, are designed to speed the transfer of freight. At major inland centers, intermodal facilities combine loading and handling facilities for containers from truck trailers and rail flatcars. *See* HOISTING MACHINES; MARINE CONTAINERS; MERCHANT SHIP.

Environmental considerations. The environment is a major consideration in planning, designing, and constructing transportation facilities. Extensive legislation at both federal and state levels has set forth requirements with the objective of protecting the health and welfare of the general public.

Beginning with the National Environmental Policy Act (NEPA) of 1970, over 25 major federal laws and countless amendments have been enacted. One provision of NEPA is the requirement for approved environmental impact statements (EIS) for federally funded projects. The EIS must document how the environment will be affected, including identification of any unavoidable adverse effects. Opportunities for public involvement in the process must be provided, and alternatives must be identified. To help control air quality, the Clean Air Act administered by the Environmental Protection Agency (EPA) requires review and approval of any new highway with an anticipated average daily traffic volume of more than 20,000 vehicles, or any new parking facility for 1000 cars or more. Approvals are also required for major modifications that increase capacity. The Noise Control Act limits noise levels from surface transportation, often requiring the construction of noise walls adjacent to highways in urban areas. To control water pollution, the Clean Water Act requires permits for discharges into streams and wetlands, both from construction activities and from storm water runoff of the completed facility. *See* AIR POLLUTION; WATER POLLUTION.

To avoid exposure of both workers and the public to lead poisoning, special precautions must be taken in the removal and containment of old, deteriorating paint systems from steel bridges. Regulations of the Occupational Safety and Health Administration (OSHA) must be met. Lead-based paint systems were used until the early 1980s, but new steel bridge construction uses either bare weathering steel or modern lead-free paint systems.

Because of the vast material quantities used to construct and maintain transportation systems, opportunities abound for the use of reclaimed and recycled resources. For example, in the production of new pavement, 20% or more of the content can be from old asphalt pavement that has been removed in repaving operations. Old concrete pavements can also be recycled, as aggregate for new concrete or for base layers. Industrial wastes such as fly ash and blast-furnace slag can be used in concrete. Many uses have been found for scrap tires, such as for erosion control devices, for safety devices (tire-sand inertial barriers), in chipped form to create a lightweight bridge approach embankment or to reduce earth pressure against abutment walls, and in the manufacture of rubber asphalt. *See* RECYCLING TECHNOLOGY.

Energy considerations. Engineers must help meet the challenge to reduce energy consumption by designing efficient transportation systems. About 97% of transportation energy in the United States is derived from oil. Indeed, the transportation system uses about 65% of all oil consumed by the United States. Although the fuel economy of vehicles has been improved significantly, this factor has been offset by a trend toward decreased vehicle occupancy rates. As a result, total usage of energy in highway passenger transportation is increasing. Despite significant gains that have been made in energy efficiency for commercial air travel and for rail freight, total energy consumption for transportation of all types continues to grow.

Efforts to curb energy use arise from a variety of concerns, including national security issues and environmental implications. Limiting carbon dioxide emissions to stem global warming is a significant consideration. In addition to fuel economy standards, legislative approaches to date have focused on promoting cleaner and more efficient alternative fuels. As an example, natural gas is now used to fuel certain commercial vehicles in some areas. However, because personal travel in light-duty vehicles consumes over 50% of transportation energy, effective efforts in energy conservation must be directed toward them. Further improvements in fuel economy will likely lead to lighter-weight vehicles with higher first costs. Efforts to relieve congestion in urban areas through incentives to make greater use of car pooling, such as special freeway lanes, and encouraging greater use of mass transit, deserve further emphasis.

Roger L. Brockenbrough

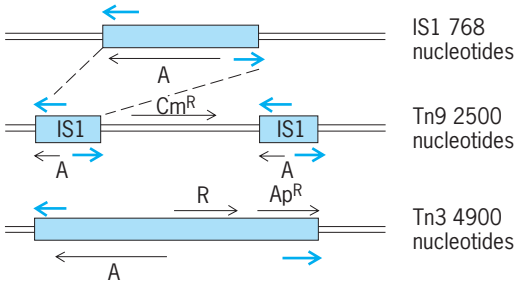
Bibliography. J. Banks, *Introduction to Transportation Engineering*, 2d ed., 2002; R. L. Brockenbrough and K. J. Boedecker, Jr., *Highway Engineering Handbook*, 2d ed., 2003; M. A. Chowdhury and A. W. Sadek, *Fundamentals of Intelligent Transportation Systems Planning*, 2003; Federal Highway Administration, *Manual on Uniform Traffic Control Devices*, 2003; C. Khisty and B. Lall, *Transportation Engineering: An Introduction*, 3d ed., 2003; R. Roess, E. Prassas, and W. McShane, *Traffic Engineering*, 3d ed., 2004; Transportation Research Board, *Highway Capacity Manual*, 2000; P. Wright and K. Dixon, *Highway Engineering*, 7th ed., 2004.

Transposons

Types of transposable elements which comprise large discrete segments of deoxyribonucleic acid (DNA) capable of moving from one chromosome site to a new location. In bacteria, the transposable elements can be grouped into two classes, the insertion sequences and the transposons. The ability of transposable elements to insert into plasmid or bacterial virus (bacteriophage) which is transmissible from one organism to another allows for their rapid spread. *See* BACTERIOPHAGE; PLASMID.

The insertion sequences were first identified by their ability to induce unusual mutations in the structural gene for a protein involved in sugar metabolism. These insertion sequences are relatively small (about 500–1500 nucleotide pairs) and can only be followed by their ability to induce these mutations. Most bacterial chromosomes contain several copies of such insertion sequence elements.

The transposons are larger segments of DNA (2000–10,000 base pairs) that encode several proteins, usually one or two required for the movement



Structure of a typical transposable element. The colored arrows indicate the terminal inverted repeats characteristics of each element. Note that Tn9 is a composite transposon derived from directly repeated IS1 elements. The black arrows indicate genes for proteins involved in transposition (A and R) or antibiotic resistance. The Tn9 transposon is resistant to chloramphenicol (Cm), while the Tn3 element encodes resistance to ampicillin (Ap) and its derivative penicillin.

of the element and often an additional protein that imparts a selective advantage to the host containing a copy of that element. The structure of many transposons suggests they may have evolved from the simpler insertion sequence elements. For example, the transposon Tn9 contains two copies of the element IS1 flanking a region of unique DNA encoding resistance to the antibiotic chloramphenicol.

All transposable elements, both the simple insertion sequence elements and the more complex transposons, have a similar structure and genetic organization (see **illus.**). The ends of the element represent recognition sites and define the segment of DNA undergoing transposition. A short sequence present at one end of the element is repeated in an inverted fashion at the other end. These terminal inverted repeats are characteristic for each element. In the case of the composite transposons like Tn9, the inverted repeats present at the end of each IS1 element result in the entire transposon also having inverted repeats. One or more proteins essential in the recognition of the inverted repeat are encoded in the body of the element.

Members of a widespread group of transposons, the Tn3 family, all have a similar structure and appear to move by a similar mechanism. Transposase, one protein encoded by the element, promotes the formation of intermediates called cointegrates, in which the element has been duplicated by replication. A second element-encoded protein, resolvase, completes the process by converting the cointegrates into the end products of transposition, a transposon inserted into a new site. A third protein encoded by the Tn3 element imparts resistance to the antibiotic ampicillin.

Transposons are known that encode resistances to almost all antibiotics as well as many toxic metals and chemicals. In addition, some transposons have acquired the ability to direct the synthesis of proteins that metabolize carbohydrates, petroleum, and pesticides. Other transposable elements produce enterotoxins that cause travelers to become ill from drinking water contaminated with bacteria carrying the element. The broad spectrum of activities encoded by the transposable elements demonstrates the strong selective advantage that has accompanied their evolution.

The bacteriophage Mu (mutator) replicates itself in a mechanism that involves transposition into many sites in the host genome. In the process of high-frequency transposition, the bacteriophage often mutates genes in the host organism. Other phages have adopted transpositionlike events for special purposes.

Transposable elements are not restricted to prokaryotes. Yeast as well as higher eukaryotes have DNA segments that move and cause mutations. In fact, the earliest models suggesting the existence of transposable DNA segments were based on genetic work by B. McClintock in the 1930s with corn plants. The eukaryotic elements have much in common with their prokaryotic counterparts: the termini of the elements are composed of inverted repeats, and many of the larger elements are composed of two small insertion sequence–like regions flanking a unique central region. One class of eukaryotic virus, the ribonucleic acid (RNA) retrovirus, also has this structure and is thought to integrate into the host chromosome through a transposition-like mechanism. *See* ANTIBIOTIC; GENE; RETROVIRUS; VIRUS.

Randall Reed

Bibliography. D. Berg and M. Howe (eds.), *Mobile DNA,* 1989; N. Federoff, Controlling elements in maize, *Sci. Amer.*, 250(6):84–91, 1984; N. Kleckner, Transposable elements in prokaryotes, *Annu. Rev. Genet.*, 15:341–404, 1981; M. E. Lambert, J. F. McDonald, and I. B. Weinstein, *Eukaryotic Transposable Elements as Mutagenic Agents,* 1988; O. Nelson (ed.), *Plant Transposable Elements,* 1988; J. A. Shapiro (ed.), *Mobile Genetic Elements*, 1983.

Transuranium elements

Those synthetic elements with atomic numbers larger than that of uranium (atomic number 92). They are the members of the actinide series, from neptunium (atomic number 93) through lawrencium (atomic number 103), and the transactinide elements (with atomic numbers higher than 103). Of these elements, plutonium, an explosive ingredient for nuclear weapons and a fuel for nuclear power because it is fissionable, has been prepared on the largest (ton) scale, while some of the others have been produced in kilograms (neptunium, americium, curium) and in much smaller quantities (berkelium, californium, and einsteinium).

The concept of atomic weight as applied to naturally occurring elements is not applicable to the transuranium elements, since the isotopic composition of any given sample depends on its source. In most cases the use of the mass number of the longest-lived isotope in combination with an evaluation of its

availability has been adequate. Good choices at present are neptunium, 237; plutonium, 242; americium, 243; curium, 248; berkelium, 249; californium, 249; einsteinium, 254; fermium, 257; mendelevium, 258; nobelium, 259; lawrencium, 260; rutherfordium, 261; dubnium, 262; seaborgium, 266; bohrium 267; and hassium 269.

The actinide elements are chemically similar and have a strong chemical resemblance to the lanthanide, or rare-earth, elements (atomic numbers 57–71). The transactinide elements, with atomic numbers 104 to 118, should be placed in an expanded periodic table under the row of elements beginning with hafnium, number 72, and ending with radon, number 86. This arrangement allows prediction of the chemical properties of these elements and suggests that they will have an element-by-element chemical analogy with the elements that appear immediately above them in the periodic table. However, deviations from this analogy are expected and are observed in specific detailed chemical properties of transactinide elements. *See* ACTINIDE ELEMENTS; PERIODIC TABLE; RARE-EARTH ELEMENTS.

The transuranium elements up to and including fermium (atomic number 100) are produced in the largest quantity through the successive capture of neutrons in nuclear reactors. The yield decreases with increasing atomic number, and the heaviest to be produced in weighable quantity is einsteinium (number 99). Many additional isotopes are produced by bombardment of heavy target isotopes with charged atomic projectiles in accelerators; beyond fermium all elements are produced by bombardment with heavy ions. Brief descriptions of transuranium elements follow. They are listed according to increasing atomic number.

Neptunium. Neptunium (Np, atomic number 93, named after the planet Neptune) was the first transuranium element discovered. In 1940 E. M. McMillan and P. H. Abelson at the University of California, Berkeley, identified the isotope ^{239}Np (half-life 2.35 days), which was produced by the bombardment of uranium with neutrons according to reaction (1).

$$^{238}\mathrm{U}(n,\gamma)^{239}\mathrm{U} \longrightarrow {}^{239}\mathrm{Np} \qquad (1)$$

The element as ^{237}Np was first isolated as a pure compound, the oxide, in 1944 by L. G. Magnusson and T. J. La Chapelle. Neptunium in trace amounts is found in nature, and is produced in nuclear reactions in uranium ores caused by the neutrons present. Kilogram and larger quantities of ^{237}Np (half-life 2.14×10^6 years), used for chemical and physical investigations, are produced as a by-product of the production of plutonium in nuclear reactors. Isotopes from mass number 227 to 244 have been synthesized by various nuclear reactions. *See* NUCLEAR REACTION.

Neptunium displays five oxidation states in aqueous solution: Np^{3+} (pale purple), Np^{4+} (yellow-green), NpO_2^+ (green-blue), NpO_2^{2+} (pink), and NpO_5^{3-} (green). The ion NpO_2^+, unlike corresponding ions of uranium, plutonium, and americium, can exist in aqueous solution at moderately high concentrations. The element forms tri- and tetrahalides such as NpF_3, NpF_4, $NpCl_3$, $NpCl_4$, $NpBr_3$, NpI_3, as well as NpF_6 and oxides of various compositions such as those found in the uranium-oxygen system, including Np_3O_8 and NpO_2.

Neptunium metal has a silvery appearance, is chemically reactive, and melts at 637°C (1179°F); it has at least three crystalline forms between room temperature and its melting point. *See* NEPTUNIUM.

Plutonium. Plutonium (Pu, atomic number 94, named after the planet Pluto) in the form of ^{238}Pu was discovered in late 1940 and early 1941 by G. T. Seaborg, McMillan, J. W. Kennedy, and A. C. Wahl at the University of California, Berkeley. The element was produced in the bombardment of uranium with deuterons according to reaction (2).

$$^{238}\mathrm{U}(d,2n)^{238}\mathrm{Np} \xrightarrow[2.1\,\mathrm{days}]{\beta^-} {}^{238}\mathrm{Pu} \qquad (2)$$

The important isotope ^{239}Pu was discovered by Kennedy, Seaborg, E. Segrè, and Wahl in 1941. Because of its property of being fissionable with neutrons, plutonium-239 (half-life 24,400 years) is used as the explosive ingredient in nuclear weapons and is a key material in the development of nuclear energy for industrial purposes. 1 lb (0.45 kg) of plutonium produced is equivalent to about 10^7 kWh of heat energy; plutonium is produced in ton quantities in nuclear reactors. The alpha radioactivity and physiological behavior of this isotope make it one of the most dangerous poisons known, but means for handling it safely have been devised. Plutonium as ^{239}Pu was first isolated as a pure compound, the fluoride, in 1942 by B. B. Cunningham and L. B. Werner. Minute amounts of plutonium formed in much the same way as naturally occurring neptunium are present in nature. Much smaller quantities of the longer-lived isotope ^{244}Pu (half-life 8.3×10^7 years) have been found in nature; in this case it may represent the small fraction remaining from a primordial source or it may be caused by cosmic rays. Isotopes of mass number 232–246 are known. The longer-lived isotopes ^{242}Pu (half-life 390,000 years) and ^{244}Pu, produced in nuclear reactors, are more suitable than ^{239}Pu for chemical and physical investigation because of their longer half-lives and lower specific activities.

Plutonium has five oxidation states in aqueous solution: Pu^{3+} (blue to violet), Pu^{4+} (yellow-brown), PuO_2^+ (pink), PuO_2^{2+} (pink-orange), and PuO_5^{3-} (blue-green). The ions Pu^{4+} and PuO_2^+ undergo extensive disproportionation to the ions of higher and lower oxidation states. Four oxidation states (III, IV, V, and VI) can exist simultaneously at appreciable concentrations in equilibrium with each other, an unusual situation that leads to complicated solution phenomena.

Plutonium forms binary compounds with oxygen (PuO, PuO_2, and intermediate oxides of variable composition); with the halogens (PuF_3, PuF_4, PuF_6, $PuCl_3$, $PuBr_3$, PuI_3); with carbon, nitrogen, and silicon

(including PuC, PuN, $PuSi_2$); in addition, oxyhalides are well known (PuOCl, PuOBr, PuOI).

The metal is silvery in appearance, is chemically reactive, melts at 640°C (1184°F), and has six crystalline modifications between room temperature and its melting point. *See* PLUTONIUM.

Americium. Americium (Am, atomic number 95, named after the Americas) was the fourth transuranium element discovered. The element as ^{241}Am (half-life 433 years) was produced by the intense neutron bombardment of plutonium and was identified by Seaborg, R. A. James, L. O. Morgan, and A. Ghiorso in late 1944 and early 1945 at the wartime Metallurgical Laboratory at the University of Chicago. By using the isotope ^{241}Am, the element was first isolated as a pure compound, the hydroxide, in 1945 by B. B. Cunningham. Isotopes of mass numbers 237–247 have been prepared. Kilogram quantities of ^{241}Am are being produced in nuclear reactors. The less radioactive isotope ^{243}Am (half-life 7400 years), also produced in nuclear reactors, is more suitable for use in chemical and physical investigation.

Americium exists in four oxidation states in aqueous solution: Am^{3+} (light salmon), AmO_2^+ (light tan), AmO_2^{2+} (light tan), and a fluoride complex of the IV state (pink). The trivalent state is highly stable and difficult to oxidize. AmO_2^+, like plutonium, is unstable with respect to disproportionation into Am^{3+} and AmO_2^{2+}. The ion Am^{4+} may be stabilized in solution only in the presence of very high concentrations of fluoride ion, and tetravalent solid compounds are well known. Divalent americium has been prepared in solid compounds; this is consistent with the presence of seven 5*f* electrons in americium (enhanced stability of half-filled 5*f* electron shell) and is similar to the analogous lanthanide, europium, which can be reduced to the divalent state.

Americium dioxide, AmO_2, is the important oxide; Am_2O_3 and, as with previous actinide elements, oxides of variable composition between $AmO_{1.5}$ and AmO_2 are known. The halides AmF_2 (in CaF_2), AmF_3, AmF_4, $AmCl_2$ (in $SrCl_2$), $AmCl_2$, $AmBr_3$, AmI_2, and AmI_3 have also been prepared.

Metallic americium is silvery-white in appearance, is chemically reactive, and has a melting point of 1176°C (2149°F). It has two crystalline forms between room temperature and its melting point. *See* AMERICIUM.

Curium. The third transuranium element to be discovered, curium (Cm, atomic number 96, named after Pierre and Marie Curie), as the isotope ^{242}Cm, was identified by Seaborg, James, and Ghiorso in 1944 at the wartime Metallurgical Laboratory of the University of Chicago. This was produced by the helium-ion bombardment of ^{239}Pu in the University of California 60-in. (152-cm) cyclotron. Curium was first isolated, using the isotope ^{242}Cm, in the form of a pure compound, the hydroxide, in 1947 by L. B. Werner and I. Perlman. Isotopes of mass number 238–251 are known. Chemical investigations have been performed using ^{242}Cm (half-life 163 days) and ^{244}Cm (half-life 18 years), but the higher-mass isotopes ^{247}Cm and ^{248}Cm with much longer half-lives (1.6×10^7 and 3.5×10^5 years, respectively) are more satisfactory for this purpose; these are all produced by neutron irradiation in nuclear reactors.

Curium exists solely as Cm^{3+} (colorless to yellow) in the uncomplexed state in aqueous solution. This behavior is related to its position as the element in the actinide series in which the 5*f* electron shell is half filled; that is, it has the especially stable electronic configuration $5f^7$, analogous to its lanthanide homolog, gadolinium. A curium IV fluoride complex ion exists in aqueous solution. Solid compounds include Cm_2O_3, CmO_2 (and oxides of intermediate composition), CmF_3, CmF_4, $CmCl_3$, $CmBr_3$, and CmI_3.

The metal is silvery and shiny in appearance, is chemically reactive, melts at 1340°C (2444°F), and resembles americium metal in its two crystal modifications. *See* CURIUM.

Berkelium. Berkelium (Bk, atomic number 97, named after Berkeley, California) was produced and identified by S. G. Thompson, Ghiorso, and Seaborg in late 1949 at the University of California, Berkeley, and was the fifth transuranium element discovered. The isotope ^{243}Bk (half-life 4.6 h) was synthesized by helium-ion bombardment of ^{241}Am. The first isolation of berkelium in weighable amount, as ^{249}Bk (half-life 314 days), produced by neutron irradiation, was accomplished in 1958 by Thompson and Cunningham; this isotope, produced in nuclear reactors, is used in the chemical and physical investigation of berkelium. Isotopes of mass number 242–251 are known.

Berkelium exhibits two ionic oxidation states in aqueous solution: Bk^{3+} (yellow-green) and somewhat unstable Bk^{4+} (yellow), as might be expected by analogy with its rare-earth homolog, terbium. Solid compounds include Bk_2O_3, BkO_2 (and oxides of intermediate composition), BkF_3, BkF_4, $BkCl_3$, $BkBr_3$, and BkI_3.

Berkelium metal is chemically reactive, exists in two crystal structure modifications, and melts at 986°C (1807°F). *See* BERKELIUM.

Californium. The sixth transuranium element to be discovered, californium (Cf, atomic number 98, named after the state and University of California, Berkeley), in the form of the isotope ^{245}Cf (half-life 44 min), was first prepared by the helium-ion bombardment of microgram quantities of ^{242}Cm by Thompson, K. Street, Jr., Ghiorso, and Seaborg at Berkeley early in 1950. Cunningham and Thompson, at Berkeley, isolated californium in weighable quantities for the first time in 1958 using a mixture of the isotopes ^{249}Cf, ^{250}Cf, ^{251}Cf, and ^{252}Cf, produced by neutron irradiation. Isotopes of mass number 239–256 are known. The best isotope for the investigation of the chemical and physical properties of californium is ^{249}Cf (half-life 350 years), produced in pure form as the beta-particle decay product of ^{249}Bk.

Californium exists mainly as Cf^{3+} in aqueous solution (emerald green), but it is the first of the actinide elements in the second half of the series to exhibit the II state, which becomes progressively

more stable on proceeding through the heavier members of the series. It also exhibits the IV oxidation state in CfF_4 and CfO_2, which can be prepared under somewhat intensive oxidizing conditions. Solid compounds also include Cf_2O_3 (and higher intermediate oxides), CfF_3, $CfCl_3$, $CfBr_2$, $CfBr_3$, CfI_2, and CfI_3.

Californium metal is chemically reactive, is quite volatile, and can be distilled at temperature ranges of 1100–1200°C (2010–2190°F). It appears to exist in three different crystalline modifications between room temperature and its melting point, 900°C (1652°F). *See* CALIFORNIUM.

Einsteinium. The seventh transuranium element to be discovered, einsteinium (Es, atomic number 99, named after Albert Einstein), was found by Ghiorso and coworkers in the debris from the "Mike" thermonuclear explosion staged by the Los Alamos Scientific Laboratory in November 1952. Very heavy uranium isotopes were formed by the action of the intense neutron flux on the uranium in the device, and these decayed into isotopes of elements 99, 100, and other transuranium elements of lower atomic number. Chemical investigation of the debris in late 1952 by workers at the University of California Radiation Laboratory, Argonne National Laboratory, and Los Alamos Scientific Laboratory revealed the presence of element 99 as the isotope ^{253}Es. Einsteinium was isolated in a macroscopic (weighable) quantity for the first time in 1961 by Cunningham, J. C. Wallman, L. Phillips, and R. C. Gatti at Berkeley; they used the isotope ^{253}Es, produced in nuclear reactors, working with only a few hundredths of a microgram. The macroscopic property that they determined in this case was the magnetic susceptibility. Isotopes of mass number 243–256 have been synthesized. Einsteinium is the heaviest transuranium element to be isolated in weighable form. Most of the investigations have used the short-lived ^{253}Es (half-life 20.5 days) because of its greater availability, but the use of ^{254}Es (half-life 276 days) will increase as it becomes more available as the result of production in nuclear reactors.

Einsteinium exists in normal aqueous solution essentially as Es^{3+} (green), although Es^{2+} can be produced under strong reducing conditions. Solid compounds such as Es_2O_3, $EsCl_3$, EsOCl, $EsBr_2$, $EsBr_3$, EsI_2, and EsI_3 have been made.

Einsteinium metal is chemically reactive, is quite volatile, and melts at 860°C (1580°F); one crystal structure is known. *See* EINSTEINIUM.

Fermium. Fermium (Fm, atomic number 100, named after Enrico Fermi), the eighth transuranium element discovered, was isolated as the isotope ^{255}Fm (half-life 20 h) from the heavy elements formed in the "Mike" thermonuclear explosion. The element was discovered in early 1953 by Ghiorso and coworkers during the same investigation which resulted in the discovery of element 99. Fermium isotopes of mass number 242–259 have been prepared.

No isotope of fermium has yet been isolated in weighable amounts, and thus all the investigations of this element have been done with tracer quantities. The longest-lived isotope is ^{257}Fm (half-life about 100 days), whose production in high-neutron-flux reactors is extremely limited because of the very long sequence of neutron-capture reactions that is required.

Despite its very limited availability, fermium, in the form of the 3.24-h ^{254}Fm isotope, has been identified in the "metallic" zero-valent state in an atomic-beam magnetic resonance experiment. This established the electron structure of elemental fermium in the ground state as $5f^{12}\,7s^2$ (beyond the radon structure).

Fermium exists in normal aqueous solution almost exclusively as Fm^{3+}, but strong reducing conditions can produce Fm^{2+}, which has greater stability than Es^{2+} and less stability than Md^{2+}. *See* FERMIUM.

Mendelevium. Mendelevium (Md, atomic number 101, named after Dmitri Mendeleev), the ninth transuranium element discovered, was identified by Ghiorso, B. G. Harvey, G. R. Choppin, Thompson, and Seaborg at the University of California, Berkeley, in 1955. The element as ^{256}Md (half-life 1.5 h) was produced by the bombardment of extremely small amounts (approximately 10^9 atoms) of ^{253}Es with helium ions in the 60-in. (152-cm) cyclotron. The first identification of mendelevium was notable in that only one or two atoms per experiment were produced. (This served as the prototype for the discovery of all heavier transuranium elements, which have been first synthesized and identified on a one-atom-at-a-time basis.) Isotopes of mass numbers 247–259 are known. Although the isotope ^{258}Md (half-life 56 days) is sufficiently long-lived, it cannot be produced in nuclear reactors, and hence it will be very difficult and perhaps impossible to isolate it in weighable amount.

The chemical properties have been investigated on the tracer scale, and the element behaves in aqueous solution as a typical tripositive actinide ion; it can be reduced to the II state with moderately strong reducing agents. *See* MENDELEVIUM.

Nobelium. The discovery of nobelium (No, atomic number 102, named after Alfred Nobel), the tenth transuranium element to be discovered, has a complicated history. For the first time scientists from countries other than the United States embarked on serious efforts to compete in this field. The reported discovery of element 102 in 1957 by an international group of scientists working at the Nobel Institute for Physics in Stockholm, who suggested the name nobelium, has never been confirmed and must be considered to be erroneous. Working at the Kurchatov Institute of Atomic Energy in Moscow, G. N. Flerov and coworkers in 1958 reported a radioactivity that they thought might be attributed to element 102, but a wide range of half-lives was suggested and no chemistry was performed. As the result of more definitive work performed in 1958, Ghiorso, T. Sikkeland, J. R. Walton, and Seaborg reported an isotope of the element, produced by bombarding a mixture of curium isotopes with ^{12}C ions in the then-new Heavy Ion Linear Accelerator (HILAC) at Berkeley. They described a novel "double recoil" technique that permitted identification by chemical means, one atom at a time, of any daughter isotope of element 102

that might have been formed. The isotope ^{250}Fm was identified conclusively by this means, indicating that its parent should be the isotope of element 102 with mass number 254 produced by the reaction of ^{12}C ions with ^{246}Cm. However, another isotope of element 102, with half-life 3 s, also observed indirectly in 1958, and whose alpha particles were shown to have an energy of 8.3 MeV by Ghiorso and coworkers in 1959, was shown later by Flerov and coworkers (working at the Dubna Laboratory near Moscow) to be of an isotope of element 102 with mass number 252 rather than 254; in other words, two isotopes of element 102 were discovered by the Berkeley group in 1958, but the correct mass number assignments were not made until later. On the basis that they identified the atomic number correctly, the Berkeley scientists probably have the best claim to the discovery of element 102; they suggest the retention of nobelium as the name for this element.

All known isotopes (mass numbers 250–259) of nobelium are short-lived and are produced by the bombardment of lighter elements with charged particles (heavy ions); the longest-lived is ^{259}No with a half-life of 58 min. All of the chemical investigations have been, and presumably must continue to be, done on the tracer scale. These have demonstrated the existence of No^{3+} and No^{2+} in aqueous solutions, with the latter much more stable than the former. The stability of No^{2+} is consistent with the expected presence of the completed shell of fourteen 5*f* electrons in this ion. *See* NOBELIUM.

Lawrencium. Lawrencium (Lr, atomic number 103, named after Ernest O. Lawrence) was discovered in 1961 by Ghiorso, Sikkeland, A. E. Larsh, and R. M. Latimer using the HILAC at the University of California, Berkeley. A few micrograms of a mixture of ^{249}Cf, ^{250}Cf, ^{251}Cf, and ^{252}Cf (produced in a nuclear reactor) were bombarded with ^{10}B and ^{11}B ions to produce single atoms of an isotope of element 103 with a half-life measured as 8 s and decaying by the emission of alpha particles of 8.6 MeV energy. Ghiorso and coworkers suggested at that time that this radioactivity might be assigned the mass number 257. G. N. Flerov and coworkers have disputed this discovery on the basis that their later work suggests a greatly different half-life for the isotope with the mass number 257. Subsequent work by Ghiorso and coworkers proves that the correct assignment of mass number to the isotope discovered in 1961 is 258, and this later work gives 4 s as a better value for the half-life.

All known isotopes of lawrencium (mass numbers 253–260) are short-lived and are produced by bombardment of lighter elements with charged particles (heavy ions); chemical investigations have been, and presumably must be, performed on the tracer scale. Work with ^{260}Lr (half-life 3 min) has demonstrated that the normal oxidation state in aqueous solution is the III state, corresponding to the ion Lr^{3+}, as would be expected for the last member of the actinide series. *See* LAWRENCIUM.

Rutherfordium. Rutherfordium (Rf, atomic number 104, named after Lord Rutherford), the first transactinide element to be discovered, was probably first identified in a definitive manner by Ghiorso, M. Nurmia, J. Harris, K. Eskola, and P. Eskola in 1969 at Berkeley. Flerov and coworkers have suggested the name kurchatovium (named after Igor Kurchatov with symbol Ku) on the basis of an earlier claim to the discovery of this element. In 1964 they bombarded ^{242}Pu with ^{22}Ne ions in their cyclotron at the Joint Institute for Nuclear Research in Dubna and reported the production of an isotope, suggested to be ^{260}Ku, which was held to decay by spontaneous fission with a half-life of 0.3 s. After finding it impossible to confirm this observation, Ghiorso and coworkers reported definitive proof of the production of alpha-particle-emitting ^{257}Rf and ^{259}Rf (half-lives 4.5 and 3 s, respectively), demonstrated by the identification of the previously known ^{253}No and ^{255}No as decay products, by means of the bombardment of ^{249}Cf with ^{12}C and ^{13}C ions in the Berkeley HILAC.

All known isotopes of rutherfordium (mass numbers 253–262) are short-lived and are produced by bombardment of lighter elements with charged heavy-ion particles. The isotope ^{261}Rf (half-life 78 s) has made it possible, by means of rapid chemical experiments, to demonstrate that the normal oxidation state of rutherfordium in aqueous solution is the IV state corresponding to the ion Rf^{4+}. This is consistent with expectations for this first "transactinide" element which should be a homolog of hafnium, an element that is exclusively tetrapositive in aqueous solution. Gas chromatographic studies of volatile halides and halides oxides of rutherfordium demonstrate characteristics as expected for a group-4 element in the periodic table. However, some detailed chemical properties of rutherfordium compounds, studied in the aqueous phase and the gas phase, resemble more the behavior of its lighter homolog zirconium (atomic number 40) than hafnium (atomic number 72). *See* RUTHERFORDIUM.

Dubnium. Dubnium (Db, atomic number 105, named after the Dubna Laboratory), the second transactinide element to be discovered, was probably first identified in a definitive manner in 1970 by Ghiorso, Nurmia, K. Eskola, Harris, and P. Eskola at Berkeley. They reported the production of alpha-particle-emitting ^{260}Db (half-life 1.6 s), demonstrated through the identification of the previously known ^{256}Lr as the decay product, by bombardment of ^{249}Cf with ^{15}N ions in the Berkeley HILAC. Again the Berkeley claim to discovery is disputed by Flerov and coworkers, who earlier in 1970 reported the discovery of an isotope thought to be dubnium, decaying by the less definitive process of spontaneous fission, produced by the bombardment of ^{243}Am with ^{22}Ne ions in the Dubna cyclotron; in later work Flerov and coworkers may have also observed the alpha-particle-emitting isotope of dubnium reported by Ghiorso and coworkers.

The known isotopes of dubnium (mass numbers 256–263) are short-lived and are produced by bombardment of lighter elements with charged heavy-ion particles. Using rapid chemical techniques and the isotope ^{262}Db (half-life 40 s), it is possible to

study the chemical properties of dubnium. The results show that dubnium exhibits the V oxidation state like its homolog tantalum. A number of chemical studies demonstrate that, in specific chemical environments, dubnium behaves more like its lighter homolog niobium or sometimes like the pentavalent actinide element protactinium (atomic number 91). *See* DUBNIUM.

Seaborgium. The discovery of seaborgium (Sg, atomic number 106, named after Glenn T. Seaborg) took place in 1974 simultaneously as the result of experiments by Ghiorso and coworkers at Berkeley and Flerov, Y. T. Oganessian, and coworkers at Dubna. The Ghiorso group used the SuperHILAC (the rebuilt HILAC) to bombard a target of californium (the isotope ^{249}Cf) with ^{18}O ions. This resulted in the production and positive identification of the alpha-particle-emitting isotope ^{263}Sg, which decays with a half-life of 0.9 ± 0.2 s by the emission of alpha particles of a principal energy of 9.06 MeV. The definitive identification consisted of the establishment of the genetic link between the seaborgium alpha-particle-emitting isotope (^{263}Sg) and previously identified daughter (^{259}Rf) and granddaughter (^{255}No) nuclides, that is, the demonstration of the decay sequence: A total of 73 ^{263}Sg alpha particles and approximately the expected corresponding number of ^{259}Rf daughter and ^{255}No granddaughter alpha particles were recorded.

The Dubna group chose lead (atomic number 82) as their target because, they believed, its closed shells of protons and neutrons and consequent small relative mass leads to minimum excitation energy for the compound nucleus and therefore an enhancement in the cross section for the production of the desired product nuclide. They bombarded ^{207}Pb and ^{208}Pb with ^{54}Cr ions (atomic number 24) in their cyclotron to find a product that decays by the spontaneous fission mechanism (a total of 51 events), with the very short half-life of 7 ms, which they assign to the isotope ^{259}Sg. Later work at Dubna and the Gesellschaft für Schwerionenforschung (GSI) laboratory in Darmstadt, Germany, has shown that this assignment is not correct. ^{259}Sg is an alpha emitter of 0.48 s. In 1984 the isotopes ^{261}Sg and ^{260}Sg were discovered at GSI, Germany.

Long-lived isotopes ^{266}Sg (21 s) and ^{265}Sg (22 s) were discovered in 1994 at Dubna bombarding ^{248}Cm with ^{22}Ne. In 2000, spontaneous fissioning isotopes ^{258}Sg (2.9 ms) and ^{262}Sg (6.9 ms) were synthesized at GSI using ^{290}Bi- and ^{207}Pb-based reactions. Thirty years after discovery of the element seaborgium, 8 isotopes are known.

Fast chemical separations, performed with the Automated Rapid Chemistry Apparatus (ARCA) and On-Line Gas Chromatography Apparatus (OLGA) on a one-atom-at-a-time scale of ^{265}Sg, allowed investigating seaborgium in aqueous solution and in the gas phase. A series of experiments were done in international collaborations at GSI, Darmstadt. As expected from its projected position in the periodic table, seaborgium shows chemical properties similar to those extrapolated from its lighter group-6 homologs, molybdenum (atomic number 42) and tungsten (atomic number 74). Seaborgium is the heaviest element that has been studied in aqueous solution. *See* SEABORGIUM.

Bohrium. Bohrium (Bh, atomic number 107, named after Niels Bohr) was synthesized and identified by G. Münzenberg and coworkers at GSI-Darmstadt. The element has a long half-life in the millisecond range, indicating an unexpected relative stability in this region of high atomic numbers. In the discovery experiment, a target of ^{209}Bi [located in the region of closed shells ($N = 126$)] was used, which when bombarded with heavy ions, led to a compound nucleus of minimum excitation energy ($\sim$15 MeV), allowing for a cooling of the nucleus by emission of a single neutron. The element was identified by the alpha-particle-emitting isotope ^{262}Bh, and the genetic links with its known alpha-particle-emitting descendants was established, as was done by Ghiorso and coworkers in their discovery of seaborgium. In 1981, Münzenberg and coworkers observed six atoms of 5 ms (the time interval for its decay) ^{262}Bh, produced by the ^{209}Bi (^{54}Cr,n) reaction. *See* NUCLEAR STRUCTURE.

The experiment was repeated, and in 1988 about 40 decay chains corroborated the discovery of bohrium. The isotope ^{261}Bh was discovered in the same reaction by the 2-neutron emission channel.

Today the isotope ^{264}Bh, a member of the decay-chain of roentgenium, is known and two isotopes, ^{266}Bh and ^{267}Bh, synthesized in the reaction of ^{22}Ne with ^{249}Bk, have been discovered. ^{265}Bh with a half-life of 0.94 s was synthesized at the Institute of Modern Physics, Lanzhou, China, in 2004. All these isotopes manifest the trend towards larger half-lives in the region of seconds.

The one and so far only chemical separation and characterization of bohrium was done at the Paul-Scherrer Institute (PSI), Villigen, Switzerland, during gas chromatography of a chloride oxide compound of ^{267}Bh. As expected, bohrium showed chemical properties similar to those extrapolated from its lighter group-7 homologs, technetium (atomic number 43) and rhenium (atomic number 75). *See* BOHRIUM.

Hassium. Hassium (Hs, atomic number 108, named after the German state of Hessen from the Latin word Hassias) was synthesized in 1984 at GSI, Darmstadt, Hessen. A target nucleus of ^{208}Pb, an isotope with two closed nucleon shells ($Z = 82$, $N = 126$) was fused with ^{58}Fe to synthesize ^{265}Hs; see also bohrium for aspects of the hassium production mechanism. The element was identified by its links to known descendants. Three atoms of 1.8 ms ^{265}Hs were produced by the ^{208}Pb(^{58}Fe,n) reaction. The reaction was confirmed in many laboratories and serves as a calibration for Pb- and Bi-based reactions. Later the isotope ^{264}Hs was discovered by the same group in the reaction ^{207}Pb(^{58}Fe,n). It links the elements with even proton numbers via a bridge of an α-decay at ^{256}Rf to lighter elements and establishes a connection of absolute binding energies up to element hassium.

The isotope ^{267}Hs was discovered at Dubna in 1992. It was observed in a 5n-channel in the fusion of ^{34}S and ^{238}U. This isotope ^{269}Hs was confirmed at GSI, being found as a descendant in the α-decay chain of ^{271}Ds. The isotope ^{269}Hs was observed first in 1996 at GSI in a decay chain of ^{277}112 and has been reproduced since in the reaction ^{248}Cm(^{26}Mg,5n) by experiments at GSI, observing its decay after a chemical hassium separation. In the reaction of magnesium-26 with curium-248, the isotope ^{270}Hs was synthesized. This is the first isotope with the closed nuclear shell $N = 162$. With increasing neutron number, a trend toward larger half-lives in the few-seconds range is observed also for hassium isotopes.

Since 2001, studies of the formation and the behavior of hassium tetroxide were done one-atom-at-a-time with the isotope ^{269}Hs in three international collaborations at GSI. All the experiments showed chemical properties of hassium similar to those of osmium (atomic number 76), the lighter homolog in group-8 of the periodic table. *See* HASSIUM.

Meitnerium. Meitnerium (Mt, atomic number 109, named after Lise Meitner) was synthesized in 1982 at GSI. (For its production see also Bohrium above.) A target nucleus of ^{209}Bi, as done before for bohrium, was fused with ^{58}Fe to synthesize ^{266}Mt. Again the element was identified by its genetic links to known descendants. In 1982, one atom of 1.7 ms ^{266}Mt was produced by the ^{209}Bi(^{58}Fe,n) reaction. Since then, the discovery has been confirmed by about 10 more decay chains.

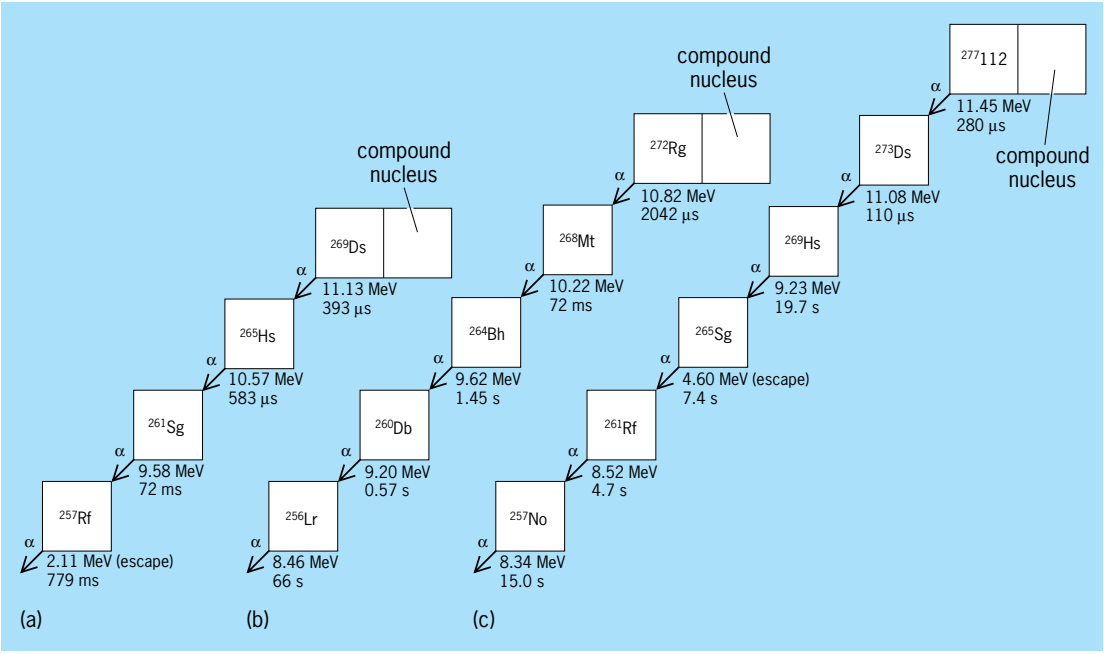
The isotopes ^{268}Mt and ^{270}Mt were observed as descendants in decay chains from heavier elements. ^{268}Mt was discovered at GSI, in 1994 in the discovery experiment of the element roentgenium. ^{270}Mt was reported from RIKEN, Japan, in 2004, together with the discovery of element 113 in the reaction ^{209}Bi(^{70}Zn,n). Half-lives stayed in the range below 1 second and the production cross section of 5×10^{-38} cm^2 is very small (the smallest cross section known, to date).

Meitnerium is expected to have chemical properties similar to those of iridium, its lighter homolog in group 9 of the periodic table. No chemical experiments have been done on meitnerium. *See* MEITNERIUM.

Darmstadtium. Darmstadtium (Ds, atomic number 110, named after Darmstadt, Germany, the location of the GSI laboratory) was synthesized in 1994 at GSI.

Darmstadtium should be a heavy homolog of the elements platinum, palladium, and nickel. It is the eighth element in the 6d shell.

Research for this element began in 1985. Experiments at Dubna, Russia; at GSI, and at Lawrence Berkeley Laboratory, Berkeley, California, failed to provide reliable evidence for a successful synthesis. However, at GSI on November 9, 1994, a decay chain was observed that proved the existence of the isotope ^{269}Ds (the isotope of darmstadtium with mass number 269). The isotopes were produced in a fusion reaction of a nickel-62 projectile with a lead-208 target nucleus. The fused system, with an excitation energy of 13 MeV, cooled down by emitting one neutron and forming ^{269}Ds, which by sequential alpha decays transformed to ^{265}Hs, ^{261}Sg, ^{257}Rf, and ^{253}No (nobelium-253). All these daughter isotopes were already known, and four decay chains observed in the following 12 days corroborated without any doubt the discovery of the element. **Illustration** *a*



Decay chains that document the discoveries of new elements. The sequence of alpha decays is shown for each element. Numbers below boxes are alpha energies and correlation times. (*a*) Darmstadtium, produced in the reaction ^{62}Ni + ^{208}Pb → ^{269}Ds + 1n. (*b*) Roentgenium, produced in the reaction ^{64}Ni + ^{209}Bi → ^{272}Rg + 1n. (*c*) Element 112, produced in the reaction ^{70}Zn + ^{208}Pb → ^{277}112 + 1n.

shows the first decay chain observed, which ended in ^{257}Rf. The isotope ^{269}Ds has a half-life of 0.2 ms and is produced with a cross section of about 3×10^{-36} cm^2.

A second isotope, ^{271}Ds, was produced in a subsequent 12-day experiment by fusion of nickel-64 and lead-208. Nine atoms, with an excitation energy of 12 MeV, were produced. They were transformed by sequential alpha decay to the known isotopes ^{267}Hs, ^{263}Sg, ^{259}Rf, and ^{255}No (nobelium-255). The half-life of ^{271}Ds is 1.1 ms, and its production cross section amounts to 1.5×10^{-35} cm^2.

The methods used to produce element Ds were the same as those already used to synthesize the three preceding elements, Bh, Hs, and Mt. Improved beam intensity and quality, improvement of the detection efficiency, and a new detector system allowing nearly complete chain reconstruction made possible the discovery after an extensive search for the optimum bombarding energy. The total sensitivity for finding a new species was increased by a factor of 20.

Two additional isotopes, ^{270}Ds and ^{273}Ds, are known. ^{273}Ds was discovered as a descendant in the α-decay chain of ^{277}112 in 1996. In the fusion of ^{207}Pb and ^{64}Ni, the isotope ^{270}Ds with a half-life of 0.1 ms was synthesized at GSI in 2001. *See* DARMSTADTIUM.

Roentgenium. Roentgenium (Rg, atomic number 111, named after W. Roentgen) was synthesized in 1994 at GSI. Roentgenium should be a homolog of the elements gold, silver, and copper. It is the ninth element in the 6*d* shell.

The element was discovered on December 17, 1994, by detection of the isotope ^{272}Rg, which was produced by fusion of a nickel-64 projectile and a bismuth-209 target nucleus after the fused system was cooled by emission of one neutron. The optimum bombarding energy for producing ^{272}Rg corresponds to an excitation energy of 15 MeV for the fused system. Sequential alpha decays to ^{268}Mt, ^{264}Bh, ^{260}Db, and ^{256}Lr (lawrencium-256) allowed identification from the known decay properties of ^{260}Db and ^{256}Lr. In the decay chain in illus. *b*, the first three members are new isotopes. The isotope ^{272}Rg has a half-life of 1.5 ms, and is produced with a cross section of 3.5×10^{-36} cm^2. Altogether, three chains were observed during the 17 days of irradiation. The methods used to produce roentgenium were the same as those used in the discovery of darmstadtium. *See* ROENTGENIUM.

Element 112. Element 112 should be a heavy homolog of the elements mercury, cadmium, and zinc. It is expected to be the last element in the 6*d* shell. The element was discovered on February 9, 1996, at GSI by detection of the isotope ^{277}112, which was produced by fusion of a zinc-70 projectile and a lead-208 target nucleus following the cooling down of the fused system by emission of a single neutron. The fused system was observed at an excitation energy of 12 MeV. Sequential alpha decays to ^{273}Ds, ^{269}Hs, ^{265}Sg, ^{261}Rf, and ^{257}No (nobelium-257) allowed unambiguous identification by using the known decay properties of the last three members of the chain. In the decay chain in illus. *c*, the first three members are new isotopes. The isotope ^{277}112 has a half-life of 0.24 ms, and it is produced with a cross section of 0.5×10^{-36} cm^2. The new isotopes of Ds and Hs are of special interest. Their half-lives and alpha energies are very different, as is characteristic of a closed-shell crossing. At the neutron number $N = 162$, a closed shell was theoretically predicted, and this closed shell is verified in the decay chain observed. The isotope ^{269}Hs has a half-life of 9 s, which is long enough to allow studies on the chemistry of this element. The methods used to produce element 112 were the same as those used for the two preceding elements, Ds and Rg. The decay chain of the new element was observed in an irradiation time of about 3 weeks. The cross section measured is the smallest observed in the production of heavy elements.

The crossing of the neutron shell at $N = 162$ is an important achievement in the field of research on superheavy elements. The stabilization of superheavy elements is based on high fission barriers, which are due to corrections in the binding energies found near closed shells. The shell at $N = 162$ is the first such shell predicted, and is now verified. Next in line are the predicted shells at proton number $Z = 114$ and neutron number $N = 184$. *See* ELEMENT 112.

Peter J. Armbruster; M. Schädel; Glenn T. Seaborg

Bibliography. G. R. Choppin and J. Rydberg, *Nuclear Chemistry: Theory and Application*, 1980; R. Eichler et al., Chemical characterization of bohrium (element 107), *Nature*, 407:63–65, 2000; Ch. E. Düllmann et al., Chemical investigation of hassium (element 108), *Nature*, 418:859–862, 2002; V. I. Goldanski and S. M. Polikanov, *The Transuranium Elements*, 1973; S. Hofmann et al., The new element 111, *Z. Phys. A*, 350:281–282, 1995; S. Hofmann et al., The new element 112, *Z. Phys. A*, 354:229–230, 1996; S. Hofmann et al., Production and decay of ^{269}110, *Z. Phys. A*, 350:277–280, 1995; S. Hofmann, *On Beyond Uranium*, 2002; C. Keller, *The Chemistry of the Transuranium Elements*, 1971; Max Planck Society for the Advancement of Science, *Transurane-Transuranium Elements*, 1975; J. V. Kratz, Critical evaluation of the chemical properties of the transactinide elements, *Pure Appl. Chem.*, 75:103–138, 2003; M. Schädel, Aqueous chemistry of transactinides, *Radiochim. Acta*, 89:721–728, 2001; M. Schädel, *The Chemistry of Superheavy Elements*, 2003; M. Schädel et al., Chemical properties of element 106 (seaborgium), *Nature*, 388:55–57, 1997; G. T. Seaborg, The new elements, *Amer. Sci.*, 68:3, 1980; G. T. Seaborg, *Transuranium Elements: Products of Modern Alchemy*, 1978; G. T. Seaborg and W. D. Loveland, *The Elements Beyond Uranium*, 1990.

Traps in solids

Localized regions in a material that can capture and localize an electron or hole, thus preventing the electron or hole from moving through the material until

supplied with sufficient thermal or optical energy. Traps in solids are associated with imperfections in the material caused by either impurities or crystal defects. *See* BAND THEORY OF SOLIDS; CRYSTAL DEFECTS; HOLE STATES IN SOLIDS.

Imperfections that behave as traps are commonly distinguished from imperfections that behave as recombination centers. If the probability for a captured electron (or hole) at the imperfection to be thermally reexcited to the conduction (or valence) band before recombination with a free hole (or free electron) is greater than the probability for such recombination, then the imperfection is said to behave like an electron (or hole) trap. If the probability for a captured electron (or hole) at the imperfection to recombine with a free hole (or free electron) is greater than the probability for being thermally reexcited to the band, the imperfection is said to behave like a recombination center. In the equilibrium state in the dark, the occupancy of all imperfections is described in terms of a Fermi distribution centered on the equilibrium Fermi level. In the nonequilibrium condition of photoexcitation, the occupancy of a recombination center is determined by recombination kinetics involving the capture cross sections of the imperfection for free carriers. Since the occupancy of an electron (or hole) trap, however, is determined by a quasiequilibrium thermal exchange with the conduction (or valence) band and not by recombination in the steady state under photoexcitation as well as in the dark equilibrium state, the electron occupancy of a trap under steady-state photoexcitation can still be described in terms of a Fermi distribution, but now centered on the quasi (steady-state) Fermi level. It is possible for a specific chemical or structural imperfection in the material to behave like a trap under one set of conditions of temperature and light intensity, and as a recombination center under another.

Traps play a significant role in many phenomena involving photoconductivity and luminescence. In photoconductors, for example, the presence of traps decreases the sensitivity and increases the response time. Their effect is detectable through changes in the rise and decay transients of photoconductivity and luminescence, thermally stimulated conductivity and luminescence in which the traps are filled at a low temperature and then emptied by increasing the temperature in a controlled way, electron spin resonance associated with trapped electrons with unpaired spins, and a variety of techniques involving the capacitance of a semiconductor junction such as photocapacitance and deep-level transient spectroscopy. *See* ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY; LUMINESCENCE; PHOTOCONDUCTIVITY; THERMOLUMINESCENCE.

Richard H. Bube

Bibliography. R. H. Bube, *Photoconductivity of Solids,* 1960, reprint 1978; R. H. Bube, *Photoelectronic Properties of Semiconductors,* 1992; S. W. S. McKeever, *Thermoluminescence of Solids,* 1985.

Trauma

Injury to tissue by physical or chemical means. Mechanical injury includes abrasions, contusions, lacerations, and incisions, as well as stab, puncture, and bullet wounds. Trauma to bones and joints results in fractures, dislocations, and sprains. Head injuries are often serious because of the complications of hemorrhage, skull fracture, or concussion.

Thermal, electrical, and chemical burns produce severe damage partly because they coagulate tissue and seal off restorative blood flow. Asphyxiation, including that caused by drowning, produces rapid damage to the brain and respiratory centers, as well as to other organs.

Frequent complications of trauma are shock, the state of collapse precipitated by peripheral circulatory failure, and also hemorrhage, infection, and improper healing. *See* SHOCK SYNDROME.

Edward G. Stuart; N. Karle Mottet

Bibliography. B. A. Landon and J. D. Goodall, *An Atlas of Trauma Management*: *The First Hour*, 1993.

Traveling-wave tube

A microwave electronic tube in which a beam of electrons interacts continuously with a wave that travels along a circuit, the interaction extending over a distance of many wavelengths. Traveling-wave tubes can provide amplification over exceedingly wide bandwidths. Typical bandwidths are 10–100% of the center frequency, with gains of 20–60 dB. Low-noise traveling-wave tube amplifiers serve as the inputs to sensitive radars or communications receivers. High-efficiency medium-power traveling-wave tubes are the principal final amplifiers used in communication satellites, the space shuttle communications transmitter, and deep-space planetary probes and landers. High-power traveling-wave amplifiers operate as the final stages of radars, wide-band radar countermeasure systems, and scatter communication transmitters. They are capable of delivering continuous-wave power levels in the kilowatt range and pulsed power levels exceeding a megawatt. *See* COMMUNICATIONS SATELLITE; ELECTRONIC WARFARE; RADAR; SPACE COMMUNICATIONS; SPACE PROBE.

Forward-wave amplifiers. In a forward-wave, traveling-wave tube amplifier (**Fig. 1**), a thermionic cathode produces the electron beam. An electron gun initially focuses the beam, and an additional focusing system retains the electron stream as a beam throughout the length of the tube until the beam is captured by the collector electrode. The microwave signal to be amplified enters the tube near the electron gun and propagates along a slow-wave circuit. The tube delivers amplified microwave energy into an external matched load connected to the end of the circuit near the collector. The slow-wave circuit serves to propagate the microwave energy along the tube at approximately

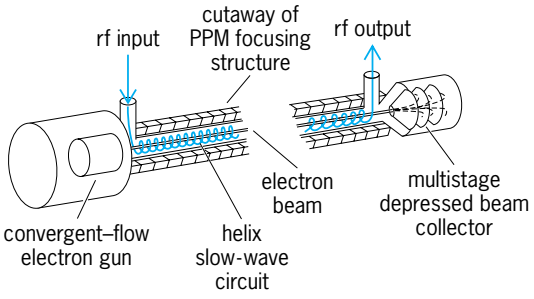


Fig. 1. Periodic-permanent-magnet (PPM) focused traveling-wave tube.

the same velocity as that of the electron beam. Interaction between beam and wave is continuous along the tube with contributions adding in phase.

Velocity and current modulations of the electron beam occur because the waves and the electrons travel in near synchronism and the amplification process takes place continuously. In the increasing electric field region of each cycle of the wave the electrons are accelerated slightly, and in the decreasing field regions the electrons are decelerated slightly. This leads to electron bunches forming with charge density in the bunches increasing with distance. The increasing charge density in the bunches induces in turn an electric field on the helix that continuously grows with distance. The electron beam is injected into the helix with a velocity slightly faster than the waves. In the ensuing interaction between fields and waves, the energy lost by the average deceleration of the electrons is the source of energy for the growing waves on the circuit. Because of the continuously distributed interaction, the power in the circuit wave grows exponentially with distance along the tube.

There are four basic elements of the traveling-wave tube: an electron gun, a means for focusing the electron beam, a slow-wave circuit, and an electron-beam collector.

Electron gun. Most modern traveling-wave tubes require electron beams, with current densities many times higher than can be achieved directly from a cathode surface. Thus, the gun design must draw the emitted electrons from the thermionic cathode and converge them into a much smaller cross-sectional area to achieve the required beam current density (**Fig. 2**).

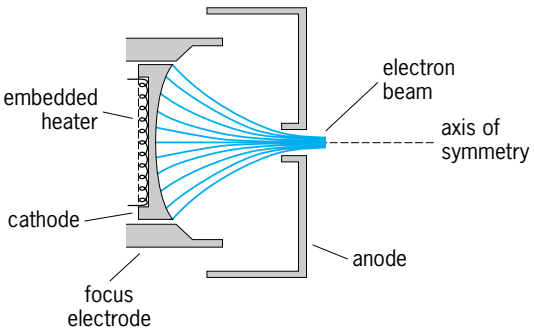


Fig. 2. Convergent-flow electron gun.

Cathode technology has continuously improved to provide high-current-density, long-life cathodes. These cathodes can produce an emitted current density of 1.0 A/cm^2 with a lifetime greater than 100,000 h. With convergent-flow electron guns and permanent-magnet focusing technology, a continuous beam current density of 75 A/cm^2 can be achieved.

Focusing methods for beam. There are two principal methods for focusing the electron beam through the length of the slow-wave circuit. The solenoid electromagnet and the periodic-permanent-magnet (PPM) focusing schemes. The former is rarely used except for special applications where consumption of power in the solenoid is acceptable. The periodic-permanent-magnet system has several advantages: (1) it consumes no power; (2) it concentrates the magnetic fields principally in the region of the electron beam and tends to cancel the fields external to the magnet system; and (3) as a result it can be made very compact and lightweight. It therefore has a great advantage for airborne and space-borne applications where weight and power consumption are important considerations.

The periodic-permanent-magnet structure can be thought of as a stack of magnetic lenses using ring magnets and disk pole pieces with alternate cells arranged in opposite polarity. This provides an axial magnetic field whose amplitude varies sinusoidally with distance. If the electron beam is introduced into this arrangement correctly, the beam can be focused over long distances with greater than 99.9% of the electron beam emerging from the end of the slow-wave circuit structure without being intercepted on it. The development of high-coercive-force permanent-magnet materials has allowed the achievement of high peak magnetic field strengths on the axis of the structure and as a result has allowed the achievement of excellent focusing of high-current-density electron beams. *See* MAGNETIC LENS.

Slow-wave circuit. In order for the electromagnetic signal wave to travel along the tube at a velocity approximately equal to the electron-beam velocity, the signal must be guided by a slow-wave circuit. Beam velocity requirements are typically 2–10% of the velocity of a free-space electromagnetic wave. The signal must be slowed down to this velocity at all wavelengths within the bandwidth of the tube. It is important that the axial electric field strength produced in the region of the electron beam be strong for a given power flow along the circuit. Certain circuits have this property in conjunction with a constant or slowly varying phase velocity as a function of frequency.

A helix is the simplest and best slow-wave circuit for low- and medium-power traveling-wave tubes and constitutes the most widely used structure (**Fig. 3*a***). It can be supported by ceramic rods within a metal vacuum envelope, and heat dissipated in the helix can be removed by conduction through

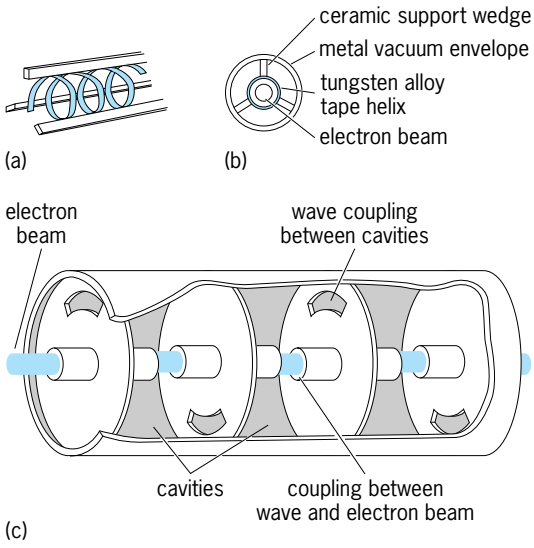


Fig. 3. Slow-wave circuits for traveling-wave tubes. (*a*) Helix circuit for medium-power tube. (*b*) Cross-sectional view of helix circuit. (*c*) Coupled-cavity circuit for high-power tube.

the rods and vacuum envelope into the surrounding magnet structure and subsequently into a heat sink.

Helix tubes have delivered up to 2500 W of continuous output power at frequencies as high as 8 GHz. For higher power output, coupled-cavity slow-wave circuits are generally used (Fig. 3*b*). This circuit has a narrower bandwidth than does a helix, 10% of center frequency being typical.

Microwave energy is usually coupled into and out of the helix through a window in the wall of the vacuum envelope of the tube. The window is usually ceramic and is designed to allow an impedance match from the external transmission system to the helix. This is usually accomplished with a direct coaxial connection, but higher-power devices often use a waveguide with a ceramic window and an internal coupling scheme to the helix or other slow-wave circuit. The coupling scheme must be impedance matched over the desired operating bandwidth to eliminate reflections. This requires very careful design. *See* COAXIAL CABLE; IMPEDANCE MATCHING; TRANSMISSION LINES; WAVEGUIDE.

To provide a stable amplifier, the input and output sections of the circuit must be isolated from each other. This is accomplished in helix-type tubes by applying lossy material to the ceramic support wedges near the center of the tube. This attenuation must have an extremely low reflection coefficient viewed from either direction not only so that stability is achieved but also so that fluctuations in gain versus frequency can meet exacting requirements. The attenuation completely absorbs the microwave signal at the downstream end of the input helix. The input and output sections of the helix are coupled by the electron beam that reexcites the growing wave in the output section due to the electron bunching that has occurred in the input section. In high-power coupled-cavity circuits, loss-impregnated ceramics are placed in selected cavities. When the lossy cavities are separated by sections of lossless circuit with 20 dB of gain, stability is good with only negligible degradation of efficiency.

Beam collector. In applications where efficiency is not important, the collector electrode and the slow-wave circuit are often connected to the same dc potential for power-supply simplicity.

A dramatic improvement in efficiency can be achieved by operating the collector electrodes at a potential that is lower than the helix. Thus, energy can be returned to the power supply as the electrons decelerate into the collector and are captured at a lower potential. The interaction that takes place between the electrons and the signal wave produces a spread in the electron velocities at the exit of the helix. Some electrons are speeded up and some are slowed down. The potential of the depressed collector can be lowered only near to the point where the slowest electrons reverse direction and return to the helix. Much effort has been expended in developing multistage collectors that allow electrons to be collected on a series of electrodes at successively lower potentials. The maximum depression and the greatest improvement in device efficiency are achieved in this manner. On the basis of the total power input to the device, overall tube efficiency of over 60% has been achieved. This technique is used extensively in satellite and deep-space applications where spacecraft power is limited and overall device efficiency is of utmost importance.

Backward-wave devices. A class of traveling-wave tubes exists called backward-wave devices in which energy on a slow-wave circuit flows in the direction opposite to the travel of electrons in the beam.

O-type oscillators. Although amplifiers are possible, the principle use of this technique is to create a voltage-tunable microwave oscillator. Typically it uses a hollow, linear electron beam and a helix circuit designed to emphasize the backward-wave fields. This represents the earliest type of voltage-tunable microwave oscillator. It is capable of generating power levels of 10–100 milliwatts with a tuning range of 2:1 in frequency. Its use has almost disappeared with the development of magnetically tuned microwave transistor oscillators using yttrium-iron-garnet (YIG) spherical resonators. *See* FERRIMAGNETIC GARNETS.

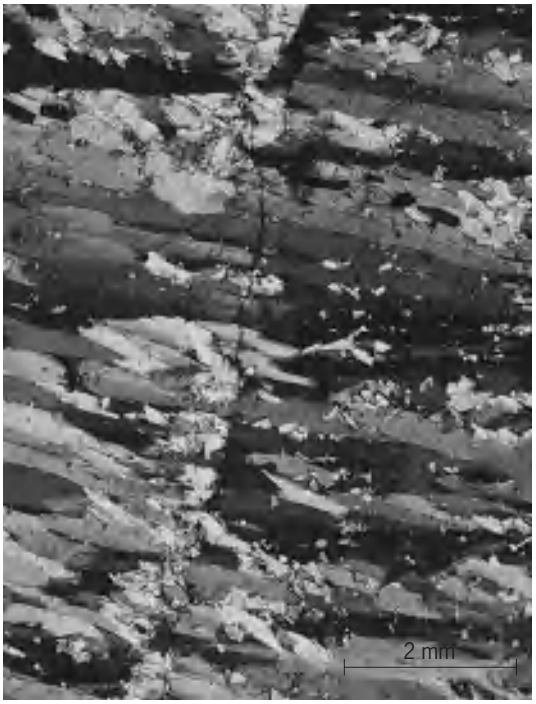
M-type oscillators. An M-type backward-wave oscillator is similar in principle to the O-type, except that focusing and interaction are through magnetic fields, as in magnetrons. Efficiency of M-type tubes is considerably higher than that of O-type tubes, typical efficiencies being 20–30%. Noise and spurious output power are also greater in the M-type tube. A continuous-wave output power of several hundred watts is typical. *See* MAGNETRON; MICROWAVE TUBE; OSCILLATOR.

Lester A. Roberts

Bibliography. A. S. Gilmour, Jr., *Microwave Tubes*, 1986; A. S. Gilmour, Jr., *Principles of Traveling Wave Tubes*, 1994; J. F. Gittens, *Power Traveling-Wave Tubes*, 1963.

Travertine

A rather dense, banded limestone (see **illus.**), sometimes moderately porous, that is formed by evaporation about springs, as is tufa, or in caves as stalac-



Travertine, Suisun, California. (*From E. W. Heinrich, Microscopic Petrography, McGraw-Hill, 1956*)

tites, stalagmites, or dripstone. Where travertine or tufa (calcareous sinter) is deposited by hot springs, it may be the result of the loss of carbon dioxide from the waters as pressure is released upon emerging at the surface; the release of carbon dioxide lowers the solubility of calcium carbonate, which precipitates. High rates of evaporation in hot-spring pools also lead to supersaturation. Travertine formed in caves is simply the result of complete evaporation of waters containing mainly calcium carbonate. *See* LIMESTONE; STALACTITES AND STALAGMITES; TUFA.

Raymond Siever

Tree

A perennial woody plant at least 20 ft (6 m) in height at maturity, having an erect stem or trunk and a well-defined crown or leaf canopy. However, no sharp lines can be drawn between trees, shrubs, and lianas (woody vines). For example, the strangler fig (*Ficus aurea*) is a climbing liana which may develop into a self-supporting tree if the host around which it twines is removed. Many large trees, such as paper birch (*Betula papyrifera*) or Alaska cedar (*Chamaecyparis nootkatensis*), become prostrate shrubs at the northern limits of their range in the boreal region or at their altitudinal limits near timberline.

Despite lack of agreement over an operational definition, the essence of the tree form is relatively large size, long life, and a slow approach to reproductive maturity. The difficulty of transporting water, nutrients, and storage products over long distances and high into the air against the force of gravity is a common problem of large treelike plants and one that is not shared by shrubs or herbs.

Age and size. Trees are the oldest and most massive living things on the Earth. Some of the bristlecone pine (*Pinus aristata*) in the mountains of California are approximately 4600 years old, though they are only small, gnarled trees clinging tenaciously to life in a rigorous alpine environment. These trees were 1000 years old in the biblical days of David and Solomon. Alerce (*Fitzroya cupressoides*) of Chile seem to be the second-oldest trees in the world; the most ancient was 3613 years old when it was cut in 1975. The redwoods (*Sequoia sempervirens*) reach an age of only about 2200 years but are the tallest trees in the world with a record height of 368 ft (112 m). The "big trees" (*Sequoiadendron giganteum*) are not as tall but are larger in diameter. The General Sherman tree, for example, has a diameter of 30 ft (9 m) at 4.5 ft (1.5 m) above the ground, weighs 6000 tons (5442 metric tons), and may be 3500 years old. Many other conifers of the Pacific Northwest frequently reach 200 ft (60 m) in height, and the eucalypts (especially *Eucalyptus regnans*) of Australia are nearly as large as the redwoods. In the eastern United States the oldest living trees are the bald cypress (*Taxodium distichum*), which may reach 1700 years of age. Most tree species, however, have a life-span of less than 200 years and heights under 100 ft (30 m). Girth or diameter is not a good indicator of age.

Zonation and distribution. With some exceptions (for example, the Joshua tree, *Yucca brevifolia*), trees are generally found in aggregates called forest stands. Because trees are the most noticeable component of the global flora, the forest cover is used to categorize world vegetation types. Forests are found from the tropics to the boreal regions and occupy one-half of the total land area. Precipitation and temperature are the major factors limiting growth of forests. In general, forests are restricted to areas where precipitation exceeds 25 in. (62.5 cm) per year. Trees are not found in the Arctic tundra and, except along water courses, they are excluded from the dry steppes of Russia, the Great Plains of the United States, and desert regions of the world. Timberline, the zone in alpine areas above which trees are not found, runs 13,000–14,000 ft (4000–4300 m) in the tropics, 10,000–12,000 ft (3000–4000 m) in the Sierra Nevada, 6000 ft (1800 m) in the Alps, 5000 ft (1500 m) in New England, and 1000–3000 ft (300–900 m) in southern Alaska. Timberline has fluctuated in concert with climate change; for example, in the Sierra Nevada region it has changed by as much as 230 ft (70 m) during the last 6300 years. *See* ALTITUDINAL VEGETATION ZONES.

As one travels from the boreal forest to the Equator, the variety of arborescent species increases from

a few forms in the subarctic to about 200 per acre in the tropics. Malaysia and Amazonia are the richest areas of the world in number of tree species. In the United States the southern Appalachians provided a refuge for migrating species during successive glacial periods and as a result have a diversity of tree forms. On the other hand, there is a relative paucity of species in the European flora. European forms became extinct when they were trapped between ice sheets advancing from the north and other glaciers originating in the Alps.

Unfortunately, the rate of deforestation is accelerating. Forest diversity, particularly in the tropics, is threatened by deforestation. On average, over 38 million acres (15 million hectares) of tropical forest was cleared and converted to other uses annually during 1981–1990.

Classification. Almost all existing trees belong to the seed plants (Spermatophyta). An exception are the giant tree ferns which were more prominent in the forests of the Devonian Period and today exist only in the moist tropical regions, where they grow to heights of 60 ft (18 m). The Spermatophyta are divided into the Pinophyta (gymnosperms) and the flowering plants, Magnoliophyta (angiosperms). The gymnosperms bear their seed naked on modified leaves, called scales, which are usually clustered into structures called cones—for example, pine cones. By contrast the seed of angiosperms is enclosed in a ripened ovary, the fruit. *See* MAGNOLIOPHYTA; PINOPHYTA; POLYPODIALES.

The orders Cycadales, Ginkgoales, and Pinales of the Pinophyta contain trees. *Ginkgo biloba*, the ancient maidenhair tree, is the single present-day member of the Ginkgoales. The Cycadales, characteristic of dry tropical areas, contain many species which are small trees. The Pinales, found throughout the world, supply much of the wood, paper, and building products of commerce. They populate at least one-third of all existing forest and include the pines (*Pinus*), hemlocks (*Tsuga*), cedars (*Cedrus*), spruces (*Picea*), firs (*Abies*), cypress (*Cupressus*), larches (*Larix*), Douglas-fir (*Pseudotsuga*), sequoia (*Sequoia*), and other important genera. The Pinales are known in the lumber trade as softwoods and are popularly thought of as evergreens, although some (for example, larch and bald cypress) shed their leaves in the winter. *See* CEDAR; CYCADALES; CYPRESS; DOUGLAS-FIR; FIR; HEMLOCK; LARCH; PAPER; PINALES; PINE; PINOPHYTA; SEQUOIA; SPRUCE.

In contrast to the major orders of gymnosperms which contain only trees, many angiosperm families are herbaceous and include trees only as an exception. Only a few are exclusively arborescent. The major classes of the angiosperms are the Liliopsida (monocotyledons) and the Magnoliopsida (dicotyledons). The angiosperm trees, commonly thought of as broad-leaved and known as hardwoods in the lumber market, are dicotyledons. Examples of important genera are the oaks (*Quercus*), elms (*Ulmus*), maples (*Acer*), and poplars (*Populus*). *See* ELM; LILIOPSIDA; MAGNOLIOPSIDA; MAPLE; OAK; POPLAR.

The Liliopsida contain few tree species, and these are never used for wood products, except in the round as posts. Examples of monocotyledonous families are the palms (Palmae), yucca (Liliaceae), bamboos (Bambusoideae), and bananas (Musaceae). *See* BAMBOO; BANANA.

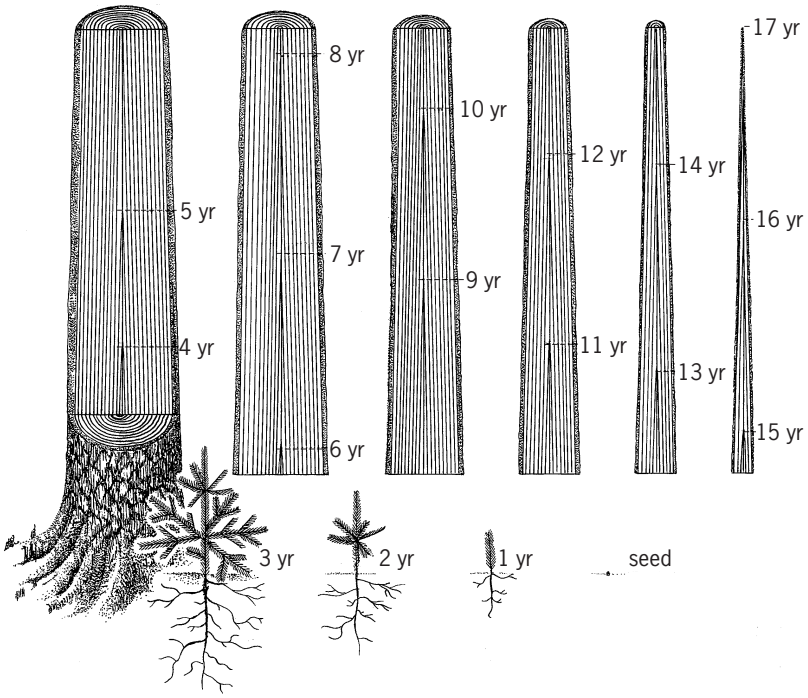
Morphology and physiology. The morphology of a tree is similar to that of other higher plants. Its major organs are the stem, or trunk and branches; the leaves; the roots; and the reproductive structures. Almost the entire bulk of a tree is nonliving. Of the trunk, branches, and roots, only the tips and a thin layer of cells just under the bark are alive. Growth occurs only in these meristematic tissues. Meristematic cells are undifferentiated and capable of repeated division. *See* FLOWER; LATERAL MERISTEM; LEAF; PLANT GROWTH; ROOT (BOTANY); STEM.

Growth. Height is a result of growth only in apical meristems at the very tips of the twigs. A nail driven into a tree will always remain at the same height, and a branch which originates from a bud at a given height will never rise higher. The crown of a tree ascends as a tree ages only by the production of new branches at the top and by the death and abscission of lower, older branches as they become progressively more shaded. New growing points originate from the division of the apical meristem and appear as buds in the axils of leaves. *See* APICAL MERISTEM; BUD; PLANT GROWTH.

In the gymnosperms and the dicotyledonous angiosperms, growth in diameter occurs by division in only a single microscopic layer, three or four cells wide, which completely encircles and sheaths the tree. This lateral meristem is the cambium. It divides to produce xylem cells (wood) on the inside toward the core of the tree and phloem cells on the outside toward the bark. In trees of the temperate regions the growth of each year is seen in cross section as a ring. Because of this yearly increment of xylem elements, the tree structure develops as conical shells over shells (see **illus.**). *See* BARK; PHLOEM; XYLEM.

Xylem elements become rigid through the thickening and modification of their cell wall material. The tubelike xylem cells transport water and nutrients from the root through the stem to the leaves. In time the xylem toward the center of the trunk becomes impregnated with various mineral and metabolic products, and it is no longer capable of conduction. This nonfunctional xylem is called heartwood and is recognizable in some stems by its dark color. The light-colored, functional outer layer of the xylem is the sapwood. *See* WOOD ANATOMY.

The phloem tissue transports dissolved carbohydrates and other metabolic products manufactured by the leaves throughout the stem and the roots. Most of the phloem cells are thin-walled and are eventually crushed between the bark and the cambium by the pressures generated in growth. The outer bark is dead and inelastic but the inner bark contains patches of cork cambium which produce new bark. As a tree increases in circumference, the old outer bark splits and fissures develop, resulting in the rough appearance characteristic of the trunks of most large trees.



Schematic drawing of a 17-year-old coniferous tree showing the manner in which the trunk increases in thickness through the addition of annual increments. (*After A. J. Panshin et al., Textbook of Wood Technology, vol. 1, 2d ed., 1964*)

In the monocotyledons the lateral cambium does not encircle a central core, and the vascular or conducting tissue is organized in bundles scattered throughout the stem. The trunk is not wood as generally conceived although it does in fact have secondary xylem.

Phylogeny and evolution. Trees have a very ancient history on Earth, first being recorded in Devonian strata deposited 300 million years ago (Ma). These trees included fernlike plants which bore sporangia rather than seeds. The first confirmed seed plants, the seed ferns (pteridosperms), appeared in the Mississippian 250 Ma and are the probable ancestors of the gymnosperms. The seed ferns vanished after the Jurassic Period. During the Pennsylvanian, dense lowland forests of giant horsetails (*Equisetum*) and club mosses (*Lycopsida*), now extinct, produced the world's coal deposits. Later in this era the Cycadales and Ginkgoales were abundant, and the ancestors of the Pinales were also present. In the late Paleozoic, 190 Ma, early conifers were fossilized. The first recognizable angiosperms were trees and appeared in the Mesozoic Era, more than 150 Ma. Few angiosperms are represented in the Jurassic flora, although some of those present, such as sycamore (*Platanus*), tulip poplar (*Liriodendron*), and sweet gum (*Liquidambar*), are still recognizable today. During the Cretaceous Period, 100 Ma, the angiosperms increased greatly in numbers. Angiosperms undoubtedly evolved from ancient gymnosperms. All the vast number of herbaceous angiosperms known today have in turn evolved from the early flowering trees. *See* EQUISETALES; LEPIDODENDRALES; PTERIDOSPERMS; TREE FERNS.

The tree form with its upright stem 100–300 ft (30–90 m) in height may have been evolutionarily successful because of the advantage it conferred in competition for sunlight. On the other hand, the longer interval between generations may mean that tree populations cannot adapt to a changing environment through the genetic process of recombination and selection as rapidly as short-lived plants. Also, trees must be physiologically capable of surviving under the varying climatic conditions experienced from year to year and cannot afford to adapt to a specific set of conditions as can annual plants. However, all modern seed-bearing plants have evolved from tree ancestors. The trees themselves provide products necessary to humanity's continued well-being. These products range from medicines and other chemicals through paper to major construction materials. *See* DENDROLOGY; FOREST AND FORESTRY; FOREST TIMBER RESOURCES; PALEOBOTANY; PLANT EVOLUTION; PLANT PHYSIOLOGY; PLANT TAXONOMY; TREE DISEASES.

F. Thomas Ledig

Bibliography. R. R. Aicher, *Growth Stresses and Strains in Trees*, 1986; W. M. Harlow, E. S. Harrar, and F. M. White, *Textbook of Dendrology*, 8th ed., 1995; M. R. Sethuraj and A. S. Raghavendra (eds.), *Tree Crop Physiology*, 1987; M. H. Zimmerman and C. L. Brown, *Trees: Structure and Function*, 1985.

Tree diseases

Diseases of both shade and forest trees have the same pathogens, but the trees differ in value, esthetics, and utility. In forests, disease is significant only when large numbers of trees are seriously affected. Diseases with such visible symptoms as leaf spots may be alarming on shade trees but hardly noticed on forest trees. Shade trees with substantial rot may be ornamentals with high value, whereas these trees would be worthless in the forest. Emphasis on disease control for the same tree species thus requires a different approach, depending on location of the tree.

Diseases of forest trees. From seed to maturity, forest trees are subject to many diseases. Annual losses of net sawtimber growth from disease (45%) are greater than from insects and fire combined. Young, succulent seedlings, especially conifers, are killed by certain soil-inhabiting fungi (damping-off). Root systems of older seedlings may be destroyed by combinations of nematodes and such fungi as *Cylindrocladium, Sclerotium*, and *Fusarium*. Chemical treatment of seed or soil with formulations containing nematicides and fungicides, and cultural practices unfavorable to root pathogens help to avoid these diseases.

Roots rots are caused by such fungi as *Heterobasidion* (= *Fomes*) *annosus* (mostly in conifers) and *Armillariella* (= *Armillaria*) *mellea* (mostly in hardwoods). These fungi cause heart rot in the roots and stems of large trees and also invade and kill young, vigorous ones. Thinning pine plantations increases

infection by *H. annosus*, which invades fresh stumps and grows through root grafts to surrounding trees. Losses are minimized by strategic timing of thinning or by stump treatment with borax, urea, or the fungus *Peniophora gigantea*. In the southern United States thinnings are made in summer when temperatures are too high for spore development. *Armillariella mellea* invades and destroys roots and stems of trees predisposed by other stress factors. Both *H. annosus* and *A. mellea* occur worldwide in temperate zones.

Leaf diseases and wilt. In natural forests, leaf diseases are negligible, but in nurseries and plantations, fungal infections cause severe defoliation, retardation of height growth, or death. *Scirrhia acicola* causes brown spot needle blight and prevents early height growth of longleaf pine in the South; it defoliates Christmas tree plantations of Scotch pine (*Pinus sylvestris*) in northern states. Fungicides and prescribed burning are used successfully for control. In Christmas tree plantations, control is possible with maneb or chlorothalonil. *Dothistroma* needle blight is severe in plantations of Austrian pine (*P. nigra*) and ponderosa pine (*P. ponderosa*), and in plantations of Monterey pine (*P. radiata*) in Australia, New Zealand, Africa, and South America. In North America, *Lophodermium pinastri* killed millions of pine seedlings in the 1970s. Maneb and chlorothalonil are effective for control when new ascospores are being released.

Oak wilt is a systemic disease, with the entire tree affected through its water-conducting system. Trees of susceptible species may be killed in a few weeks to a year or more through plugging of the water-conducting vessels. The causal fungus, *Ceratocystis fagacearum*, spreads to nearby healthy trees by root grafts and to trees at longer distances by unrelated insects, including bark beetles of the Nitidulidae. The sporulating mats of the fungus develop between bark and wood, producing asexual and, sometimes, sexual spores, which are disseminated by insects attracted by aromatic odors of the fungus. Control is possible by eradicating infected trees and by disruption of root grafts by trenching or by chemicals.

Rust diseases and cankers. Stem rust diseases occur as cankers or galls on coniferous hosts and as minor lesions on other ones. A few, such as white pine blister rust (**Fig. 1**) and southern fusiform rust, are epidemic, lethal, and economically important. Others of less immediate importance (such as western gall rust) are capable of serious, widespread infection, requiring no other hosts to complete their life cycles. Most rust fungi require two unrelated hosts, such as the currant and white pine for blister rust. Eradication of hosts of secondary importance is cost-prohibitive, and resistant varieties are favored for control. Other control measures include pruning out early infections and spraying nursery trees with chemicals during periods favoring needle infection.

Except for stem rust, cankers of conifers are generally of minor importance. However, scleroderris canker is a new, serious threat to pine stands (especially hard pines) in Canada and the United States. The causal fungus (*Gremeniella abietina*) is favored by cold weather, and is epidemic and lethal in forest reproduction, nurseries, plantations, and young stands. Control in nurseries is possible by successive applications of chlorothalonil; pruning and sanitation are recommended for plantations.

Fig. 1. White pine blister rust (*Cronartium ribicola*) on white pine. The blisters are masses of aeciospores that are wind-disseminated for many miles to *Ribes* plants. (*Photograph by Robert Campbell*)

Stem infections by numerous fungi, resulting in localized death of cambium and inner bark, range from lesions killing small stems in a year (annual) to gross stem deformities (perennial), where cankers enlarge with stem growth. Chestnut blight, first known in the United States in 1904, destroyed the American chestnut as a commercial species (**Fig. 2**), and is an example of the annual lesion type. The less dramatic or devastating *Nectria* canker destroys stems of timber value, and is an example of the perennial lesion type. Large, living chestnut trees are rare in the United States, but disease-resistant roots maintain the chestnut as a shrub or small tree in native stands in the eastern hardwood forest. Resistant varieties of chestnut lack the superior qualities of the native American species. Hope for biological control of chestnut blight is in newly discovered hypovirulent strains of the blight fungus that prevent the killing of trees infected with virulent ones. The hypovirulent strains may be themselves diseased by a viruslike agent of double-stranded ribonucleic acid.

Tree decay. All tree species, including decay-resistant ones such as redwood, are subject to ultimate disintegration by fungi. Decay fungi (Hymenomycetes) are associated with nondecay fungi (Deuteromycetes) and bacteria. These microflora enter the tree through wounds, branch stubs, and roots, and are confined to limited zones of wood

Fig. 2. Chestnut blight in American chestnut tree (*USDA*). Insert is mycelial fan of chestnut blight caused by *Endothia parasitica*, advancing through bark of American chestnut. The tip of the fan on the left is surrounded by cortical tissue. The contents of the cortical cells back from the tip of the fan are discolored to a yellowish brown, as indicated by the darkened cells (*from J. S. Boyce, Forest Pathology, 3d ed., McGraw-Hill, 1961*).

by anatomical and wound-stimulated tissue barriers (**Fig. 3**). The extent of decay is limited by compartmentalization of decay in trees. Trees aged beyond maturity are most often invaded by wood-rotting fungi; losses can be minimized by avoiding wounds and by shortening cutting rotations. Losses from rot are especially serious in overmature coniferous stands in the western United States, Canada, and Alaska. *Fomes pini*, the cause of white pocket rot in most commercial conifers, and *Echinodontium tinctorium*, the cause of stringy brown butt rot in fir and hemlock, are two of the most destructive heart rot fungi. The processes of discoloration and decay in trees that reduce the quality of wood are initiated by wounds. If the tree response to wounding by chemical and physiological reactions is fast, compartmentalization will limit damage to a small discolored zone without decay. But if the tree response is slow, bacteria and nondecay fungi will predispose the wood to decay by the wood decay fungi. Soon after wounding, the cambium is stimulated to produce a special layer of cells that prevents invasion of newly formed tissue by internal microflora. Thus, new wood rings are free of defect. *See* WOOD DEGRADATION.

Fig. 3. Shelf fungus (*Fomes applanatus*) on a dead aspen tree. This wood-rotting fungus enters through roots and wounds and decays the heartwood and sapwood of both hardwoods and softwoods.

Dwarf mistletoe disease. Dwarf mistletoes are small seed plants that are obligate parasites on coniferous trees; they cause losses of more than 3×10^9 board feet of lumber annually. More than 20 species occur on millions of acres of western conifers, and on spruce in the eastern United States. Once limited by natural fires, they have increased in distribution and severity with better fire control. Dwarf mistletoe plants may be $^1/_2$ in. (13 mm) to several inches in height and are leafless and scalelike. They absorb nutrients and water from host plants through haustoria and sinkers in bark and wood. They depend on their hosts for most of their food and produce fruit with seeds that are forcibly discharged up to 30 ft (9 m). Limited seed dispersal distance results in circular infection patterns around source trees; spread is slow except for occasional transmission by birds.

Miscellaneous diseases. Another series of tree diseases is recognized for which there is no satisfactory explanation. Such diseases as birch dieback, little-leaf disease of shortleaf pine, pole blight of western white pine, and ohia decline in Hawaii represent complex combinations of common symptoms and multiple causal factors, defying positive diagnosis. One or more known pathogenic organisms that appear likely as potential first causes are later regarded only as secondary agents. Several adverse environmental factors cause stress, predisposing trees to colonization by weakly pathogenic fungi. These diseases reflect the degree to which human-induced change of environment may be regarded as a primary cause of tree disease. *See* REFORESTATION.

Disease of shade trees. Many shade trees are grown under conditions for which they are poorly adapted, and are subject to environmental stresses not common to forest trees. Both native and exotic trees planted out of natural habitats are predisposed to secondary pathogens following environmental stress of noninfectious origins. They are also susceptible to the same infectious diseases as forest trees. Appearance is more important than the wood produced, and individual value is higher per tree than for forest trees. Thus, disease control methods differ from those recommended for forest trees.

Fungal diseases. The most important and destructive shade tree disease known is Dutch elm disease, introduced from Europe to North America before 1930 (**Fig. 4**). Elms represented millions of prized shade trees in monoculture plantings in the urban

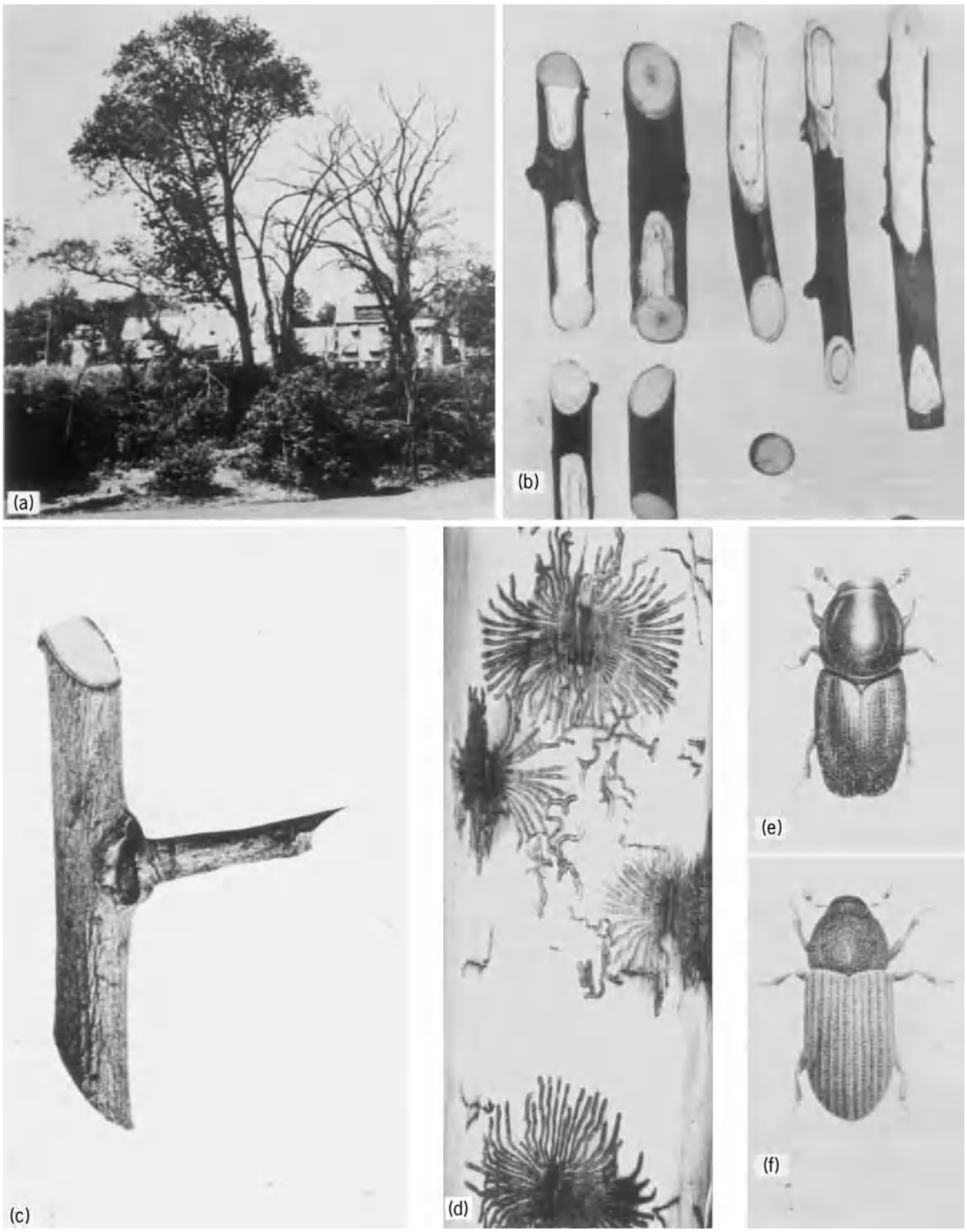


Fig. 4. Dutch elm disease. (*a*) Group of trees affected by the disease. (*b*) Discoloration in the sapwood of infected trees. (*c*) Feeding scar in small elm crotch made by an adult of the smaller European elm bark beetle (*Scolytus multistriatus*). (*d*) Brood galleries made by female beetles and larvae. (*e*) European elm bark beetle, the most important carrier of the Dutch elm disease. (*f*) Native elm bark beetle. (*a, c, d, USDA*; *b, e, f, Michigan State University*)

United States and Canada, and losses and control efforts cost millions of dollars annually. The causal fungus, *Ophiostoma ulmi* (*Ceratocytis ulmi*), is introduced to the water-conducting system of healthy elms by the smaller European elm bark beetle (*Scolytus multistriatus*) or the American elm bark beetle (*Hylurgopinus rufipes*). One or more new and more aggressive strains of the fungus have arisen since 1970. More devastating than the original ones, they are destroying the elms in North America and Europe that survived earlier epidemics. Effective means of prevention are sanitation (destroying diseased and dying and dead elm wood); insecticidal sprays (methoxychlor); disruption of root systems; and early pruning of new branch infections. Of much promise are resistant varieties of elm, systemic fungicides (such as solubilized benzimidazoles), and insect pheromones. Reports (subject to confirmation)

Fig. 5. Elm infected with bacterial wetwood. The bacterial ooze is coming from a wound.

on infectious agents (such as bacteria and viral particles) inside the Dutch elm disease fungus, as well as antifungal bacteria, offer hope for biological control.

Bacterial diseases. The most common bacterial disease of shade trees is wetwood (**Fig. 5**) of elm and certain other species. It is reported to be caused by a single bacterial species (*Erwinia nimipressuralis*), although causal associations of other bacteria are now suspected. The bacteria are normally present in the heartwood of mature elms and cause no disease unless they colonize sapwood by exterior wounds. Fermented sap under pressure bleeds from wounds and flows down the side of the tree. Sustained bleeding kills underlying cambial tissue. Internal gas pressure and forced spread of bacterial toxins inside living tissues of the tree can be reduced by strategic bleeding to avoid seepage into bark and cambium.

A second bacterial disease of elm, elm yellows (= elm phloem necrosis), is caused by a mycoplasmalike organism, considered to be a unique kind of bacterium. Elm yellows is as lethal as Dutch elm disease but is more limited in distribution. Foliage of infected trees turns yellow and falls off; infected trees die. Diagnostic symptoms are browning of inner bark near the base of the tree and wintergreen odor of affected tissue. The pathogen is carried by the elm leafhopper (*Scaphoideus luteolus*), which sucks phloem juice from leaf veins. Spread of disease also occurs through grafted root systems. Control measures include early destruction of infected trees, disruption of root systems, and insecticidal sprays. Injection with tetracycline and other antibiotics helps to slow the progress of the bacterium.

Leaf spot and heart rot. Leaf diseases of minor importance on forest trees often appear alarming or ominous on shade trees. These diseases are always present and become conspicuous following a wet spring. They are mostly cosmetic in impact, are rarely serious, and may be minimized by elimination of fallen leaves and timely fungicidal sprays in the early spring.

Heart rot is of minor importance in shade trees. Many trees with extensive heart rot remain healthy for decades. Nothing will stop this internal decay, but exposure to air may decrease its rate; sealing with cavity work may accelerate the decay.

Noninfectious disease. The most common and complex diseases of shade trees are diebacks and declines (such as maple decline). Many species show similar patterns of symptoms caused by multiple factors, but no single causal factor is known to cause any one of these diseases. Noninfectious agents of shade tree diseases are drought, soil compaction, mineral deficiency, soil pollution from waste or salt, air pollution, and so on. Trees affected experience chlorosis, premature fall coloration and abscission, tufting of new growth, dwarfing and sparseness of foliage, progressive death of terminal twigs and branches, and gradual to rapid decline in growth. Such trees are often infested with borers and bark beetles, and infected by branch canker (such as *Cytospora* and *Nectria* spp.) and root rot fungi (such as *Armillariella mellea*). Noninfectious stress predisposes trees to infectious disease that is caused by different kinds of weakly parasitic fungi as secondary pathogens. *See* PLANT PATHOLOGY.

Richard J. Campana

Tree ferns

Plants belonging to the families Cyatheaceae and Dicksoniaceae, whose members typically develop tall trunks crowned with leaves (fronds) which often reach some 20 ft (6 m) in length and 5–6 ft (1.5–1.8 m) in width (**Figs. 1** and **2**). Tree ferns reach their greatest development in the rainforests and cloud forests of the mountainous tropics. *See* RAINFOREST.

Distribution. Two families and 13 genera are recognized. The Cyatheaceae are represented by the genera *Lophosoria* (tropical America, 1 species); *Metaxya* (tropical America, 1 species); *Sphaeropteris* (tropical America, India and southeastern Asia to New Zealand, the Marquesas, and Pitcairn Island, about 120 species); *Alsophila* (pantropic, about 230 species); *Nephelea* (tropical America, about 30 species); *Trichipteris* (tropical America, about 90 species); *Cyathea* (tropical America, about 110 species); and *Cnemidaria* (tropical America, about 40 species). The other family, Dicksoniaceae, is represented by *Dicksonia* (tropics and southern subtropics in Malaysia, Australasia, America, Hawaii, St. Helena, about 25 species); *Cystodium* (Malaysia, 1 species); *Thyrsopteris* (Juan

Fig. 1. ***Cyathea arborea*** **in the Luquillo Forest, Puerto Rico. (*Courtesy of M. Canoso*)**

Fernández, 1 species); *Culcita* (tropical America, Azores, Malaysia, Australasia, about 7 species); and *Cibotium* (Southeast Asia, Malaysia, Hawaii, Central America, about 12 species).

Stems and leaves. Tree fern trunks may reach 65 ft (19.8 m) in height, as in *Alsophila australis* of Australia. Their diameters vary from 0.4 in. (1.0 cm) in *A. biformis* of New Guinea to 4 ft (1. 2 m) in the *Cibotium* trunks of Hawaii. The lower trunk is often densely covered with matted adventitious roots which greatly increase its diameter. Certain specimens branch near the base of the trunk or higher up; perhaps this branching occurs in response to injury.

Fig. 2. Upper part of the 30-ft (9-m) trunk of ***Cyathea arborea*** **showing the crown of arching 7-ft (2-m) leaves (called fronds). (*Courtesy of T. Plowman*)**

The degree of division of the leaves varies from simple in *A. sinuata* of Ceylon to four or five pinnate in *Lophosoria*. The leaflets (pinnae) are usually smaller toward the base of the leaf; when these basal leaflets are branched into threadlike divisions, they are called aphlebiae.

In most species the old leaves and petiole bases are eventually deciduous, leaving distinctive scars on the trunk. These scars are usually arranged in spirals whose spacing varies with the growth rate. In *Nephelea aureonitens* of Costa Rica, however, the scars are in rings at intervals of several inches. This pattern may relate to the reported occurrence of totally leafless resting periods in this species and certain others. The majority of species are evergreen, however, and produce a new flush of leaves during each wet season.

Scales and hairs. In the Dicksoniaceae and *Lophosoria* and *Metaxya*, simple septate hairs, often quite thick at the base, are found on the leaves and trunks. All species of *Sphaeropteris, Alsophila, Nephelea, Trichipteris, Cyathea*, and *Cnemidaria*, however, have scales, with hairs also on the upper and sometimes the lower leaf surface. Many species of Cyatheaceae are also characterized by variously developed spines on the leaf buds (croziers), petioles, and trunk.

Sori. The Dicksoniaceae have marginal sori terminal on the veins and protected by a bivalved indusium. The Cyatheaceae produce sori well away from the margin, usually seated at the forking of a vein or midway along a simple vein. The indusia may be absent, a small scale, a cup, or a completely enclosing membrane. Sporangia are relatively small and in most cases contain 64 trilete, variously ornamented spores. The indurated annulus is more or less oblique and uninterrupted by the attachment of the stalk.

Anatomy. The typical vascular system of both families is dictyostelic. The Cyatheaceae have accessory vascular strands in the pith and cortex. Fibrous sheaths around the vascular tissue and just inside the epidermis provide mechanical support. The xylem consists of scalariform tracheids and parenchyma, the phloem of sieve tubes and parenchyma. Numerous mucilage canals are embedded in the pith and cortex. *See* STEM.

Fossils. The earliest satisfactory fossils of this group are Jurassic and Cretaceous leaf compressions and trunk petrifactions. Cyatheo-Dicksoniaceous tree ferns had worldwide distribution during the Jurassic and Cretaceous, some occurring as far north as Greenland. In the Tertiary they began to disappear from the northern regions and became restricted to their present tropical distribution. *See* PALEOBOTANY.

Uses. The young leaf buds and pith have been used as food and fermented into alcohol. The woolly hairs of *Cibotium* are used to stanch wounds and stuff cushions. In southern China and Taiwan, apices of

Cibotium barometz are fashioned into "vegetable lambs" and sold to tourists. The trunks, highly resistant to termites and decay, are used in building and fencing. In New Zealand, the trunks are carved into ornamental objects called ponga ware. The adventitious roots are much sought for use in orchid culture, either as solid slabs or broken in potting mixtures.

Classification. Although the tree ferns have traditionally been classified in the Dicksoniaceae and Cyatheaceae, some studies unite them in the inclusive family Cyatheaceae with four subfamilies, and others maintain both families, considering each as an independent evolutionary line. Within the traditional Cyatheaceae, the former genera *Alsophila*, *Hemitelia*, and *Cyathea* have long been recognized as unnatural. Research indicates that the cellular detail of the petiole scales has very strong correlative value in distinguishing natural evolutionary lines. This is reflected in the modern classification of the scaly Cyatheaceae into the genera *Sphaeropteris, Alsophila, Nephelea, Trichipteris, Cyathea*, and *Cnemidaria*. *See* POLYPODIALES; TREE.

Gerald J. Gastony

Bibliography. L. R. Atkinson and A G. Stokey, Comparative morphology of the gametophyte of homosporous ferns, *Phytomorphology*, 14(1):51–70, 1964; F. O. Bower, *The Ferns*, vol. 2, 1926; G. Dunk, *Ferns*, 1994; R. E. Holttum, Cyatheaceae, *Flora Malesiana*, ser. 2, 1(2):65–176, 1963; R. E. Holttum and U. Sen, Morphology and classification of the tree ferns, *Phytomorphology*, 11(4):406–420, 1961; W. R. Maxon, The tree ferns of North America, *The Smithsonian Report for 1911*, pp. 463–491, 1912; R. M. Tryon, The classification of the Cyatheacea, *Contributions from the Gray Herbarium of Harvard University*, vol. 200, 1970.

Tree growth

Trees grow to a larger size at maturity than other woody perennials, have a comparatively long period of development to maturity, and live a long time in the mature state. Trees, like other vascular plants, are made up of cells; growth is the result of adding more cells through cell division, and of the elongation and maturation of those cells into functional tissues. Cells and the tissues they compose differ in structure and function. Some tissues, such as the corky bark, mechanically protect the tree and insulate it against rapid temperature changes. Xylem tissues conduct water, minerals, and some hormones up the tree, and the phloem tissues conduct photosynthate sugar solutions and other organic molecules in most cases down the tree. That fraction of the xylem and phloem which conducts liquids is termed vascular tissue. There are also other nonconducting cells such as parenchyma within the xylem and phloem tissues, which is important in photosynthesis and food storage.

Tissues may also be compared by their origin. Primary tissues are derived from apical meristems of the stem and root, whereas secondary tissues arise from secondary cambia. Two important secondary cambia in trees are the vascular cambium which produces cells which become secondary xylem (the wood of commerce) to the inside and secondary phloem tissue to the outside, and the cork cambium or phellogen which produces corky bark tissues to the ouside and usually phelloderm tissue to the inside. Collectively, these secondary cambia account for most of the diameter growth in trees. *See* PLANT TISSUE SYSTEMS.

Photoperiod has a large influence on the growth of trees. Height growth rate, flowering, date of vegetative budset, and in some species leaf abscission are affected. Plant response is actually to length of the night rather than of the day. Thus, when it is said that growth is promoted by long photoperiods, in fact the plant senses short nights and so continues to grow rapidly. The glycoprotein-pigment complex phytochrome, located in the plasmalemma, is responsible for transducing the night length into the plant's physiological system. The exact nature of the biological action of the phytochrome system is not known. *See* PHOTOPERIODISM.

The photoperiod which affects growth or developmental responses is a function of the photoperiod of the location where the tree species evolved. Trees that have similar photoperiodic responses are termed photoperiodic ecotypes. Research has shown the heritable nature of photoperiodic response, since hybrids of trees from different photoperiodic ecotypes show intermediate responses.

Growth and the cell wall. In vascular plants, growth generally refers to the development of organized structures. Longitudinal growth depends upon cell division in apical meristems and upon subsequent elongation of these cells. Unlike animal cells, plant cells are encased in a rigid cell wall. Growth begins by breaking of chemical bonds in the cellulosic framework of the wall, followed immediately by insertion of new cellulose molecules into the wall. This type of cell-wall growth is termed intussusception and allows the primary cell wall to grow in area while retaining the same thickness. Primary walls are laid down by immature cells during the growth process, but as growth ends, secondary walls are formed in some types of cells (for example, the secondary xylem tracheids and vessels).

Primary cell walls have similar structure in a variety of higher plants. The composition and ultrastructure of secondary walls, however, vary considerably even from one cell type to another. In general, the secondary wall is considerably thicker than the primary wall and is laid down in three layers which differ in orientation of cellulose microfibrils. The secondary wall is formed by adding wall material to the inner surface of the enlarged cell. This is termed apposition and differs from the intussusception growth process of the primary cell wall. During early stages of secondary wall development, lignin precursors synthesized by the cell itself or diffused from the vascular cambium polymerize in spaces between the cellulose microfibrils of the primary and secondary walls to form the rigid, lignified, mature cell wall. *See* CELL WALLS (PLANT).

Meristems. Shoot and root apices, originating in the seed embryo, are major centers of organization in development of the tree. Cell divisions occur in distinct patterns in these apices. These patterns are the basis for development of tissues from which the mature axial structures of the shoot and root are formed. The shoot apex is the source of cells which develop in very predictable patterns into leaf primordia, axillary buds that can grow into lateral branches, and reproductive structures. The root apex is a less complex structure producing only cells which will develop into the internal tissues of the root. Lateral roots are not formed at the apex, but in pericycle tissues some distance behind the apex.

Plant cells are totipotent; that is, they have the total genetic information necessary to program for any cell type, organ, or the entire plant. Cells produced at apices can differentiate into cell types to form different tissues, depending on the chemical and physical environment to which they are exposed. The totipotency concept was proposed at the turn of the century. It was first demonstrated in the 1950s when single cells of carrot were cultured to whole plants. In trees, somatic embryos have been produced successfully in sweetgum, which is an angiosperm. Techniques for inducing somatic embryogenesis in conifers have been developed in Norway spruce (*Picea abies*), white spruce (*Picea glauca*), sugar pine (*Pinus lambertiana*), Douglas-fir (*Pseudotsuga menziesii*), and loblolly pine (*Pinus taeda*). These methods offer possibilities for the efficient vegetative propagation of these species.

Procambial strands differentiate, connecting new leaf and bud primordia to the continuous vascular system of the main stem or branch. In eastern cottonwood (*Populus deltoides*) it has been shown that substances transported from leaves through the mature vascular system affect the development of new leaf primordia and the primary vascular tissues serving them. Research on ponderosa pine (*Pinus ponderosa*) supports the hypothesis that the number of secondary needle fasicles determines the amount of the hormone auxin transported basipetially from the elongating shoot, and that the hormone concentration determines the amount of stored carbohydrates tranported into that shoot to support growth.

Apices and tissues subjacent to them are centers of organization in the plant because cell division and differentiation occur there. Biosynthetic patterns, transport and polar transport tissues, and polar transport gradients create biochemical environments necessary for specific differentiation of the new cells. *See* APICAL DOMINANCE; APICAL MERISTEM.

Lateral meristems and secondary vascular tissues. The vascular cambium is the source of cells that, as they enlarge, cause most of the diameter growth of trees. The vascular cambium is formed from procambial strands. Cell division in the vascular cambium adds cells to both the inside and the outside of the meristem. Cells just inside the cambium differentiate into secondary xylem tissue which primarily conducts water, whereas those outside the cambium develop into secondary phloem tissue (commonly called the inner bark) which conducts a solution of photosynthetically derived sugars. Studies of stem tissues and tissue callus cultures indicate that sucrose levels and levels of the plant hormones auxin (indoleacetic acid), gibberellic acid, and cytokinins control whether xylem or phloem cells develop from cambial derivatives. The sugar-conducting phloem tissue develops in an environment low in auxin, high in gibberellic acid, and high in sucrose, whereas the water-conducting xylem develops at higher auxin, lower gibberellic acid, and moderate sucrose levels. In many species, an early season predominance of xylem production has been reported, while in others primarily phloem is produced early in the season. The diameter of xylem is controlled by the level of auxin present. There is a gradient in auxin concentration from the top of a tree, where the level is high, to the roots, where the level is lower. In the presence of high auxin levels, cells differentiate rapidly; therefore, they do not attain as large a diameter as in locations having lower auxin levels. Thus, there is a gradient in the diameter of xylem from small near the top of the shoot to progressively larger down the shoot and in the roots. *See* AUXIN; GIBBERELLIN; PHLOEM; XYLEM.

Formation of vascular cambium in roots involves a specific balance of an auxin, a cytokinin, a cyclitol, and sucrose. Vascular cambial development obtained by feeding the above mixture to the cut end of excised roots in culture is similar to that found in intact plants.

The amount of annual radial growth in trees is a function of the number of cells produced by the vascular cambium. Fast-growing hemlock (*Tsuga canadensis*) trees have higher rates of cell division and a greater number of cambial initial cells per radial file than do slow-growing trees, and the difference in rate increases during the growing season. Rates of cell elongation, deposition of secondary cell wall material, and lysis of the cytoplasm to form the mature, dead, water-conducting tracheid or vessel do not vary with tree vigor. In the latter part of the growing season, increased levels of auxin, possibly due to modification of auxin transport by increased abscisic acid (ABA) levels in the area adjacent to the cambium where xylem cell maturation is occurring, cause a delay in the onset of lysis of the cytoplasm, resulting in a longer period of secondary cell wall formation. Cells which mature during this period have comparatively thick secondary cell walls and form the latewood.

Hormones in some cases affect the development of angiosperms and gymnosperms differently. Effects of auxin on incorporation of sugar into cell walls of pine (*Pinus sylvestris*) and cottonwood (*Populus deltoides*) are quite different; thus the effects of this hormone on angiosperm and gymnosperm wood formation could also differ. Evidence indicates a probable role of abscisic acid (ABA) in regulating winter dormancy of Douglas-fir and poplar. *See* LATERAL MERISTEM.

The shoot apex of the main stem of conifers and of some dicotyledons consistently outgrows the apices of lateral branches. This causes the conical, or sometimes even spirelike, appearance termed excurrent branching. In contrast, in most dicotyledons, lateral branch apices grow as fast as or faster than the apex of the main shoot, thus causing the main stem to eventually fork repeatedly and form a large spreading crown. This is termed decurrent growth form.

Shoot growth and dormancy. With few exceptions, periods of rapid stem elongation in trees alternate with periods of either very slow or no elongation even in environments favorable for continuous growth.

Tropical trees do not exhibit seasonal growth periodicity as is commonly observed in north temperate species. A few species grow continuously, forming new leaves and stems throughout the year. Intermittently growing tropical trees form true resting buds enclosed in bud scales at the end of each shoot growth period. Some species produce growth flushes only once per year while others have many, and periodicity does not necessarily follow a seasonal pattern. A third type of growth in tropical trees is termed manifold, referring to lack of synchrony between branches on the same tree during a given period.

At germination, seedling trees produce leaf primordia and elongate them coincidentally. This process is termed free growth. For loblolly pine, once free growth ends, the correlation between the subsequent growth of a genetic family to age one and the growth to age eight is much improved. This suggests a change in the genetic control of growth between the free-growth stage and the sequential growth that follows.

During sequential growth, leaf primordia are produced on the edges of the stem apex, and there is a period of delay until they elongate, which can be as short as several days and as long as over winter.

Trees of temperate regions all display intermittent growth. Periods of shoot growth vary with species and genotype, but always occur during periods of favorable temperature and adequate moisture. Shoot-extension patterns of these trees vary with the species but can be grouped into four basic types: (1) A single flush followed by formation of a resting bud that normally remains dormant through the winter (Douglas-fir, *Pseudotsuga menziesii*, and Georgia buckeye, *Aesculus georgiana*). (2) Recurrent growth flushes with bud formation at the end of each successive flush (loblolly pine, *Pinus taeda*, and slash pine, *Pinus elliottii*). (3) A flush of growth followed by shoot-tip abortion (beech, *Fagus grandifolia*, and black willow, *Salix nigra*). (4) A sustained flush of growth with extension of all primordia into leaves prior to bud scale elongation (sweetgum, *Liquidambar styraciflua*, and yellow poplar, *Liriodendron tulipifera*).

Under favorable moisture, temperature, photoperiod, and nutrient conditions, young trees of species that normally exhibit only one growth flush can have

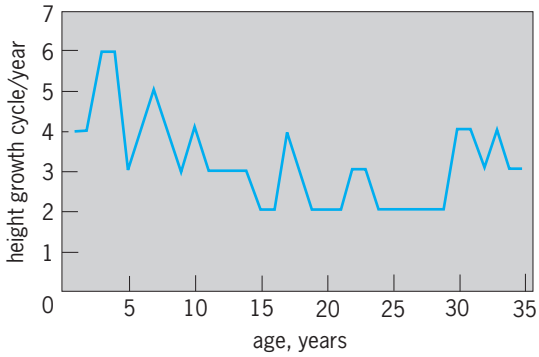


Fig. 1. Number of height growth cycles per year over the first 35 years in a loblolly pine tree. As many as four cycles can occur as late as age 33 in this tree.

two or three growth flushes. This also occurs in older trees, but less frequently (**Fig. 1**). Staining techniques have been developed that allow each growth flush to be identified by marking the crown region that forms at the base of the bud. Such techniques are very useful in allowing retrospective growth analysis on trees that do not develop a whorl of branches at the end of each growth flush.

In trees which normally exhibit only one growth flush per year, the end of the period of shoot extension usually occurs in mid or late summer, a considerable period of time prior to the onset of cool temperatures and frost danger of late fall and winter. After bud set, a period of rapid leaf initiation occurs at the apex, followed by a period of slower leaf initiation. The leaf primordia formed during this period will be elongated the following growing season. In early December (Douglas-fir), cell division in the developing bud ceases and the bud becomes dormant. The buds then require a substantial period of chilling before growth will resume under any environmental conditions. The amount of chilling hours (between 32 and 41°F or 0 and 5°C) required to break dormancy varies among species and between ecotypes within species (Douglas-fir approximately 2000 h and western hemlock, *Tsuga heterophylla*, 660 h).

Physiological processes controlling vegetative bud set, the onset of dormancy, and the effects of chilling on release from dormancy are under intensive study, but are not well understood. Early research indicated the accumulation of a chemical complex, termed beta-inhibitor, during onset of dormancy, and a reduction of this factor during chilling-induced release from dormancy. Subsequent analytical work indicated that ABA was present in the beta-inhibitor. Work on the Douglas-fir in Oregon showed that the concentration of ABA peaked in October and was lowest in the 3 months immediately prior to bud burst. In related studies the ratio of auxin to ABA was at its annual minimum when buds entered dormancy in December. These factors suggest a role for ABA and perhaps auxin in regulating the dormancy cycle in gymnosperms. Studies of hormonal control of dormancy in sycamore maple indicate that neither total ABA nor free or bound ABA concentrations in vegetative buds are correlated with emergence

from dormancy. This suggests that factors other than ABA in the beta-inhibitor complex change during the chilling period of dormancy conditioning prior to emergence from dormancy. If hormonal control of dormancy occurs, then the controlling factor is probably the balance between several compounds rather than simply a change in any one. Differences in mechanism could also exist between angiosperms and gymnosperms. Cell division begins in Douglas-fir in mid-March, a few weeks before budbreak occurs. At this time the chilling requirement has been fulfilled and the photoperiod is lengthening. Relationships between these events have not yet been experimentally defined, but it has been hypothesized that budbreak (extension of the overwintered primordial shoot in the bud) begins when soil temperatures warm to the point where the root system can supply the apical bud with necessary amounts of the hormone gibberellic acid. *See* ABSCISIC ACID; COLD HARDINESS (PLANT); DORMANCY; PLANT HORMONES.

Onset of reproductive competency. Trees usually grow more rapidly and longer when they are young than later when they have achieved a larger size. Research suggests that the cessation of vegetative growth early in the season allows the onset of reproductive competency.

It was suggested in the early 1960s that seasonal fluctuations in date of flushing could be one of the factors controlling the number of female strobili (cone primordia) differentiated on *Pinus banksiana* seedlings. It has been demonstrated that in *Pinus taeda* (loblolly pine), out-of-phase dormancy (induced by prolonging vegetative growth during the winter with long photoperiods and warm temperatures, then in early spring inducing vegetative budset by reducing photoperiod and temperature) will induce development of large numbers of female and male strobili at age 3 years. This finding led to the hypothesis that a critical factor in strobilus production (or, more generally, the onset of reproductive competency in trees) is the slowing of vegetative growth, to allow sufficient time for the comparatively slow process of initiation and differentiation of reproductive structures.

Growth hormones are also known to affect reproductive development. Whereas out-of-phase dormancy treatment on *Pinus taeda* induces production of both male and female strobili, treatment with gibberellin and water stress induced development of only female strobili. Gibberellin also promotes the development of mainly female strobili in most other members of the Pinacae. Such hormone treatment may act primarily by increasing the size of the vegetative bud, but it may also directly affect differentiation.

Treatments inducing earlier reproductive competency in trees are extremely valuable in the genetic improvement of commercial forest trees, as they allow shorter time periods for each breeding cycle. *See* FOREST GENETICS.

Root growth periodicity. Root apices exhibit growth periodicity, but do not appear to have the uniform and predictable dormancy cycles of the shoot. Root dormancy is preceded by cessation of root growth. In many conifers, superficial browning of roots progresses toward the root tip, culminating in formation of a layer of suberized and lignified cells—the metacutinization layer—around the root apex. Patterns vary substantially between species and locations; however, in general, root growth is at its peak for several weeks prior to shoot elongation, slows during shoot elongation, and resumes more rapid growth after shoot elongation has ceased (**Fig. 2**). Root grooth periodicity appears to occur at the individual apex level; that is, one apex can be growing while another is dormant. Sitka spruce (*Picea sitchensis*) roots respond individually to dormancy-inducing stimuli such as exogenous ABA, moisture stress, or nutrient stress. It is thus likely that the mechanism of dormancy resides in individual apices rather than in the shoot.

Correlative inhibition does occur within root systems such that if the main root apex becomes dormant or is destroyed, initiation and growth of lateral branches on that particular root are stimulated. Tree root gross morphology is heavily influenced by correlative inhibition. Initial morphology in trees under 10 years old in several species is taprooted (Douglas-fir, **Fig. 3***a*), whereas when the taproot loses viability, as by penetrating the water table or growing against an impermeable layer, lateral roots develop more rapidly (Fig. 3*b*), finally attaining the mature bell-shaped form (Fig. 3*c*). *See* ROOT (BOTANY).

Correlative relations in shoot and root development. The aboveground portion of the tree—the stem, branches, and leaves—has a geometrical structure well adapted to receiving incident sunlight and taking up carbon dioxide. Carbon dioxide is taken up through stomatal openings in leaves and, in the process, water is unavoidably lost. The belowground portion of the tree, the root system, has a geometrical structure adapted to anchorage and water and mineral nutrient uptake. While the shoot supplies the root with photosynthate, some vitamins such as thiamin, and perhaps the major portion of the hormone auxin, the root supplies the shoot with water, minerals, and some plant hormones, probably including cytokinins. Some studies suggest that some of the gibberellins translocated from the root affect shoot development.

The role of the root in water uptake is passive. The loss of water from the leaves by transpiration causes a tension in the water-conducting xylem stream which pulls water into the root and upward to the leaves. Severing the root and leaving the stem in water will increase water uptake since the resistance in the water conduction path of the root is eliminated. In nature, however, ramification of the root system through the soil is very important as it allows access to soil water. Deep root systems have been shown to be very important in delaying the onset of drought in trees subjected to dry weather. Soil water is usually available only within a few millimeters of the root, although it has been shown that fungal hyphae could extend the zone for mycorrhizal roots. *See* PLANT-WATER RELATIONS.

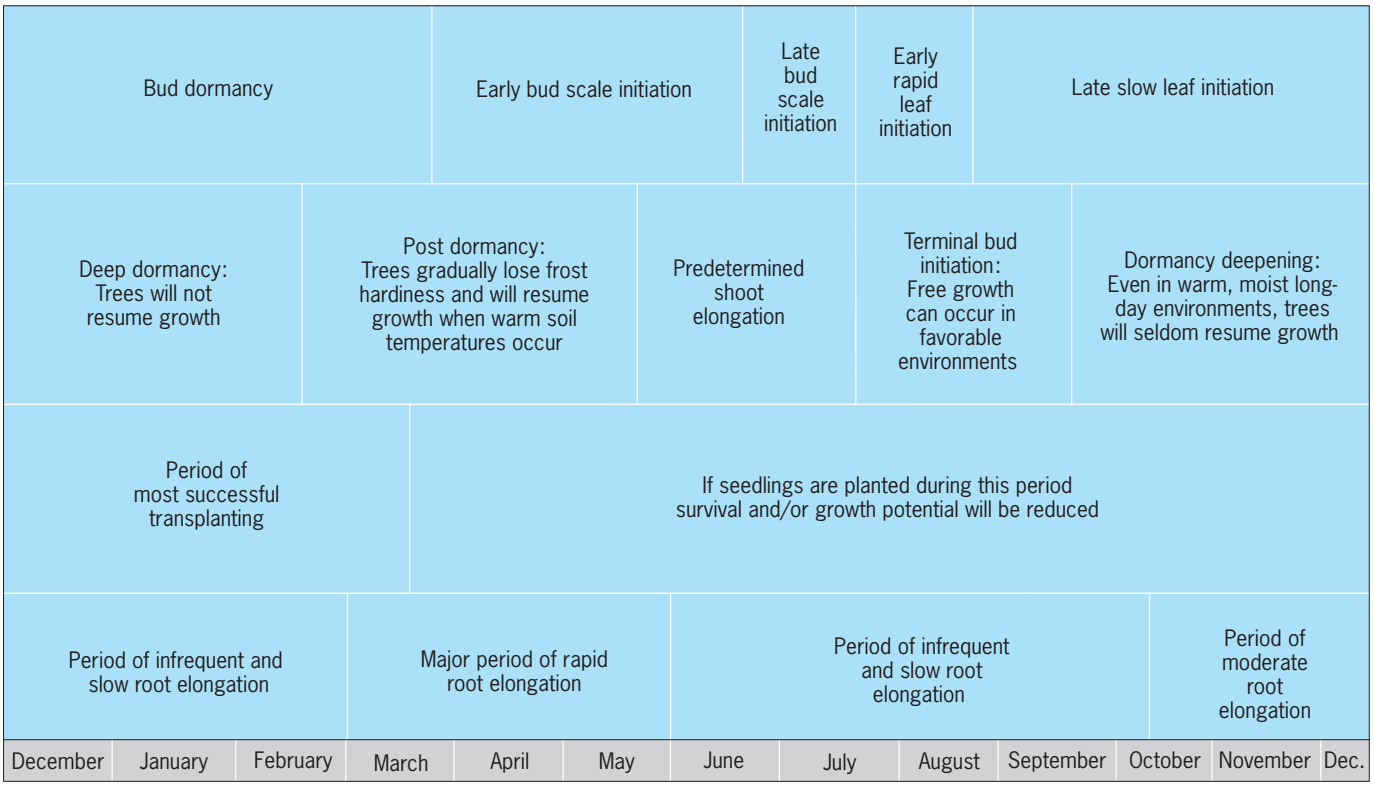


Fig. 2. Growth and dormancy cycle in Douglas-fir.

The interdependence between the shoot and root implies that they must grow in a predictable proportion to each other. The shoot-root dry weight ratio has been most often utilized to study this interdependent growth; however, directly functional parameters would be more descriptive. In many species, the amount of shoot dry matter increases more rapidly than the amount of root dry matter as seedlings grow into large trees. Root growth of peach trees can be limited by pot size. If root growth is limited, then a proportional reduction in shoot growth also occurs. Normal shoot growth can be obtained on such trees by applying the cytokinin 6-benzylamino-purine to the shoot.

Root grafting. It is often assumed that trees in a forest are distinct individuals that compete with one another for water and minerals belowground and light aboveground. However, for many years it has been observed that root grafting between trees of the same species is very common and that grafting between individuals of different species occurs but at a much lower frequency. When a root-grafted tree is cut and other trees of the union left intact, the stump of the severed tree will often remain alive and develop a ring of callus on the cut surface.

Occurrence of root grafts thus suggests that grafted root systems could act as a physiological unit, reducing the degree of belowground competition

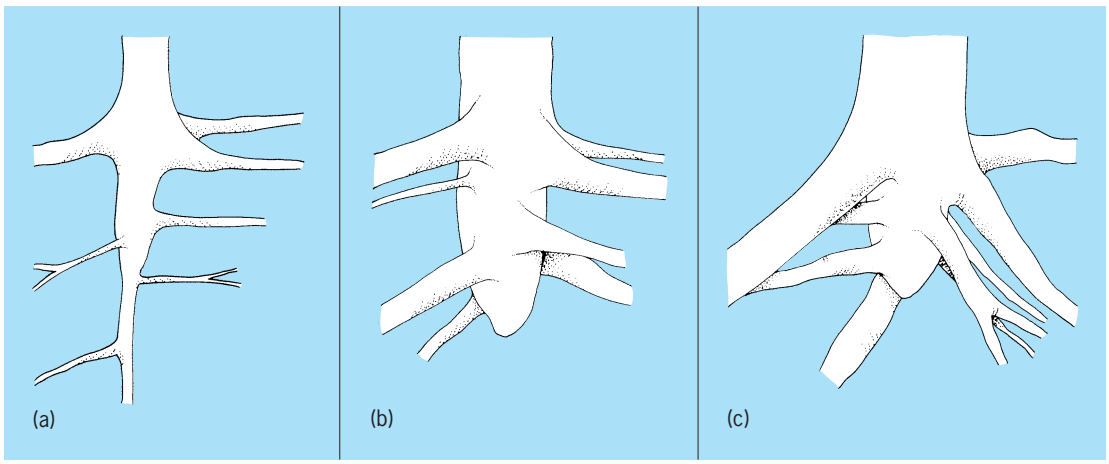


Fig. 3. Development of gross root system morphology in Douglas-fir trees. (*a*) Taprooted juvenile form. (*b*) Taproot dies. (*c*) Bell-shaped mature form.

within grafted groups of trees. In eastern white pine (*Pinus strobus*) the transport of organic substances occurs readily between grafted trees, but water and minerals, although transportable via grafts, might not be exchanged in physiologically significant volumes due to the nature of the graft union which requires water movement radially through xylem cell walls. Red pine (*Pinus resinosa*) root grafts excised from the tree and studied in the laboratory can readily transport water, requiring only a small gradient in xylem water tension to do so. It is apparent that root grafting and its effect on competitive interactions is a little-known area in which fundamental relationships are yet to be defined.

Mycorrhizal symbiosis. Mycorrhizae are symbiotic associations between plants and fungi in which short secondary lateral roots are invaded by specific fungi while the root is actively growing. Only the root cortical tissue is colonized; in ectomycorrhizae, hyphae grow over the surface of the root and between the cortical cells, whereas in endomycorrhizae, hyphae only grow directly into cortical cells. The apical meristem and vascular cylinder of the root remain fungus-free. Root cells retain a healthy appearance in mycorrhizal associations and develop no pathological symptoms. The role of mycorrhizae in ecosystem dynamics is not well described. It is known that mycorrhizal fungi obtain carbohydrates from the tree roots, and in return they increase nutrient absorption by the root, protect the root from some pathogens, and produce plant hormones. Mycorrhizal fungal hyphae can take up and transport water, but the magnitude of this function in nature is undefined. An important role of mycorrhizae is in uptake of phosphorus, which is relatively immobile in the soil. Fungal hyphae from mycorrhizal roots tremendously increase the volume of soil from which phosphorus can be obtained. It is apparent that the mineral nutrition of trees cannot be accurately described without consideration of the role of mycorrhizal fungi. *See* MYCORRHIZAE; PLANT MINERAL NUTRITION.

Certain mycorrhizal fungi allowed tree species to survive and grow well on coal mine spoils in eastern Pennsylvania that formerly could not be revegetated. This knowledge led to new approaches in revegetation of particularly adverse sites. *See* PLANT GROWTH.

William C. Carlson

Bibliography. P. J. Gregory, J. V. Lake, and D. A. Rose, *Root Development and Function: Effects of the Physical Environment*, 1987; L. R. Roberts, P. B. Gahan, and R. Aloni, *Vascular Differentiation and Plant Growth Regulators,* 1988; M. H. Zimmerman, *Xylem Structure and the Ascent of Sap,* 1983.

Tree physiology

The study of how trees grow and develop in terms of genetics; biochemistry; cellular, tissue, and organ functions; and interaction with environmental factors. While many physiological processes are similar in trees and other plants, particularly at the molecular and biochemical levels, trees possess unique physiologies that help determine their outward appearance. These physiological processes include carbon relations (photosynthesis, carbohydrate allocation), cold and drought resistance, water relations, and mineral nutrition.

Three characteristics of trees that define their physiology are longevity, height, and simultaneous reproductive and vegetative growth. Trees have physiological processes that are more adaptable than those in the more specialized annual and biennial plants. Height, exceeding 330 ft (100 m) in some species, allows trees to successfully compete for light, but at the same time this advantage creates transport and support problems. These problems were solved by the evolution of the woody stem which combines structure and function into a very strong transport system. Simultaneous vegetative and reproductive growth in adult trees causes significant competition for carbohydrates and nutrients, resulting in decreased vegetative growth. Trees accommodate both types of growth by having cyclical reproduction: one year many flowers and seeds are produced, followed by a year or more in which few or no flowers are produced.

Carbon relations. While biochemical processes of photosynthesis and carbon assimilation and allocation are the same in trees and other plants, the conditions under which these processes occur in trees are more variable and extreme. In evergreen species, photosynthesis can occur year round as long as the air temperature remains above freezing, while some deciduous species can photosynthesize in the bark of twigs and stem during the winter.

Carbon dioxide fixed into sugars moves through the tree in the phloem and xylem to tissues of high metabolism which vary with season and development. At the onset of growth in the spring, sugars are first mobilized from storage sites, primarily in the secondary xylem (wood) and phloem (inner bark) of the woody twigs, branches, stem, and roots. The sugars, stored as starch, are used to build new leaves and twigs, and if present, flowers. Once the new leaves expand, photosynthesis begins and sugars are produced, leading to additional leaf growth. Activation of the vascular cambium occurs at the same time, producing new secondary xylem and phloem. In late spring, the leaves begin photosynthesizing at their maximum rates, creating excess sugars which are translocated down the stem to support further branch, stem, and root growth. From midsummer through fall until leaf abscission (in deciduous trees) or until temperatures drop to freezing (in evergreen trees), sugars replenish the starch used in spring growth. Root growth may be stimulated at this time by sugar availability and warm soil temperatures. Throughout the winter, starch is used for maintenance respiration, but sparingly since low temperatures keep respiration rates low. *See* PHLOEM; PHOTOSYNTHESIS; XYLEM.

In adult trees, reproductive structures (flowers in angiosperms or strobili in gymnosperms) develop along with new leaves and represent large carbohydrate sinks. Sugars are preferentially utilized at

the expense of leaf, stem, and root growth. This reduces the leaf area produced, affecting the amount of sugars produced during that year, thereby reducing vegetative growth even further. The reproductive structures are present throughout the growing season until seed dispersal and continually utilize sugars that would normally go to stem and root growth.

Cold resistance. The perennial nature of trees requires them to withstand low temperatures during the winter. At higher latitudes and elevations, temperatures can reach −58°F (−50°C). Trees that evolved in these regions have physiological processes that protect their cells from damage to temperatures as low as −328°F (−200°C), whereas tropical species lacking these processes exhibit damage at above-freezing temperatures. Cell damage occurs from ice crystal formation within the cell, which disrupts the plasma membrane. Trees develop resistance to freezing through a process of physiological changes beginning in late summer. The degree of cold resistance within a tree depends on the tissue; buds and twigs typically tend to be more resistant than roots.

A tree goes through three sequential stages to become fully cold resistant. The process involves reduced cell hydration along wtih increased membrane permeability. The first stage is initiated by shortening days and results in shoot growth cessation, bud formation, and metabolic changes. Trees in this stage can survive temperatures down to 23°F (−5°C). The second stage requires freezing temperatures which alter cellular molecules. Starch breakdown is stimulated, causing sugar accumulation. Trees can survive temperatures as low as −13°F (−25°C) at this stage. The last stage occurs after exposure to very low temperatures (−22 to −58°F or −30 to −50°C), which increases soluble protein concentrations that bind cellular water, preventing ice crystallization. Trees can survive temperatures below −112°F (−80°C) in this stage. A few days of warmer temperatures, however, causes trees to revert to the second stage.

Water relations. Trees are likely to experience drought conditions during their lifetime. Unlike annual plants that survive drought as seeds, trees have evolved traits that allow them to avoid desiccation. These traits include using water stored in the stem, stomatal closure, and shedding of leaves to reduce transpirational area. All the leaves can be shed and the tree survives on stored starch. Another trait of some species is to produce a long tap root that reaches the water table, sometimes tens of meters from the soil surface.

On a daily basis, trees must supply water to the leaves for normal physiological function. If the water potential of the leaves drops too low, the stomata close, reducing photosynthesis. To maintain high water potential, trees use water stored in their stems during the morning which is recharged during the night. *See* PLANT-WATER RELATIONS.

Transport and support. Trees have evolved a means of combining long-distance transport between the roots and foliage with support through the production of secondary xylem (wood) by the vascular cambium. In older trees the stem represents 60–85% of the aboveground biomass (and about 40–50% of the total tree biomass). However, 90% of the wood consists of dead cells. These dead cells function in transport and support of the tree. As these cells develop and mature, they lay down thick secondary walls of cellulose and lignin that provide support, and then they die with the cell lumen becoming an empty tube. The interconnecting cells provide an efficient transport system, capable of moving 106 gal (400 liters) of water per day. The living cells in the wood (ray parenchyma) are the site of starch storage in woody stems and roots.

Further specialization within the secondary xylem results from variable sugar levels, and possibly hormonal levels, at the time of development. During spring growth when sugars are being used primarily for leaf growth, xylem cells form thinner secondary walls that have large lumens. These cells facilitate rapid water movement to the expanding foliage. Later in the year as more sugars are transported to the stem and roots, these cells form thick secondary walls and small lumens which function for support. *See* PLANT TRANSPORT OF SOLUTES.

Mineral nutrition. Mineral nutrients are required in proper concentrations and in relative proportions to one another by trees for normal growth and development. Nutrient deficiencies are similar in trees and other plants because of the functions of these nutrients in physiological processes. Tree nutrition is unique because trees require lower concentrations, and they are able to recycle nutrients within various tissues.

Trees adapt to areas which are low in nutrients by lowering physiological functions and slowing growth rates. In addition, trees allocate more carbohydrates to root production, allowing them to exploit large volumes of soil in search of limiting nutrients. Proliferation of fine roots at the organic matter–mineral soil interface where many nutrients are released from decomposing organic matter allows trees to recapture nutrients lost by leaf fall.

Internal cycling keeps nutrients from leaving the tree (then the nutrients could be lost to neighboring trees or competing vegetation) and accounts for a majority of the nutrients used annually for growth. Trees translocate nutrients out of leaves during senescence, prior to abscission. These nutrients can be translocated to newly expanding leaves or to storage sites located in either the twigs (deciduous trees) or younger foliage (evergreen trees). As an example, nitrogen is readily transported as amino acids during leaf senescence and, in deciduous trees, is stored as a special protein in twig bark. This process is triggered by shortening days in the fall. As buds enlarge in the spring, these proteins break down into their constituent amino acids and are translocated to newly developing leaves. *See* PLANT MINERAL NUTRITION; PLANT PHYSIOLOGY; TREE.

Jon D. Johnson

Bibliography. P. J. Kramer and T. T. Kozlowski, *Physiology of Woody Plants*, 1979; A. S. Raghavendra, *Physiology of Trees*, 1991.

Trematoda

A loose grouping of acoelomate, parasitic flatworms of the phylum Platyhelminthes formerly accorded class rank and containing the subclasses (or orders) Digenea, Monogenea, and Aspidobothria. Although neither the name Trematoda nor the classification scheme appears as valid as formerly in demonstrating helminth phylogenetic relationships (in light of recent information, covered in the Taxonomy section below), a discussion of the three older "trematode" groups just mentioned is profitable. These organisms commonly occur as adults in or on all vertebrate groups. They exhibit cephalization, bilateral symmetry, and well-developed anterior and ventral, or anterior and posterior, holdfast structures. Electron-microscope studies of some parasitic platyhelminths by D. L. Lee in 1966 and K. M. Lyons in 1970 indicate that the outer covering of these worms, once thought to be a secreted, nonliving cuticle, is probably a living epidermis. The mouth is anterior, and usually a blind, forked gut occurs, as well as three muscle layers. The excretory system consists of flame cells and collecting tubules. These animals are predominantly hermaphroditic and oviparous with operculated egg capsules. The life histories of the Digenea are complex, while those of the Monogenea and Aspidobothria are simple.

Taxonomy. Systematic studies are based mainly on comparative morphology of adults and larvae, although life history and ecological data are also used. In 1957 G. R. LaRue presented an interesting, detailed history of trematode systematics and a classification of digenetic trematodes based upon larval characteristics.

Also in 1957 B. E. Bychowsky, a recognized leader in parasite research, reiterating his ideas of 1937, commented extensively on trematode classification. J. Llewellyn in 1965 and H. W. Manter in 1969 also examined their systematics of the groups involved. H. W. Stunkard presented in 1962 a revised classification which was quite similar to that of L. H. Hyman in 1951:

Class Trematoda
  Order Monogenea
    Suborder: Monopisthocotylea
              Polyopisthocotylea
  Order: Aspidobothria
         Digenea

It is noteworthy that Stunkard separated the aspidobothrids and the digeneids more distinctly as follows:

Class Trematoda
  Subclass Pectobothridia
    Order: Monopisthocotylea
           Polyopisthocotylea
  Subclass Malacobothridia
    Order: Aspidobothrea
           Digenea

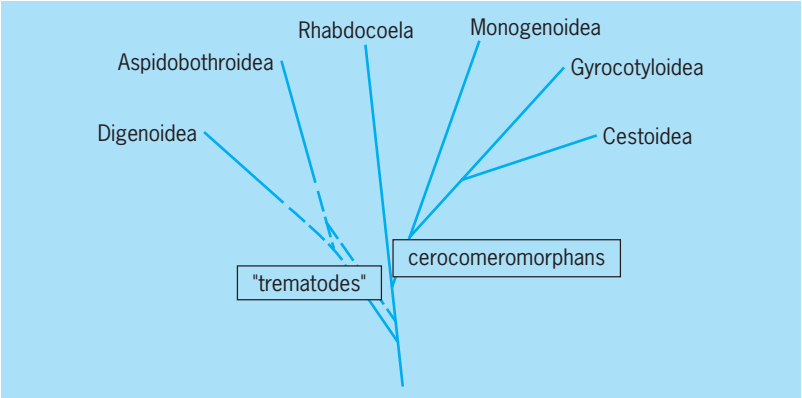


Fig. 1. Probable lines of evolution of Digenoidea, Aspidobothroidea, Monogenoidea, Gyrocotyloidea, and Cestoidea from rhabdocoel turbellarians as based on recent information.

In 1965 Llewellyn, basing some of his conclusions on Bychowsky's earlier works, recommended separation of the grouping Trematodes, which included the classes Digenoidea and Aspidobothroidea, from those platyhelminth parasites which generally have hooked larvae. Within the grouping Cercomeromorphae, he housed Monogenoidea, Gyrocotyloidea, Cestodaria, and Cestoda, which appear to have been accorded class ranking (**Fig. 1**).

Manter's classification encompasses many of the concepts of these later workers, differing somewhat in rank accorded Aspidogastrea (Aspidobothrea) and Digenea, which he called subclasses.

After extensive research in these laboratories the author is inclined to place some credence in the "cercomeromorph concept" and, keeping this in mind, to prefer the arrangement of Manter, rearranging the order so that the class Cestoidea are nearer the Gyrocotyloidea and Monogenoidea. Without more experience and evidence the author prefers not to address the class Mesozoa, a highly aberrant grouping of organisms. The position of Aspidobothroidea is less clear than some of the other groupings because of their general anatomical similarity to the Digenoidea and their monogeneid life histories. However, aspidobothreids seem sufficiently different from both groups to deserve equal status. The arrangement would appear as follows:

Class: Monogenoidea
       Gyrocotyloidea
       Cestoidea
  Subclass: Amphilinidea
            Cestoda
Class: Digenoidea
       Aspidobothroidea

The current phylogenetic concept is represented in Fig. 1.

The Digenoidea and Aspidobothroidea are probably phylogenetically older than the Monogenoidea, since they are entoparasitic, parasitize both invertebrates and vertebrates, and have more complicated life histories. Since early mollusks were marine

forms, these trematodes are probably marine in origin. Monogenoidea probably originated in the sea also because a greater divergence of forms has taken place among marine species. All trematodes are strikingly similar to rhabdocoel turbellarians, from which they probably evolved. Indeed, some turbellarians have simple sucking holdfasts and tend toward parasitic habits. The major differences between free-living turbellarians and trematodes are the addition of complex holdfasts, development of a cuticle, loss of photoreceptors, and an increase in reproductive capacities with concomitant alterations in the reproductive system, all of which are modifications for the parasitic habit. *See* TURBELLARIA.

Morphology. Although the three classes share a common internal morphological plan, they differ considerably externally. The Digenoidea are commonly more elongate, with continuous outlines, while the Monogenoidea have enlarged anterior and posterior holdfasts and more irregular outlines, and the Aspidobothroidea are modified ventrally. Digeneids usually have oral and ventral suckers as adult holdfasts. Aspidobothreids have large, ventral, adhesive disks with many depressions, or alveoli. Monogenoidea have paired suckers, or adhesive glands, anteriorly and armed suckers or wedges, called opisthaptors, posteriorly. Though all classes have longitudinal, transverse, and oblique muscle layers, monogeneids usually have more complex arrangements of longitudinal muscle bands to operate the holdfasts. Digestive systems consist of an anterior mouth, muscular pharynx, short esophagus, and forked, blind gut, with the exceptions of *Udonella* and the Aspidobothroidea, which have single, median digestive sacs. In some digeneids and many monogeneids the main intestinal branches are much ramified. The nervous system has an enlarged anterior esophageal ganglion or brain. Dorsal nerve trunks usually lacking in Monogenoidea, as well as ventral, longitudinal, and transverse trunks, also occur. Photosensitive eyespots occur in the larvae of many Digenoidea and Monogenoidea, and in some adult monogeneids. Digenoidea usually have a single nephridiopore, while monogeneids possess two. Aspidobothroidea feature both conditions. Most trematodes possess turbellarianlike hermaphroditic reproductive systems with single testis or multiple testes, a single ovary, a common gonopore, and a protrusible, or eversible, copulatory organ. A few digeneids, the blood flukes, have separate sexes. All three groups may have homologous accessory ducts running from the female tracts to the outside, known as Laurer's canal in the Digenoidea and Aspidobothroidea; or they may have accessory tracts to the gut, like the genitointestinal canal of the Monogenoidea. Cross-fertilization and self-fertilization are possible and, depending on the availability of mates, both occur, but the former probably predominates. Oviparity is the rule, but a few monogeneids, such as *Gyrodactylus* and *Isancistrinae*, bear living young. The ciliated embryo, which usually does not occur in the Aspidobothroidea, is enclosed in an operculated egg capsule containing nourishing vitelline cells.

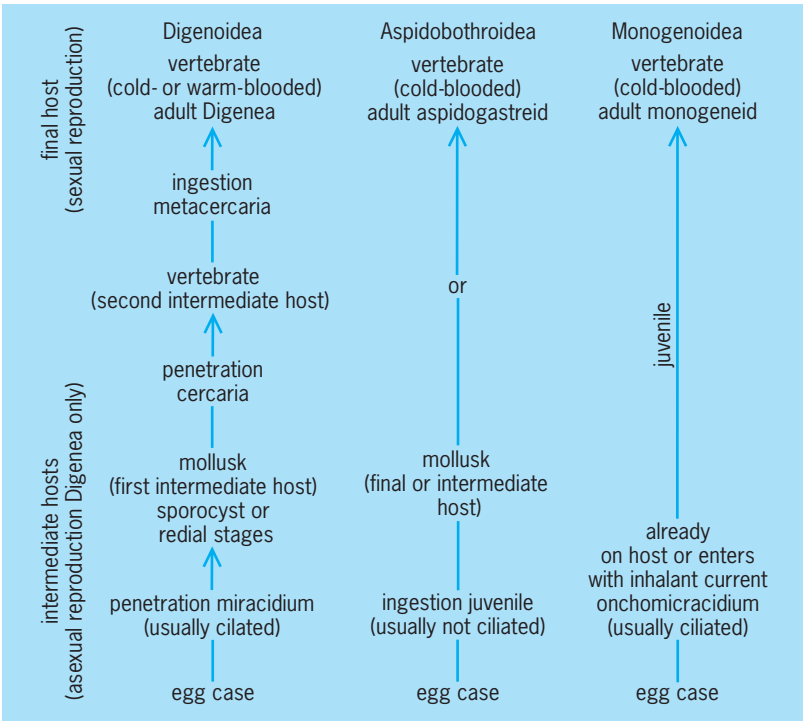


Fig. 2. Stylized diagrams of the life histories of three trematode classes.

Life history. Egg production of most digeneids is high and, coupled with asexual multiplication of the larval stages, results in enormous numbers of young. Most probably succumb to the hazards of their complicated life cycles; therefore large numbers are necessary. The Monogenoidea produce fewer embryos (**Fig. 2**).

The complicated life cycles of the Digenoidea involve asexual reproduction in the first intermediate host, which is generally a mollusk, and sexual reproduction in the final hosts. Aspidobothroidea usually develop directly, and they find lodgment in the final molluscan host. Some aspidobothreids, however, have an intermediate molluscan, or decapod, host with turtles or fishes as the final host. Monogeneids possess ciliated larvae, which undergo simple metamorphosis into the juvenile form on a single host.

Physiology. Comparatively little is known of helminth physiology. Such studies are complicated by the parasitic habit, small size, and complex life histories of the worms. Helminth metabolism is not as simple as once believed, and involves great adaptability to varying oxygen tensions and pH ranges, among other factors. Many digeneids are capable of respiration under low-oxygen or anaerobic conditions. The physiological activities of larvae and juveniles are probably very different from those of the adults. Trematodes feed upon host tissues, body fluids, and exudates, and both extracellular and intracellular digestion takes place. Transcutaneous absorption of nutrients may occur. Efforts at culturing trematodes on artificial media have met with some success, and several investigators are pursuing transplantation studies.

Ecology. Trematodes parasitize a wide variety of invertebrates and vertebrates and occupy almost every available niche within these hosts. The adaptations demanded of the worms for survival are as varied as the characteristics of the microhabitats. Over the millions of years of coevolution of the hosts and their parasites, delicate balances have, for the most part, been attained, and under normal conditions it is probable that trematodes rarely demand more than the host can supply without undue strain. This has involved adjustments between the antigenic properties of the parasite and the defense reactions of the host. In many entoparasitic trematodes, delicate balances have developed between the protein digestive enzymes of the host and the resistant properties of the cuticle and other tissues of the parasite. In *Paragonimus* the cuticle actually appears to be digested by the host but is continually renewed by the parasite. Monogeneids do not have the same problems but have had to develop efficient adhesive mechanisms to keep from being swept from the surfaces of their hosts. This struggle has produced many interesting holdfast organs.

It seems axiomatic that the host must survive until the parasite can again gain access to another host or until the life cycle is completed. Those parasites which cause the least disruption of the host's activities are probably the oldest as well as the most successful. Immunities are sometimes developed by the hosts. Many trematodes seem to possess such rigid requirements and such responses to particular hosts that host specificity is a phenomenon of considerable significance. Monogeneids appear more host-specific than digeneids, and aspidobothreids seem less specific than both. Trematodes are of considerable veterinary and medical importance because under certain conditions they cause debility, even death. *See* ASPIDOGASTREA; DIGENEA; MONOGENEA; PLATYHELMINTHES.

William J. Hargis, Jr.

Bibliography. J. G. Baer, *Ecology of Animal Parasites*, 1951; B. E. Bychowsky, *Monogenetic Trematodes, Their Classification and Phylogeny*, 1957; B. E. Bychowsky, Ontogeny and phylogeny of parasitic platyhelminths, *Izv. Akad. Nauk. SSSR, Ser. Biol.*, no. 4, 1353–1383, 1937; B. Dawes, *The Trematoda*, 1946; L. H. Hyman, *The Invertebrates*, vol. 2, 1951; A. V. Ivanov, *Udonella caligorum* Johnston, 1835, a representative of a new class of flatworms, *Zool. Zh.*, 31(2):175–178, 1952; G. R. LaRue, The classification of digenetic Trematoda: A review and a new system, *Exp. Parasitol.*, 6(3):306–349, 1957; D. L. Lee, The structure and composition of the helminth cuticle, *Adv. Parasitol.*, 4:187–254, 1966; J. Llewellyn, The evolution of parasitic platyhelminths, *Evolution of Parasites: 3d Symposium of the British Society for Parasitology*, pp. 47–78, 1965; K. M. Lyons, The fine structure and function of the adult epidermis of two skin parasitic monogeneans, *Entobdella soleae* and *Acanthocotyle elegans, Parasitology*, 60:39–52, 1970; H. W. Manter, Problems in systematics of trematode parasites, *Problems in Systematics of Parasites*, pp. 91–105, 1969; H. W. Stunkard, *Taeniocotyle* nom. nov. for *Macraspis* Olsson, 1869, preoccupied, and systematic position of the Aspidobothrea, *Biol. Bull.*, 122(1):137–148, 1962; S. Yamaguti, *The Digenetic Trematodes of Vertebrates*, in *Systema Helminthum*, vol. 1, pts. 1 and 2, 1958; S. Yamaguti, *Monogenea and Aspidocotylea*, in *Systema Helminthum*, vol. 4, 1963.

Tremolite

The name given to magnesium-rich monoclinic calcium amphibole $Ca_2Mg_5Si_8O_{22}(OH)_2$. The mineral is white to gray, but colorless in section, and optically negative. It usually exhibits long prismatic crystals with prominent (110) amphibole cleavage. Unlike other end-member compositions of the calcium amphibole group, very pure tremolite is found in nature. Substitution of Fe for Mg is common, but pure ferrotremolite, $Ca_2Fe_5Si_8O_{22}(OH)_2$, is rare. Intermediate compositions between tremolite and ferrotremolite are referred to as actinolites, and are green in color and encompass a large number of naturally occurring calcium amphiboles. The substitution of Na, Al, and Fe^{3+} ions into the amphibole structure is common in actinolites. The nature of these substitutions is complex and leads, under some conditions, to miscibility gaps between actinolites and aluminous calcium amphiboles called hornblendes. For the most part, however, the physical and chemical variations between actinolite and hornblende are continuous so that an arbitrary actinolite-hornblende division is necessary. Most of the amphibole classifications divide actinolite from the hornblende series on the basis of Al substitution for Si with actinolite containing less than 0.5 atom of Si replaced by Al per formula unit. *See* HORNBLENDE.

The basic building block for the tremolite-actinolite crystal structure is the silicon tetrahedron. In all amphiboles, double chains of SiO_4 tetrahedrons are formed through joining two or three tetrahedrons at their corners by consecutively sharing two or three oxygens in alternating fashion along the entire chain length. These double chains form two anionic layers: a nearly coplanar layer of oxygens at the base of the tetrahedrons, and a second layer of oxygens and associated OH and F along the apices of the tetrahedrons. Silicon, and to a lesser extent aluminum, atoms lie between these basal and apical anionic layers in fourfold coordination. The double chains are arranged so that along the *b* crystallographic direction the basal oxygen layer of one chain is approximately coplanar with the apical anionic layer of the adjacent chains. Along the *a* crystallographic direction, the apical oxygen layers of adjacent chains face each other, as do the basal oxygen layers. The *c* axis parallels the long axis of the double chains. The double-chain structure is held together by bonding apical oxygens of facing double chains to a cation strip of 5Mg and 2Ca which have six- to eightfold coordination. Tremolite is usually described in terms of the (001) face-centered cell (space group C2/m). Typical lattice parameters for tremolite are

$a = 9.83$; $b = 18.05$; and $c = 5.27$ (all in angstroms; 10 Å = 1 nanometer); and β angle = 104.5°. Actinolite, due to its higher Fe^{2+} content, has a larger cation strip and hence a proportionately larger b dimension.

Tremolite in pure form is a product of thermal and regional metamorphism of siliceous dolomites and marbles as shown by reaction (1). In similar rocks

$$\underset{\text{Dolomite}}{5CaMg(CO_3)_2} + \underset{\text{Quartz}}{8SiO_2} + H_2O \longrightarrow \underset{\text{Tremolite}}{Ca_2Mg_5Si_8O_{22}(OH)_2} + \underset{\text{Calcite}}{3CaCO_3} + 7CO_2 \quad (1)$$

at higher grades of metamorphism, in the presence of both calcite and quartz, tremolite breaks down to form diopside, as shown by reaction (2). Actinolites, owing to a more variable chemistry, are more

$$\underset{\text{Tremolite}}{Ca_2Mg_5Si_8O_{22}(OH)_2} + \underset{\text{Calcite}}{3CaCO_3} + \underset{\text{Quartz}}{2SiO_2} \longrightarrow \underset{\text{Diopside}}{5CaMgSi_2O_6} + 3CO_2 + H_2O \quad (2)$$

ubiquitous than tremolite in occurrence. They are most commonly found in regionally metamorphosed mafic igneous rocks such as basalts and are known to occur in a wide range of pressure conditions. Both tremolite and actinolite can form through the breakdown of olivine and pyroxenes in regionally metamorphosed ultrabasic rocks; associated minerals are talc, chlorite, and carbonates. Actinolitic amphiboles formed as a breakdown product of pyroxenes are referred to as uralite. Intergrown fibrous crystals of tremolite are known as nephrite, a form of jade widely used for centuries, in making of artifacts and jewelry. Highly fibrous tremolite is used in commercial asbestos. *See* AMPHIBOLE; ASBESTOS; DIOPSIDE; JADE; METAMORPHISM.

Barry L. Doolan

Trepostomata

An extinct order of bryozoans in the class Stenolaemata. Trepostomes possess generally robust colonies, composed of tightly packed, moderately complex, long, slender, tubular or prismatic zooecia, with solid calcareous zooecial walls. Colonies show a moderately gradual transition from endozone to exozone regions, and they are exclusively free-walled. *See* BRYOZOA; STENOLAEMATA.

Morphology. Trepostome colonies range from small and delicate to large and massive; they can be thin to thick encrusting sheets; tabular, nodular, hemispherical, or globular masses; or bushlike or frondlike erect growths. Most colonies are divisible into very distinct endozone and exozone regions, with that portion of each zooecium lying within the endozone relatively long compared with that within the exozone in most erect colonies. Trepostome colonies usually bear mesopores (or exilapores) but lack other polymorphs. There is no unallocated colonial skeleton (extrazooidal skeleton) secreted between zooecia where colonies were actively growing. However, secondary extrazooidal skeleton may cover over colony surfaces in older areas. Colony surfaces may be smooth, but more commonly there are regularly spaced maculae (spots), generated by clusters of polymorphs, enlarged zooecia, or even groups of normal zooecia. In many trepostomes, maculae generate small, conspicuous bumps (monticules).

Individual trepostome zooecia are long, straight to markedly curved, cylindrical to prismatic tubes. The walls of adjacent zooecia are fused, forming one continuous skeletal element in which a boundary between zooecia may or may not be visible. These walls are generally extensively calcified, nonporous, pronouncedly laminated in microstructure, thin in the endozone, and thick in the exozone. Commonly, colony surfaces have small to large spines projecting from the skeletal surface; these are seen as conspicuous rod-shaped structures (styles) in thin slices cut through the skeleton. Most zooecia are crossed inside by diaphragms; however, in some trepostomes they are crossed by hemiphragms, heterophragms, or cystiphragms. Apertures are round to polygonal in outline.

History and classification. Apparently exclusively marine, the trepostomes first appeared about the start of the Middle Ordovician; they apparently share a common ancestor with cystoporates based on similar ranges of colony forms, zooidal arrangements and shapes, and simple colony organization. The trepostomes immediately rose to dominance within the bryozoan fauna and are among the most common macrofossils of the early Paleozoic; in places, trepostomes contributed to construction of small reefs, as well as being abundant in level-bottom environments. They remained abundant through Silurian time, declined during the Devonian, and died out in the Late Triassic. Their assignment to family, genus, and species usually requires preparation of orientated thin sections or peels that clearly show details of colony interiors. *See* CYSTOPORATA.

Roger J. Cuffey; Frank K. McKinney

Bibliography. R. S. Boardman, Indications of polypides in feeding zooids and polymorphs in lower Paleozoic Trepostomata (Bryozoa), *J. Paleont.*, 73:803–815, 1999; R. S. Boardman, A. H. Cheetham, and A. J. Rowell (eds.), *Fossil Invertebrates*, Blackwell Scientific Publications, Palo Alto, CA, 1987; P. D. Taylor and G. P. Larwood, Major evolutionary radiations in the Bryozoa, in P. E. Taylor and G. P. Larwood (eds.), *Major Evolutionary Radiations*, pp. 209–233, Clarendon Press, Oxford, 1990.

Trestle

A succession of towers of steel, timber, or reinforced concrete supporting the horizontal beams of a roadway, bridge, or other structure. Little distinction can be made between a trestle and a viaduct, and the terms are used interchangeably by many engineers. A viaduct is defined as a long bridge consisting of a series of short concrete or masonry spans

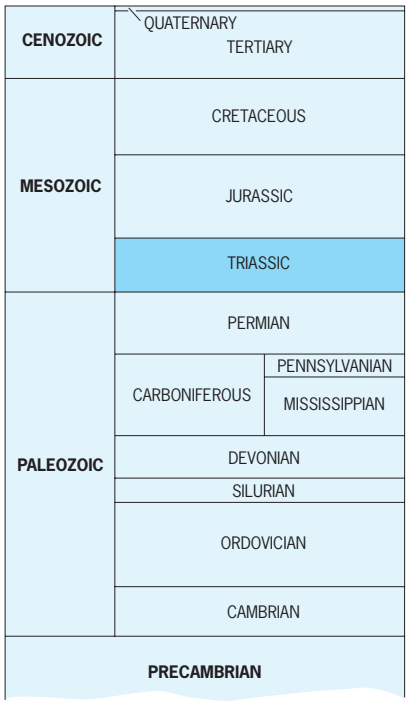
supported on piers or towers, and is used to carry a road or railroad over a valley, gorge, another roadway, or across an arm of the sea. A viaduct may also be constructed of steel girders and towers. It is even more difficult to draw a distinction between a viaduct and a bridge than it is between a viaduct and a trestle. *See* BRIDGE.

A trestle or a viaduct usually consists of alternate tower spans and spans between towers. For low trestles the spans may be supported on bents, each composed of two columns adequately braced in a transverse direction. A pair of bents braced longitudinally forms a tower. The columns of one bent of the tower are supported on planed base plates or movable shoes to allow horizontal movement in the longitudinal direction of the trestle. Struts connect the column bases and force the movable shoes to slide. The width of the base of a bent is usually not less than one-third the height of the bent. This width is sufficient to prevent excessive uplift at windward columns when the trestle is unloaded. *See* STRUCTURE (ENGINEERING); TOWER.

Charles M. Antoni

Triassic

The oldest period of the Mesozoic Era, encompassing an interval between about 248 and 206 million years ago (Ma). It was named in 1848 by F. A. von Alberti for the threefold division of rocks at its type locality in central Germany, where continental redbeds and evaporites of the older Buntsandstein and younger Keuper formations are separated by marine limestones and marls of the Muschelkalk formation. These carbonates were laid down in a shallow tongue of the Tethys seaway that extended from the Himalayas through the Middle East to the Pyrenees,



where more than 10,000 ft (3000 m) of carbonate were deposited. The German section was an unfortunate choice because it is atypical of other Triassic sections and carries a sparsely preserved fossil record. It was subsequently replaced by a marine carbonate sequence in the Alps as the standard for global Triassic reference and correlation. The North American standard marine section is in the western Cordilleras of British Columbia and the Sverdrup Basin of the Arctic. *See* MESOZOIC.

Major events. Triassic strata record profound paleontologic changes that reflect major physical changes in Earth history. Two of the five most catastrophic extinctions of the Phanerozoic Eon mark the beginning and end of the Triassic. More than 50% of all Permian families died out at the beginning of the period, including 85–90% of all marine species and 75% of land species; and more than 50% of all marine genera became extinct at the end of the period. The Triassic was also when many new families of plants and animals evolved, including the earliest known mammals.

As a very brief interval of geologic time (about 40 million years), the Triassic Period uniquely embraces both the final consolidation of Pangaea and the initial breakup of the landmass, which in the Middle Jurassic led to the opening of the Central Atlantic Ocean and formation of modern-day continental margins. The Triassic marks the beginning of a new Wilson cycle of ocean-basin opening through lithospheric extension and oceanic closing through subducting oceanic lithospheres along continental margins. The cycle was named for J. Tuzo Wilson, a pioneer of modern plate tectonic theory. The initial breakup of Pangaea occurred in the western Tethys (precursor of the Mediterranean Sea) between Baltica and Africa and in eastern Greenland between Baltica and the North American craton. Rifting then proceeded into the Central Atlantic, separating the North American and African cratons that led to the separation of Laurasia from Gonwanaland (**Fig. 1**). Rifting also occurred in Argentina, east Africa, and Australia. In the central Atlantic region, extensional tectonics was accompanied by a huge outpouring of continental flood basalts, forming the Central Atlantic magmatic province (CAMP), whose remnants are now found as feeder dikes and flood basalts on four circum-Atlantic continents, separated by thousands of miles of younger basalts of the oceanic crust (**Fig. 2***a*). *See* LITHOSPHERE; PLATE TECTONICS.

Final consolidation of Pangaea. The initial consolidation of Pangaea, which was marked by the formation of the Allegheny-Mauritanide-Variscan mountain chain in the middle Carboniferous (320 Ma), resulted from the collision of Gondwanaland and the combined Laurasia-Baltica-Siberian-Kazakhstania landmass (Fig. 1). Major plate accretion continued into the Middle-to-Late Triassic (230 ± 5 Ma), when southern China and Cimmeria (Asia Minor) were sutured to the northern margin of the Tethys seaway. Smaller terranes, called suspect or exotic, were also accreted to the western margin of North

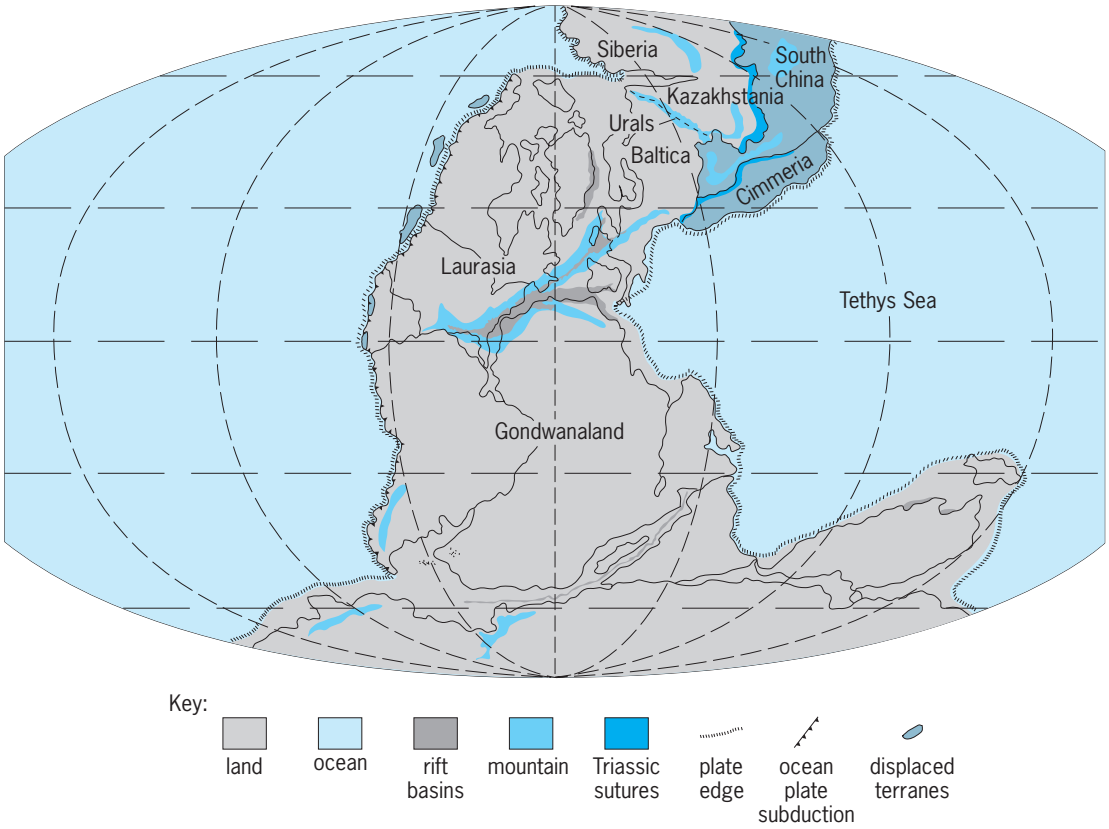


Fig. 1. Paleogeography of the Late Triassic Period: after the accretion of south China and Cimmeria (Turkey, Iran, and Tibet) to Laurasia; during the Incipient rifting of Pangaea in eastern North America and northwest Africa along the Allegheny-Mauritanide-Variscan orogeny; and concurrent with oceanic subduction and formation of deep-sea trenches and magmatic arc along the western plate boundary of North America. (*After R. K. Bambach, C. R. Scotese, and A. M. Ziegler, Before Pangea: The geographics of the Paleozoic world, Amer. Sci., 68:26–38, 1980*)

America at this time. Disconnected patches of Triassic strata occur from California through western British Columbia into Alaska, where they appear to be displaced island-arc terranes, microcontinents, and ocean-ridge segments, as inferred from paleomagnetic data in the lavas and by the exotic character of their Permian faunas. *See* PALEOGEOGRAPHY; PALEOMAGNETISM.

Pangaean supercontinent. The final phase of deformation produced a broadly convex continental plate that extended from the north to the south paleo poles, covered about 25% of the Earth's surface, and was surrounded by a global ocean called Panthalassa. It had a central arch standing about 1 mi (1.6 km) high and an average elevation of more than 4300 ft (1300 m) above the early Mesozoic sea level (Fig. 1). Because of its size, location, and pronounced orographic peaks that probably rivaled the Himalayas, the Pangaea landmass had a major impact on global climates. During the Middle Triassic, Florida lay about 5° south of the Equator, whereas Grand Banks (now off southeastern Newfoundland) was located about 20°N. Pangaea's climatic zones ranged from tropical savanna along its extensive coasts to arid and semiarid across its vast interior. *See* DESERT; SAVANNA; SUPERCONTINENT.

As the plate migrated north, transgressing about 10° of latitude between the Middle Triassic and Middle Jurassic, the plate was subjected to increased aridity as it moved under the influence of the subtropical high-pressure cell. Because of its large size, the landmass must have been subjected to monsoon circulation. Winters along the future central Atlantic probably were dominated by subtropical high-pressure cells bringing in cool dry air from aloft, whereas summers were dominated by equatorial low-pressure systems bringing in warm moist air from the Tethys seaway to the east. As moist air was uplifted almost 1.2 mi (2 km) over the Alleghenian-Variscan chain, it would have cooled adiabatically, yielding rainwater that fed major rivers (for example, the Congo River) flowing thousands of miles away from the axis of uplift across broad alluvial plains to the coastal regions of Alaska, Patagonia, India, and Siberia. *See* MONSOON METEOROLOGY.

With the onset of rifting in the Late Triassic and subsequent topographic changes, small ephemeral streams flowed into the rift valleys, creating huge lakes that may have been comparable in size to present Lake Tanganyika of the East African rift system. Where air masses descended into low-lying rift basins, along the Central Atlantic axis, they warmed adiabatically, causing evaporation and precipitation of evaporite minerals (for example, halite, gypsum, and anhydrite) in marginal epicontinental seas and in continental lacustrine basins. The Triassic and Lower

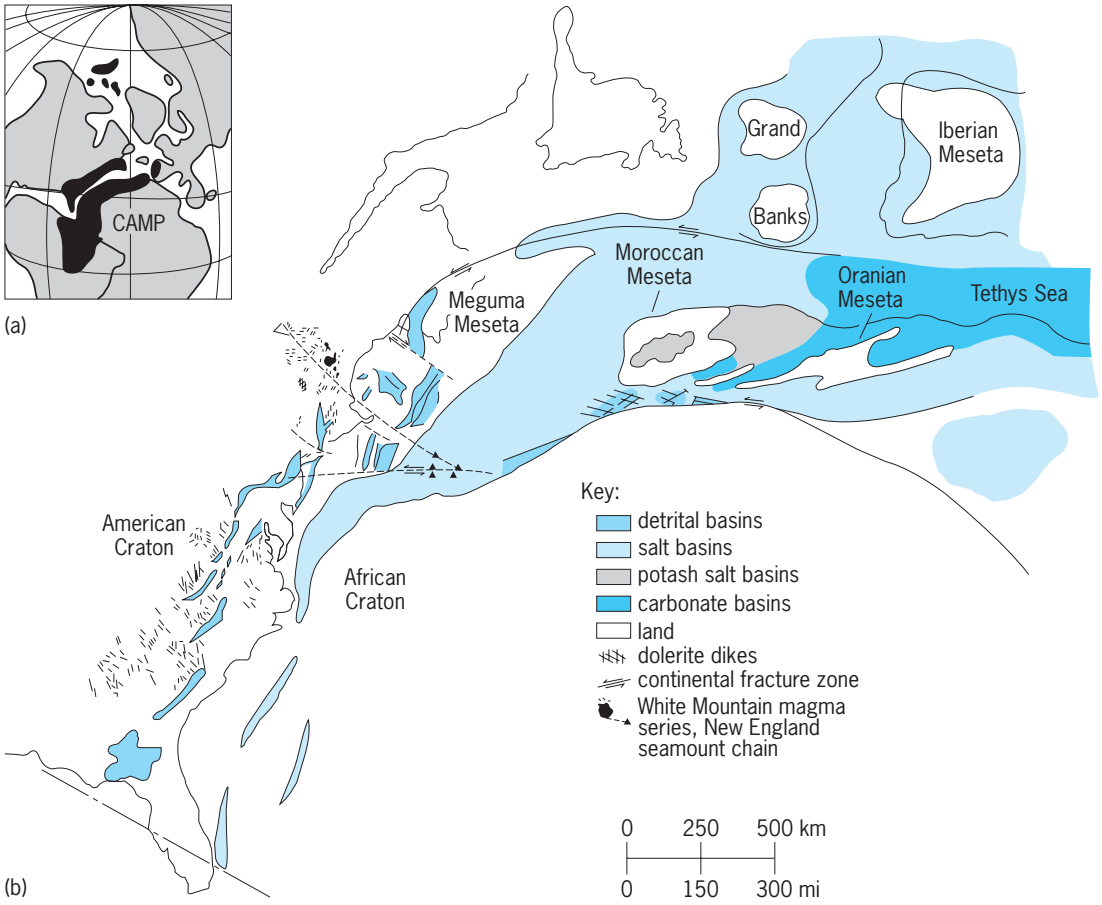


Fig. 2. (*a*) Map of CAMP. (*b*) Late Triassic–Early Jurassic reconstruction of eastern North America and northwest Africa, outlining rift basins, lithofacies, and tectonic elements, during breakup but prior to sea-floor spreading of the Central Atlantic in the Middle Jurassic. Arrows indicate the direction of intraplate movement along continental fracture zones. (*After W. Manspeizer, ed., Triassic-Jurassic Rifting: Continental Breakup and the Origin of the Atlantic Ocean and Passive Margins, Elsevier, 1988*)

Jurassic lake deposits show a pervasive cyclical pattern of wetting and drying, wherein lakes expanded and contracted with periodicities of 21,000, 42,000, 100,000, and 400,000 years. These intervals agree with the Holocene Milankovitch astronomical theory of climates that are related to small variations in the Earth's orbit and rotation. *See* BASIN; JURASSIC; RIFT VALLEY; SALINE EVAPORITES.

Crustal extension. The most important tectonic event in the Mesozoic Era was the rifting of the Pangaea craton, which began in the Late Triassic, culminating in the Middle Jurassic with the formation of the Central Atlantic ocean basin and the proto-Atlantic continental margins (Fig. 2*b*). Rifting began in the Tethys region in the Early Triassic, and progressed from western Europe and the Mediterranean into the Central Atlantic off Morocco and eastern North America by the Late Triassic. As crustal extension continued throughout the Triassic, the Tethys seaway spread farther westward and inland. Although marine palynomorphs from deep wells on Georges Bank indicate that epicontinental seas, from Tethys on the east or Arctic Canada (through eastern Greenland) on the north, transgressed the craton to the coast of Massachusetts in the Late Triassic, an ocean sea floor did not form in this region until the Middle Jurassic. By that time, rifting and sea-floor spreading extended into the Gulf of Mexico, separating North and South America. Africa and South America did not separate until the Early Cretaceous, when sea- floor spreading created the South Atlantic ocean basin, the great flood basalts of the Amazon and Karoo (Africa), and those of Transarctic and Tasmania record that Gondwanaland had begun to break up by the Triassic Period. *See* BASALT; CRETACEOUS; PALYNOLOGY.

Atlantic rift basins. Continental rift basins, passive continental margins, and ocean basins form in response to divergent stresses that extend the crust. Crustal extension, as it pertains to the Atlantic, embraces a major tectonic cycle marked by Late Triassic–Early Jurassic rifting and Middle Jurassic to Recent (Holocene) drifting. The rift stage, involving heating and stretching of the crust, was accompanied by uplift, faulting, basaltic igneous activity, and rapid filling of deep elongate rift basins. The drift stage, involving the slow cooling of the lithosphere over a broad region, was accompanied by thermal subsidence with concomitant marine transgression of the newly formed plate margin. The transition from rifting to drifting, accompanied by sea-floor spreading, is recorded by the postrift

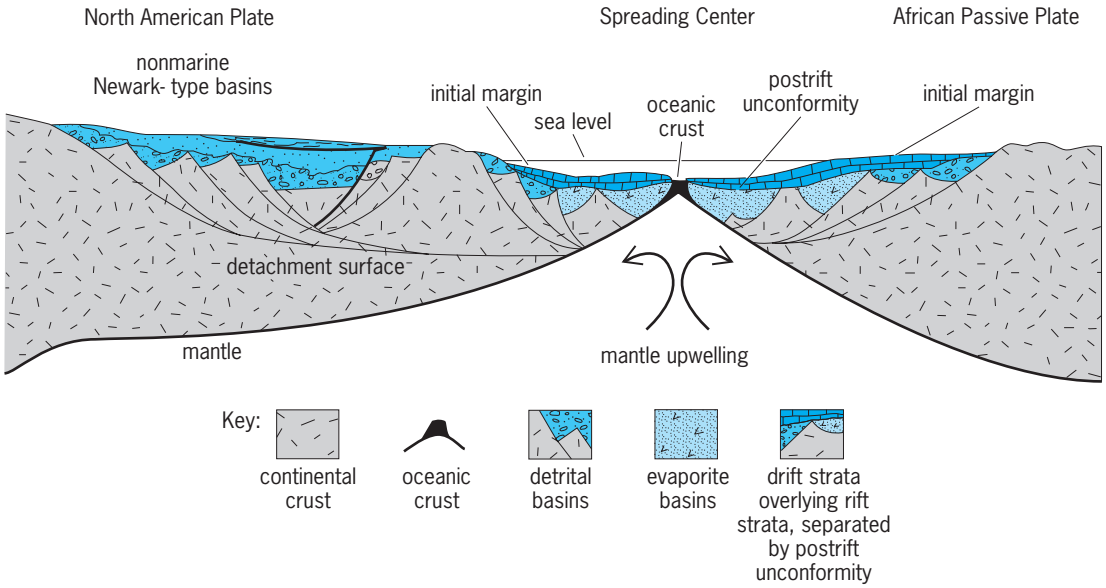


Fig. 3. Diagrammatic cross section of the Atlantic-type continental passive margine of North America and North Africa, taken at the beginning of the Middle Jurassic with the onset of sea-floor spreading that resulted from crustal thinning and mantie upwelling Note the setting of the Late Triassio-Early Jurassic continental and marine rift basins and their relation to the future passive margins, the postrift unconformity, and the overtying Middle Jurassic drift sequence. (*After W. Manapelzer, ed., Triassic-Jurassic Rifting: Continental Breakup and the Origin of the Atlantic Ocean and Passive Margins, pt. A, Elsevier, 1988*)

unconformity (**Fig. 3**). Late Triassic Proto-Atlantic rift basins occur in eastern North America, Greenland, the British Isles, north and central West Africa, and South America. *See* CONTINENTAL DRIFT; CONTINENTAL MARGIN; HOLOCENE; UNCONFORMITY.

Within the proto-Atlantic, off eastern North America and Morocco, lie about 50 northeast- to southwest-trending elongate rift basins, called the Newark rift basins, whose trend follows the fabric of the Alleghenian-Variscan orogen (Fig. 2). Some of these basins are exposed on the land, while others occur beneath the Coastal Plain and under the continental shelf. Almost all of them have developed along reactivated late Paleozoic thrust faults (Fig. 3). Seismic reflection surveys of both the onshore and offshore rift basins show that they are asymmetric half-grabens, bounded on one side by a system of major high-angle normal faults, and on the other side by a gently sloping basement with sedimentary overlap. These basins contain Late Triassic to Early Jurassic strata, which comprise the Newark Supergroup. At the end of the Triassic and into the Early Jurassic, the Newark strata of the Atlantic region were uplift, tilted, faulted, and intruded by tholeiitic sills and dikes. Subsequently, they were eroded and unconformably overlain by younger Jurassic post rift or drift strata. This episode of deformation, known as the Palisade disturbance, is most evident by the postrift unconformity in the offshore basins. *See* FAULT AND FAULT STRUCTURES; GRABEN.

Figure 2, showing a predrift paleogeographic reconstruction of the circum-Atlantic region, outlines the major Triassic basins and lithofacies. Two major basin types are recognized (Fig. 3): Newark-type detrital basins, which are exposed onshore as half-grabens and contain a thick (approximately 2.5–5 mi or 4–8 km) sequence of fluvial-lacustrine strata and border fanglomerates; and evaporite basins, which occur seaward of the string of detrital basins and contain a thick evaporite facies with interbeds of red mudstones and carbonates. As more than 3300 ft (1000 m) of salt was concentrated, these basins must have acted as huge evaporating pans. The Triassic-Jurassic systemic boundary, throughout the broad region of the Atlantic, typically is marked by tholeiitic lava flows and intrusives that are dated about 200 Ma (Early Jurassic), or only slightly older than the oldest dated crust of the Atlantic Ocean. *See* FACIES (GEOLOGY).

Central Atlantic magmatic province. The breakup of Pangaea was accompanied by the most extensive outpouring of continental basaltic lava known, covering an area estimated to be about 4 million mi^2 (10 million km^2). Basaltic remnants (flood basalts and feeder dikes) of this igneous province, named the Central Atlantic Magmatic Province (CAMP), are found on the rifted margins of four circum-Atlantic continents, particularly eastern North America, South America, western Africa, and southwestern Europe (Fig. 2*a*). Almost all of CAMP rocks are mafic tholeoiites that were intruded into or extruded onto clastic rocks in Newark-type rift basins. The Palisades Sill, along the west shore of the lower Hudson River in Northern New Jersey, is an example of this magmatic event. It is thought that the immensely thick and widespread seaward-dipping basaltic wedges, manifested by the East Coast magnetic anomaly, are linked to CAMP. *See* GEOMAGNETISM.

Recent multidisciplinary studies in stratigraphy, palynology, geochronology, paleomagnetism, and petrography indicate CAMP formed as a singular episodic event in Earth's history, occurring during

a very brief interval of geologic time, perhaps no longer than 4 million years. Importantly, this event occurred about 200 Ma and was contemporaneous with widespread mass extinction at the Triassic-Jurassic boundary. A causal relationship is postulated by many scientists to climate change that was forced by emission of huge quantities of volcanic gasses, estimated by researchers to be in the order of from $1–5 \times 10^{12}$ metric tons. Radical shifts in climate due to the ejection of aerosols into the atmosphere and destruction of environments by lava flows, ash falls, fires, and toxic pollution of soil and streams are suggested consequences of this event. A similar explanation has been offered for the mass extinctions that occurred at the end of the Paleozoic and Mesozoic Eras, with major volcanic eruptions of the Siberian Traps and Deccan Traps, respectively.

Western North America. Permian-to-Triassic consolidation of Pangaea in western North America led to the Sonoma orogeny (mountain building), which resulted from overthrusting and suturing of successive island-arc and microcontinent terranes to the western edge of the North American Plate. However, toward the end of the Triassic Period, as crustal extension was occurring in the Central Atlantic region, the plate moved westward, overriding the Pacific Plate along a reversed subduction zone. This created, for the remainder of the Mesozoic Era, an Andean-type plate edge with a subducting sea floor and associated deep-sea trench and magmatic arc. These effects can be studied in the Cordilleran mountain belt, from Alaska to California, where great thicknesses of volcanics and graywackes were derived from island arcs to the west, and in Idaho and eastern Nevada, where thick Lower Triassic marine limestones and sandstones were laid down adjacent to the rising Cordilleras on the west and interfinger with continental redbeds derived from the stable interior to the east. *See* CORDILLERAN BELT; LIMESTONE; OROGENY; REDBEDS; SANDSTONE.

As the epicontinental seas regressed westward, nonmarine fluvial, lacustrine, and windblown sands were deposited on the craton. Today many of these red, purple, ash-gray, and chocolate-colored beds are some of the most spectacular and colorful scenery in the American West. For example, the Painted Desert of Arizona, known for its petrified logs of conifer trees, was developed in the Chinle Formation, and the windblown sands of the Wingate and Navajo formations are exposed in the walls of Zion National Park in southern Utah. *See* PETRIFIED FORESTS.

Life. The Triassic is bracketed by two major biotic crises that terminated many groups of organisms. Triassic marine faunas can be distinguished from their predecessors by the absence of groups that flourished in the Permian, such as the fusulinid foraminiferans, the tabulate and rugose corals, the trepostome and cryptostome bryozoans, the productid and other brachiopod groups, the trilobites, and certain groups of echinoderms. Owing to a very low stand of sea level, Early Triassic marine faunas are not common, and show very small diversity except for ammonites. This is partly ecologic. The reef community, for example, is not known from the Early Triassic deposits; yet when it reappeared in mid-Triassic time, it contained sponges that were major members of Permian reefs, and that must have survived in settings that have not been found. *See* BRACHIOPODA; BRYOZOA; ECHINODERMATA; FORAMINIFERIDA; FUSULINACEA; PERMIAN; RUGOSA; TABULATA; TRILOBITA.

Triassic faunas are also distinguished from earlier ones by newly evolved groups of plants and animals. In marine communities, molluscan stocks proliferated vigorously. Bivalves diversified greatly and took over most of the niches previously occupied by brachiopods; ammonites proliferated rapidly from a few Permian survivors. The scleractinian (modern) corals appeared, as did the shell-crushing placodont reptiles and the ichthyosaurs. In continental faunas, various groups of reptiles appeared, including crocodiles and crocodilelike forms, the mammallike reptiles, and the first true mammals, as well as dinosaurs. *See* CEPHALOPODA; CROCODYLIA; DINOSAUR; MAMMALIA; MOLLUSCA; PLACODONTIA; SCLERACTINIA.

The Jurassic faunas lack numerous stocks lost in the Rhaeto-Liassic faunal crises. These include survivors of the Permian crises, such as the orthoceratid cephalopods and the conodonts. However, stocks that had flourished greatly in the Triassic also became extinct (phytosaurs, placodonts) or nearly extinct: ammonites were reduced to one or two surviving lineages, which then underwent no other great evolutionary surge in Jurassic time. Furthermore, new groups such as the plesiosaurs and pterosaurs appeared. Triassic land plants contain survivors of many Paleozoic stocks, but the gymnosperms became dominant and cycads appeared. The Permo-Triassic and Rhaeto-Liassic crises record a severe stressing of the biosphere, but the nature and origin of these stresses have not been established. *See* CONODONT; CYCADEOIDALES; EXTINCTION (BIOLOGY); INDEX FOSSIL; PALEOBOTANY; PALEOECOLOGY; PALEONTOLOGY; PINOPHYTA; PTEROSAURIA.

Warren Manspeizer

Bibliography. A. Hallam, The end-Triassic bivalve extinction event, *Paleogeo. Paleoclimatol. Paleoecol.*, vol. 35, pp. 1–44, 1981; W. E. Hames et al. (eds.), *The Central Atlantic Magmatic Province: Insights from Fragments of Pangea*, American Geophysical Union, Washington, D.C., 2003; G. D. Klein (ed.), *Pangea: Paleoclimate, Tectonics, and Sedimentation During Accretion, Zenith, and Breakup of a Supercontinent* Spec. Pap. No. 288, Geological Society of America, Boulder, 1994; P. M. Letournea and P. E. Olsen (eds.), *The Great Rift Valleys of Pangea in Eastern North America*, vol. 1: *Tectonics, Structure and Volcanism*, Columbia University Press, New York, 2003; W. Manspeizer (ed.), *Triassic-Jurassic Rifting: Continental Breakup and the Origin of the Atlantic Ocean and Passive Margins*, pt. A, Elsevier Science, Amsterdam, 1988; D. R. Prothero and R. H. Dott, *Evolution of the Earth*, 7th ed., McGraw-Hill, New York, 2004; S. M. Stanley, *Earth System History*, Freeman, New York, 1999.

Tribology

The science and technology of interactive surfaces in relative motion. It incorporates various scientific and technological disciplines such as surface chemistry, fluid mechanics, materials, lubricants, contact mechanics, bearings, and lubrication systems. It is customarily divided into three branches: friction, lubrication, and wear.

Friction. This phenomenon is encountered whenever there is relative motion between contacting surfaces, and it always opposes the motion. As no mechanically prepared surfaces are perfectly smooth, when the surfaces are first brought into contact under light load, they touch only along the asperities (real area of contact). The early theories attributed friction to the interlocking of asperities; however, it is now understood that the phenomenon is far more complicated. Although the theories of friction are new, its laws have changed little since Leonardo da Vinci's time: (1) The force of friction is proportional to the normal load pressing the surfaces together. (2) The force of friction is independent of the apparent area of contact. The coefficient of friction (the ratio of friction force to normal force) ranges, under normal, dry conditions, from small (for example, 0.04 for Teflon on steel) to large (for example, 1.1 for cast iron on cast iron). *See* FRICTION; IRON; STEEL.

Lubrication. When clean surfaces are brought into contact, their coefficient of friction decreases drastically if even a single molecular layer of a foreign substance (for example, an oxide) is introduced between the surfaces. For thicker lubricant films, the coefficient of friction can be quite small and no longer dependent on the properties of the surfaces but only on the bulk properties of the lubricant. Most common lubricants are liquids and gases, but solids such as molybdenum disulfide or graphite may be used. The machine elements that accommodate relative motion through the introduction of a lubricant between the contacting surfaces are named bearings. Bearings lubricated by liquids or gases operate either in the hydrostatic (externally pressurized) mode or in the hydrodynamic (self-acting) mode. Conformal hydrostatic or hydrodynamic bearings (for example, thrust or journal bearings) remain rigid, as the pressures are relatively small. In self-acting counterformal bearings (ball or roller bearings), the pressures are large (1–4 GPa), necessitating that both the elastic deformation of the surfaces and the pressure dependence of viscosity be taken into account during performance calculations; these bearings are said to operate in the elastohydrodynamic mode. *See* ANTIFRICTION BEARING; LUBRICATION; VISCOSITY.

Wear. This is the progressive loss of substance of one body because of rubbing by another body. There are many different types of wear, including sliding wear, abrasive wear, corrosion, and surface fatigue. The scientific study of wear is of recent origin and is made possible by the development of sensitive experimental equipment and techniques. The same type of complicated surface interactions that cause friction also cause wear; thus quantitative prediction of wear rates involves the same difficulties as that of friction. Actually, the situation is more severe: whereas the friction coefficient changes at most by one order of magnitude from one metal pair to another, the wear coefficient changes by several orders of magnitude depending on environmental and other conditions. *See* ABRASIVE; CORROSION; WEAR.

Andras Z. Szeri

Bibliography. B. Bhushan (ed.), *Fundamentals of Tribology and Bridging the Gap between the Macro and Micro/Nanoscales*, 2001; B. Bhushan (ed.), *Modern Tribology Handbook*, 2 vols., 2000; B. Bhushan, *Principles and Applications of Tribology*, 1999; K. C. Ludema, *Friction, Wear, Lubrication: A Textbook in Tribology*, 1996; E. Rabinowicz, *Friction and Wear of Materials*, 1995; A. Z. Szeri, *Fluid Film Lubrication: Theory and Design*, 1998; A. Z. Szeri, *Tribology: Friction, Lubrication and Wear*, 1980.

Trichasteropsida

A monospecific order of Asteroidea established for *Trichasteropsis wiessmanni*, the only Triassic asteroid known from articulated specimens. It is a small starfish with a relatively large disc and short arms. The skeleton is differentiated into marginals, actinals, and abactinals, the marginals comprising a single series of blocklike ossicles. The abactinal surface is composed of larger stellate plates and smaller rods and granules; a carinal row is present. Oral plates are large and well developed. Pedicellariae are not present. *Trichasteropsis* is the most primitive post-Paleozoic asteroid known and lies close to the latest common ancestor of all living asteroids. It comes from the Muschelkalk (Middle Triassic, Anisian-Ladinian) of Germany. *See* ASTEROIDEA; ECHINODERMATA.

Andrew B. Smith

Bibliography. D. B. Blake, *J. Nat. Hist.*, 21:481–528, 1987.

Trichomycetes

A polyphyletic class of Eumycota in the subdivision Zygomycotina, containing the orders Amoebidiales Asellariales, Eccrinales, and Harpellales. These orders are grouped together because they usually exist only as commensals in the mid- and hindgut of arthropods (*Amoebidium parasiticum* can be found on the outside of its hosts).

Asexual reproduction is accomplished by the formation of trichospores (Harpellales), arthrospores (Asellariales), sporgiospores (Eccrinales), or ameboid cells, cystospores, or rigid-walled spores (Amoebidiales). Sexual reproduction (zygospore formation) is known only in Harpellales, although conjugations have been observed in Asellariales.

The thallus of Amoebidiales is aseptate, but is regularly septate in the other three orders; septa with plugs in the lenticular cavities are produced by

Asellariales and Harpellales. Similar septa and plugs are formed by Dimargaritales and Kickxellales (Zygomycetes). *See* ZYGOMYCETES.

Classification is based on the type of reproduction; thallus branching pattern, complexity, and septation; and nature of the holdfast. Although most mycologists include Amoebidiales in the class Trichomycetes because of morphology and habitat, possibily the formation of ameboid spores renders the order unrelated to the other members. Eccrinales are treated either as the most advanced Trichomycetes or as nonrelated organisms. Because of similarities in septum and septal plug structure, and in general architecture of the asexual reproductive structures, Asellariales and Harpellales (Trichomycetes) and Dimargaritales and Kickxellales (Zygomycetes) may be more closely related to one another than to other orders in the two classes.

Trichomycetes occur worldwide and may be found anywhere a suitable host exists; they inhabit a more or less equatic environment (the gut); and the spores are discharged with feces. A few species have been cultured [*A. parasiticum* (Amoebidiales), some Harpellales], but most species are found in the host. Infection of a host is probably by the chance ingestion of a spore that lands on a food source.

More than 30 species of the Harpellales and one species of Amoebidiales (*A. parasiticum*) have been cultured. Most species are known only from their existence in the host. *See* EUMYCOTA; FUNGI; ZYGOMYCOTA.

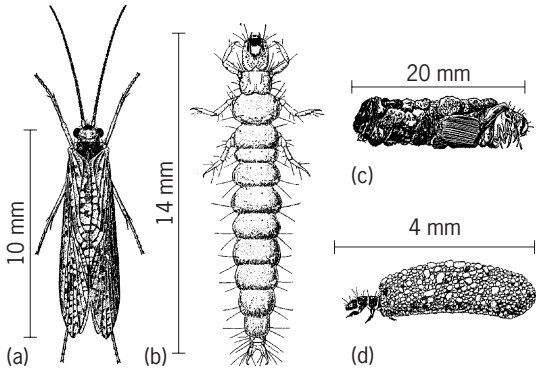
Gerald L. Benny

Bibliography. L. R. Batra, *Insect-Fungus Symbiosis*, 1979; D. H. Howard and J. D. Miller (eds.), *The Mycota*, vol. 6, 1996; R. W. Lichtwardt, *The Trichomycetes*, 1986; D. J. McLaughlin, E.G. McLaughlin, and P. A. Lemke (eds.), *The Mycota*, vol. 7, 1999; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Trichoptera

An aquatic order of the class Insecta commonly known as the caddis flies. The adults have two pairs of well-veined hairy wings, long antennae, and mouthparts capable of lapping only liquids (**illus.** *a*). The larvae are wormlike, with distinct heads, three pairs of legs on the thorax, and a pair of hook-bearing legs at the end of the body (illus. *b*). The pupae are delicate, with free appendages held close to the body, and have a pair of sharp mandibles, or jaws, which are used to cut and exit from the cocoon.

The adults live several weeks to several months, the females rapidly becoming mature. The crawl into water and lay eggs under stones and other objects. The eggs hatch into aquatic larvae. Some larvae construct a fixed retreat and some type of nest in cracks or crevices. Others build a portable case in which to live, and a few build neither, crawling about in moss and under stones. Most caddis-fly larvae are omnivorous, feeding on algae, other microorganisms, or other aquatic animals which are small enough for them to devour. A few forms are entirely predacious and live on other aquatic insect larvae. When full grown, the free-living and retreat-making larvae spin an oval cocoon under a rock or in a crevice and pupate in it. The case-makers anchor the case securely to some object in the water and pupate inside it. When mature, the pupa, using its jaws, cuts its way out of the cocoon (illus. *c* and *d*), swims to the surface, sometimes climbing out of the water on a stem or stone, and there the adult emerges from the pupal skin.



Life cycle of Trichoptera. (*a*) Adult and (*b*) the free-living larva of the widespread genus *Rhyacophila*. (*c*) Head and thorax of larva protruding from purselike case of a micro caddis fly, *Ochrotrichia*. (*d*) Head and thorax of larva protruding from case of a large caddis fly, *Limnephilus*. (*Illinois Natural History Survey*)

Except for a brackish-water species in New Zealand and a few moss-inhabiting species in Europe and North America, caddis flies occur only in fresh water. They abound in cold or running water relatively free from pollution. Altogether they compose a large and important segment of the biota of such habitats and of the fish feed economy.

The Trichoptera include about 10,000 described species, divided into 34 families, and occur in practically all parts of the world. The order probably arose over 2×10^8 years ago in early Mesozoic time, for fossils of typical trichopteran wings have been found in Late Triassic deposits. Many existing genera are probably of Cretaceous origin; hence, it seems almost certain that representatives of all the diverse family lines evolved during the middle part of the Mesozoic Era at the same time that the dinosaurs were proliferating. The Trichoptera were originally cool-adapted animals, as most of the primitive forms still are, but many warm-adapted lines have evolved. As a result, the caddis flies are found in arctic, temperate, and tropical habitats. *See* INSECTA.

Herbert H. Ross

Bibliography. C. Betten, *The Caddis Flies or Trichoptera of New York State*, N. Y. State Mus. Bull. 292, 1934; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; H. H. Ross, *The Caddisflies, or Trichoptera, of Illinois*, Illinois Nat. Hist. Surv. Bull. 23, 1944; H. H. Ross, *The Evolution and Classification of the Mountain Caddisflies*, 1956; G. Ulmer, Trichoptera, *Genera Insectorum*, 60:1–259, 1907; R. L. Usinger (ed.), *Aquatic Insects of California*, 1956.

Trichroism

When certain optically anisotropic transparent crystals are subjected to white light, a cube of the material is found to transmit a different color through each of the three pairs of parallel faces. Such crystals are sometimes termed trichroic, and the phenomenon is called trichroism. This expression is used only rarely today since the colors in a particular crystal can appear quite different if the cube is cut with a different orientation with respect to the crystal axes. Accordingly, the term is frequently replaced by the more general term pleochroism. Even this term is being replaced by the phrase linear dichroism or circular dichroism to correspond with linear birefringence or circular birefringence. *See* BIREFRINGENCE; DICHROISM; PLEOCHROISM.

Cordierite is a typical trichroic crystal. In light with a vibration direction parallel to the *X* axis of the index ellipsoid, the crystal appears yellow. With the vibration direction parallel to the *Y* axis, the crystal is dark violet. In the *Z* direction the crystal is clear.

The phenomena of trichroism can be explained crudely as follows. Classically, one can consider an electron in a biaxial crystal as having three different force constants associated with a displacement directed along each of the principal axes. Linear polarized light traveling along the *X* axis with its electric vector parallel to the *Y* axis will displace the electron against the *Y* force constant and will experience a certain absorption and retardation. It will be unaffected by the force constants in the *X* and *Z* directions. Similarly, polarized light traveling in the *Y* direction will experience absorption and retardation. Unpolarized light will also be absorbed in a different fashion depending on the direction of propagation. In this case, light traveling in the *X* direction can be considered as composed of an equal mixture of light polarized parallel to the *Y* axis and the *Z* axis. The absorption will be intermediate between the two polarization directions. *See* CRYSTAL OPTICS; POLARIZED LIGHT.

Bruce H. Billings

Bibliography. E. E. Wahlstrom, *Optical Crystallography*, 5th ed., 1979; E. A. Wood, *Crystals and Light: An Introduction to Optical Crystallography*, 1977; A. Yariv and P. Yeh, *Optical Waves in Crystals*, 1983.

Tricladida

An order of the Turbellaria (of the phylum Platyhelminthes) known commonly as planaria, which are several millimeters to 20 in. (50 cm) or more in length. They have a diverticulated intestine with a single anterior branch and two posterior branches separated by a plicate pharynx or pharynges. Rhabdites are numerous and, except in cave planarians, two to many eyes are present. The much branched protonephridial tubules form a network with numerous nephridiopores on each side of the body. The female reproductive system includes a single pair of small anteriorly located ovaries, numerous minute yolk glands arranged in clusters along either side of the body, the common ducts, the female antrum, and usually one or more bursae. The male system has several to many testes, which are lateral in position and connected with a single sperm duct on each side. These ducts empty either directly or after fusion into the copulatory organ which lies in the male antrum. Following copulation and mutual insemination, capsules containing several fertilized eggs are attached to objects in the water and hatch in 2 or more weeks into young worms. Asexual reproduction by fission is common in forms such as the cosmopolitan *Dugesia tigrina* which has been much used in studies on regeneration. Fragmentation and regeneration is the usual method of reproduction in some land planarians such as *Bipalium kewense*, an exotic species which has become established through much of the southern United States. The marine planarian *Bdelloura candida* is a commensal on the horseshore crab. *See* TURBELLARIA.

E. Ruffin Jones

Bibliography. L. von Graff, Tricladida, in H. G. Bronn (ed.), *Klassen und Ordnungen des Tierrechs*, vol. 4, pt. 2, 1912–1917; R. C. Harrel, D. L. Bechler, and R. E. Ogren, First Texas record of *Geoplana arkalabamensis* (Turbellaria: Tricladida: Terricola) with a review of other land planarians reported in the state, *Tex. J. Sci.*, 46:45–49, 1994; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; L. Winsor, The biodiversity of terrestrial flatworms (Tricladida: Terricola) in Queensland: A preliminary report, *Mem. Mus. Victoria*, 56:575–579, 1997.

Trigger circuit

An electronic circuit that generates or modifies an existing waveform to produce a pulse of short time duration with a fast-rising leading edge. This waveform, or trigger, is normally used to initiate a change of state of some relaxation device, such as a multivibrator. The most important characteristic of the waveform generated by a trigger circuit is usually the fast leading edge. The exact shape of the falling portion of the waveform often is of secondary importance, although it is important that the total duration time is not too great. A pulse generator such as a blocking oscillator may also be used and identified as a trigger circuit if it generates sufficiently short pulses. *See* PULSE GENERATOR.

Peaking (differentiating) circuits. These circuits, which accent the higher-frequency components of a pulse waveform, cause sharp leading and trailing edges and are therefore used as trigger circuits. The simplest form of peaking circuits are the simple *RC* (resistance-capacitance) and *RL* (resistance-inductance) networks shown in **Fig. 1**. If a steep wavefront of amplitude *V* is applied to either of these circuits, the output will be a sudden rise followed by an exponential decay according to the equation $v_0 = V\epsilon^{-kt}$, where $k = 1/RC$ or R/L.

These circuits are often called differentiating

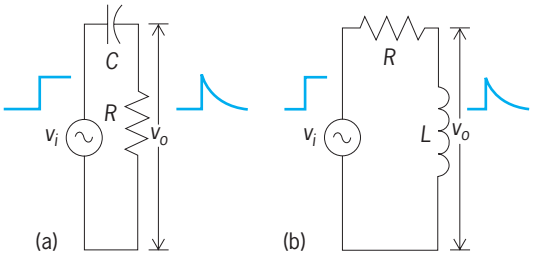


Fig. 1. Simple peaking circuits: (a) resistance-capacitance network; (b) resistance-inductance network.

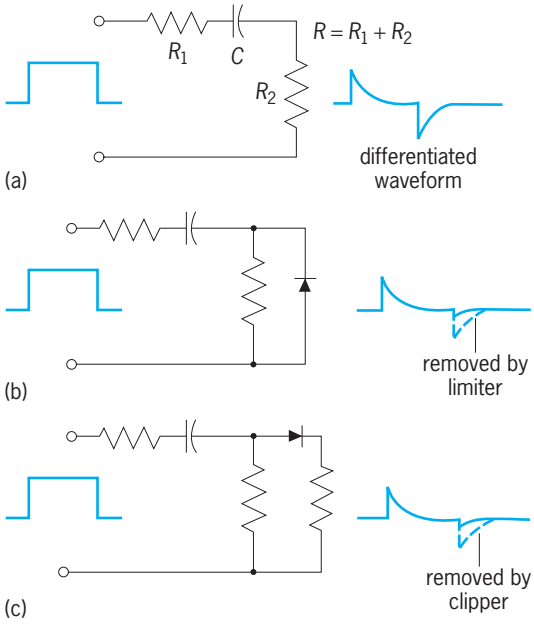


Fig. 2. Differentiated pulses. (*a*) Basic circuit. (*b*) Limiting unwanted portion. (*c*) Clipping unwanted portion.

circuits because the outputs are rough approximations of the derivative of the input waveforms, if the *RC* or *R/L* time constant is sufficiently small.

If the pulse is applied to the differentiating circuits, the resultant waveform shown in **Fig. 2** may be used as a trigger. It is sometimes necessary, however, to remove by limiting or clipping the undesired portion of the waveform to prevent circuits from responding to it.

The *RL* circuit of Fig. 1 cannot be considered in its simplest form when extremely fast rise times are required because of the distributed capacitance and small series resistance associated with the inductance. A more accurate representation of the circuit is that in **Fig. 3**. The response is limited as shown for a fixed value of L and C. The value for $k = 1$ is referred to as critical damping. A value of k slightly less than unity provides a pulse that is a suitable trigger for many applications.

Ringing circuits. A circuit of the form shown in Fig. 3 that is highly underdamped, or oscillatory ($k \ll 1$), and is supplied with a step or pulse input is often referred to as a ringing circuit. When used in the output of a field-effect or bipolar transistor as in **Fig. 4**, this circuit can be used as a trigger circuit. When the input pulse is applied, current in the output circuit is immediately cut off. Since the current in L cannot change instantaneously, it flows in the LC circuit in an oscillatory manner, gradually decaying because of the resistance in the circuit. However, if the diode is in the circuit, the circuit will be highly overdamped for the negative portion of the oscillatory waveform, and the oscillations will be damped out as shown.

If a transistor is used as the current source and operated near saturation, damping will take place when the transistor goes into saturation as shown in **Fig. 5**.

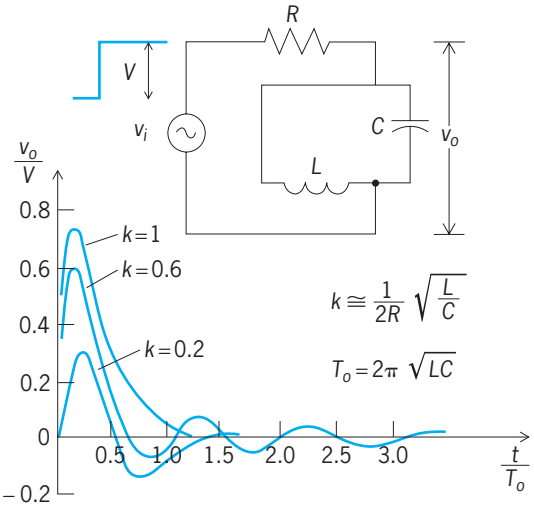


Fig. 3. Resistance-inductance-capacitance peaking circuit.

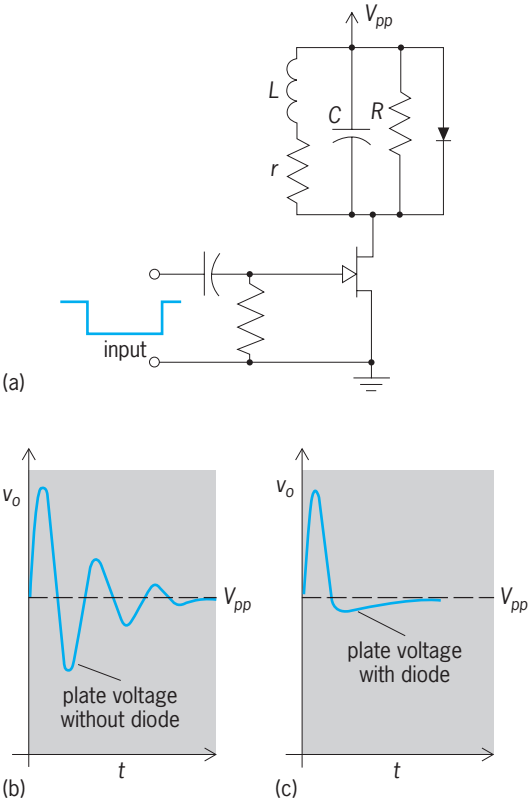


Fig. 4. Ringing circuit as trigger source. (*a*) Circuit diagram. (*b*) Drain-voltage waveform without diode limiter. (*c*) Drain-voltage waveform with diode limiter.

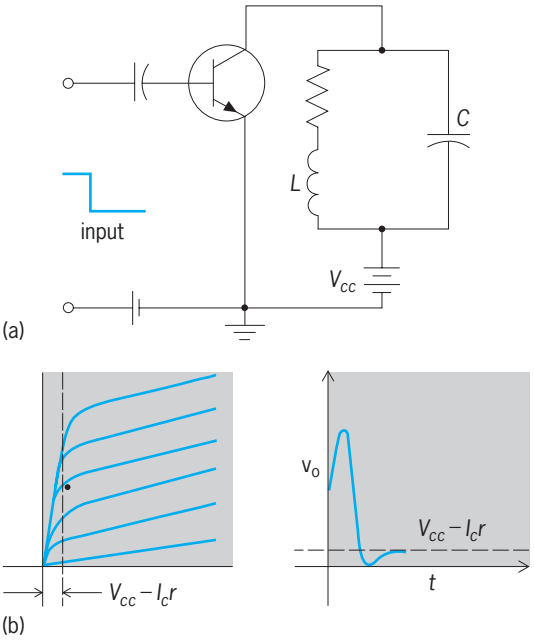


Fig. 5. Ringing circuit with transistor saturation damping: (*a*) circuit diagram; (*b*) waveforms.

The diode is not required. For other waveforms *See* WAVE-SHAPING CIRCUITS. Glenn M. Glasford

Bibliography. D. Christiansen, *Electronics Engineers' Handbook*, 4th ed., 1996; J. Millman and A. Grabel, *Microelectronics*, 2d ed., 1987; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997.

Triglyceride (triacylglycerol)

A simple lipid. Triglycerides are fatty acid triesters of the trihydroxy alcohol glycerol which are present in plant and animal tissues, particularly in the food storage depots, either as simple esters in which all the fatty acids are the same or as mixed esters in which the fatty acids are different. The triglycerides constitute the main component of natural fats and oils.

The generic formula of a triglyceride is shown below, where RCO_2H, $R'CO_2H$, and $R''CO_2H$ repre-

$$\begin{array}{l} CH_2-OOC-R \\ | \\ CH-OOC-R' \\ | \\ CH_2-OOC-R'' \end{array}$$

sent either the same or different fatty acids, such as butyric or caproic (short chain), palmitic or stearic (long chain), oleic, linoleic, or linolenic (unsaturated). Saponification with alkali releases glycerol and the alkali metal salts of the fatty acids (soaps). The triglycerides in the food storage depots represent a concentrated energy source, since oxidation provides more energy than an equivalent weight of protein or carbohydrate.

Animal and vegetable triglycerides contain predominantly even-chain-length fatty acids, with palmitic and oleic acids as the main components. Since n fatty acids may be esterified in $(n^3 + n^2)/2$ ways into glycerol, and since natural fats contain a variety of fatty acids, the number of component triglycerides of a relatively simple natural fat or oil may be high. Some pure simple and mixed triglycerides have been isolated from natural fats by fractional crystallizations at low temperatures, but in general physical methods are not yet available for the separation of naturally occurring mixtures. Several theories, such as those of even distribution and partial random distribution, have been advanced to account for the distribution of the fatty acids in the triglycerides. Many synthetic triglycerides have been prepared, and the study of the physical properties of these compounds has provided much useful information. Melting-point, x-ray-diffraction, and infrared-spectroscopy investigations have shown that triglycerides may exist in at least three polymorphic modifications. *See* MOLECULAR STRUCTURE AND SPECTRA.

The physical and chemical properties of fats and oils depend on the nature of the fatty acids present. Saturated fatty acids give higher-melting fats and represent the main constituents of solid fats, for example, lard and butter. Unsaturation lowers the melting point of fatty acids and fats. Thus, in the oil of plants, unsaturated fatty acids are present in large amounts, for example, oleic acid in olive oil and linoleic and linolenic acids in linseed soil. Oils are hydrogenated commercially to produce the proper consistency and melting point for use as edible fats. *See* CARBOXYLIC ACID; FAT AND OIL (FOOD); LIPID.

Roy H. Gigg; Herbert E. Carter

Bibliography. G. Fuller and W. D. Nes (eds.), *Ecology and Metabolism of Plant Lipids*, 1986; F. J. Mead et al., *Lipids: Chemistry, Biochemistry, and Nutrition*, 1986; P. Quinn and J. Harwood (eds.), *Plant Lipid Biochemistry, Structure, and Utilization*, 1991.

Trigonometry

The study of triangles and the trigonometric functions. One common use for trigonometry is to measure heights and distances that are awkward or impossible to measure by ordinary means. Surveyors use it to find heights of mountains and distances

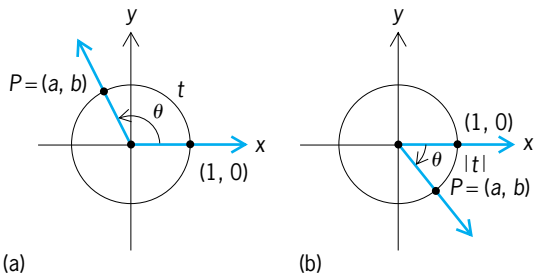


Fig. 1. Point *P* on the unit circle corresponding to $\theta = t$ radians. (*a*) $t \geq 0$: length of arc from (1,0) to *P* is *t* units. (*b*) $t < 0$: length of arc from (1,0) to *P* is $|t|$ units.

TABLE 1. Values of trigonometric functions at integral multiples of $\pi/4$ (90°)

θ, radians	θ	$\sin\theta$	$\cos\theta$	$\tan\theta$	$\csc\theta$	$\sec\theta$	$\cot\theta$
0	0°	0	1	0	Not defined	1	Not defined
$\pi/2$	90°	1	0	Not defined	1	Not defined	0
π	180°	0	−1	0	Not defined	−1	Not defined
$3\pi/2$	270°	−1	0	Not defined	−1	Not defined	0

across lakes and countries; engineers use it in the design of large structures and roads; astronomers use it in accurate measurements of the time and in locating the position of objects in the sky; and navigators on the sea and in the air use it to find latitudes, longitudes, and direction. Trigonometry has evolved from use by surveyors, engineers, and navigators to applications involving ocean tides, the rise and fall of food supplies in certain ecologies, brainwave patterns, the analysis of alternating-current electricity, and many other phenomena of a vibratory character.

Plane trigonometry. Plane trigonometry mostly deals with the relationships among the three sides and three angles of a triangle that lies in a plane.

A ray is that portion of a line that starts at a point on the line and extends indefinitely in one direction. The starting point of a ray is called its vertex. If two rays are drawn with a common vertex, they form an angle. One of the rays of an angle is called the initial side, and the other ray is the terminal side. The angle that is formed is identified by showing the direction and amount of rotation from the initial side to the terminal side. If the rotation is in the counterclockwise direction, the angle is positive; if the rotation is clockwise, the angle is negative.

The angle formed by rotating the initial side exactly once in the counterclockwise direction until it coincides with itself (1 revolution) is said to measure 360 degrees, written 360°. Thus, one degree, 1°, is 1/360 of a revolution. One-sixtieth of a degree is called a minute, written 1′.

By using a circle of radius r, an angle can be constructed whose vertex is at the center of this circle and whose rays subtend an arc on the circle whose length equals r. Such an angle measures 1 radian. For a circle of radius r, a central angle of θ radians subtends an arc whose length s is given by Eq. (1).

$$s = r\theta \tag{1}$$

Because a central angle of 1 revolution (360°) subtends an arc equal to the circumference of the circle ($2\pi r$), it follows that an angle of 1 revolution equals 2π radians; that is, 2π radians $= 360°$. *See* PLANE GEOMETRY; RADIAN MEASURE.

Trigonometric functions. A unit circle is a circle whose radius is one and whose center is at the origin of a rectangular system of coordinates. For the unit circle, Eq. (1) states that a central angle of θ radians subtends an arc whose length $s = \theta$. If t is any real number, let θ be the angle equal to t radians and P be the point on the unit circle that is also on the terminal side of θ. If $t \geq 0$, then the point P is reached by moving counterclockwise along the unit circle, starting at the point with coordinates $(1, 0)$, for a length of arc equal to t units (**Fig. 1***a*). If $t < 0$, this point P is reached by moving clockwise along the unit circle beginning at (1,0), for a length of arc equal to $|t|$ units (Fig. 1*b*). Thus, to each real number t there corresponds a unique point $P = (a, b)$ on the unit circle. The coordinates of this point P are used to define the six trigonometric functions: If $\theta = t$ radians, the sine, cosine, tangent, cosecant, secant, and cotangent of θ, respectively abbreviated as $\sin\theta$, $\cos\theta$, $\tan\theta$, $\csc\theta$, $\sec\theta$, $\cot\theta$, are given by Eqs. (2), (3), and (4).

$$\sin\theta = b \qquad \cos\theta = a \tag{2}$$

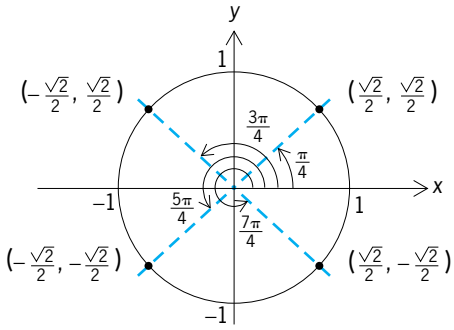
$$\text{if } a \neq 0, \quad \tan\theta = b/a, \quad \sec\theta = 1/a \tag{3}$$

$$\text{if } b \neq 0, \quad \tan\theta = a/b, \quad \sec\theta = 1/b \tag{4}$$

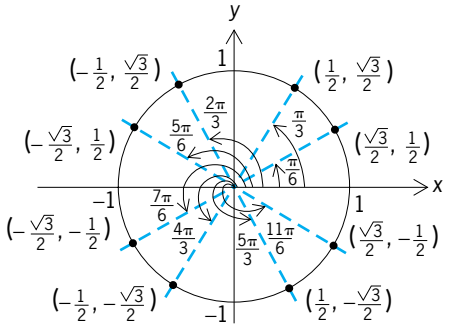
See COORDINATE SYSTEMS.

TABLE 2. Values of trigonometric functions at integral multiples of $\pi/6$ (30°), $\pi/4$ (45°), and $\pi/3$ (60°)

θ, radians	θ	$\sin\theta$	$\cos\theta$	$\tan\theta$	$\csc\theta$	$\sec\theta$	$\cot\theta$
$\pi/6$	30°	1/2	$\sqrt{3}/2$	$\sqrt{3}/3$	2	$2\sqrt{3}/3$	$\sqrt{3}$
$\pi/4$	45°	$\sqrt{2}/2$	$\sqrt{2}/2$	1	$\sqrt{2}$	$\sqrt{2}$	1
$\pi/3$	60°	$\sqrt{3}/2$	1/2	$\sqrt{3}$	$2\sqrt{3}/3$	2	$\sqrt{3}/3$
$2\pi/3$	120°	$\sqrt{3}/2$	−1/2	$-\sqrt{3}$	$2\sqrt{3}/3$	−2	$-\sqrt{3}/3$
$3\pi/4$	135°	$\sqrt{2}/2$	$-\sqrt{2}/2$	−1	$\sqrt{2}$	$-\sqrt{2}$	−1
$5\pi/6$	150°	1/2	$-\sqrt{3}/2$	$-\sqrt{3}/3$	2	$-2\sqrt{3}/3$	$-\sqrt{3}$
$7\pi/6$	210°	−1/2	$-\sqrt{3}/2$	$\sqrt{3}/3$	−2	$-2\sqrt{3}/3$	$\sqrt{3}$
$5\pi/4$	225°	$-\sqrt{2}/2$	$-\sqrt{2}/2$	1	$-\sqrt{2}$	$-\sqrt{2}$	1
$4\pi/3$	240°	$-\sqrt{3}/2$	−1/2	$\sqrt{3}$	$-2\sqrt{3}/3$	−2	$\sqrt{3}/3$
$5\pi/3$	300°	$-\sqrt{3}/2$	1/2	$-\sqrt{3}$	$-2\sqrt{3}/3$	2	$-\sqrt{3}/3$
$7\pi/4$	315°	$-\sqrt{2}/2$	$\sqrt{2}/2$	−1	$-\sqrt{2}$	$\sqrt{2}$	−1
$11\pi/6$	330°	−1/2	$\sqrt{3}/2$	$-\sqrt{3}/3$	−2	$2\sqrt{3}/3$	$-\sqrt{3}$



(a)



(b)

Fig. 2. Angles and coordinates of corresponding points on the unit circle. (*a*) Angles that are integral multiples of $\pi/4$ (45°). (*b*) Angles that are integral multiples of $\pi/6$ (30°) and $\pi/3$ (60°).

For example, for $\theta = 0$, the point (1,0) is on the terminal side of θ and is on the unit circle so that Eqs. (5) hold, with csc 0 and cot 0 not defined. The

$$\sin 0 = 0 \quad \cos 0 = 1 \quad \tan 0 = 0 \quad \sec 0 = 1 \qquad (5)$$

trigonometric functions of angles that are integral multiples of $\pi/2$ (90°) are found similarly (**Table 1**).

The coordinates of points on the unit circle that are on the terminal sides of angles that are integral multiples of $\pi/6$ (30°), $\pi/4$ (45°), and $\pi/3$ (60°) can be found (**Fig. 2**). With Eqs. (2), (3), and (4), the trigonometric functions of these angles are obtained (**Table 2**).

It is not necessary to use a unit circle to define the trigonometric functions. This can be done by considering a point $P = (a,b)$ on the terminal side of the angle θ a distance r from the origin, so that P lies on the circle $x^2 + y^2 = r^2$ (**Fig. 3**). There is a corresponding point $P^* = (a^*,b^*)$ that is also on the terminal side of the angle θ but at unit distance from the origin, so that P^* lies on the unit circle. To P and P^* correspond the points $A = (a,0)$ and $A^* = (a^*,0)$, which lie at the feet of perpendiculars to the x axis from P and P^* respectively. Since the triangles OA^*P^* and OAP are similar, ratios of corresponding sides are equal and Eqs. (6), (7), and (8) hold.

$$\sin\theta = b^*/1 = b/r \quad \cos\theta = a^*/1 = a/r \qquad (6)$$

if a $\neq 0$,

$$\tan\theta = b^*/a^* = b/a \quad \sec\theta = 1/a^* = r/a \qquad (7)$$

if b $\neq 0$,

$$\cot\theta = a^*/b^* = a/b \quad \csc\theta = 1/b^* = r/b \qquad (8)$$

Properties of trigonometric functions. Based on Eqs. (2) and the above geometric construction (Fig. 1) for $\sin\theta$ and $\cos\theta$, θ can be any angle, so the domain of the sine and cosine functions is all real numbers. In Eqs. (3), if $a = 0$, the tangent and secant functions are not defined, so the domain of these functions is all real numbers, except odd multiples of $\pi/2$ (90°). In Eqs. (4), if $b = 0$, the cotangent and cosecant functions are not defined, so the domain of these functions is all real numbers, except multiples of π (180°). Also, since $|a| \leq 1$ and $|b| \leq 1$, the range of the sine and cosine functions is -1 to 1 inclusive. Since $|b| = |\sin\theta| \leq 1$ and $|a| = |\cos\theta| \leq 1$, it follows that $|\csc\theta| = 1/|b| \geq 1$ and $|\sec\theta| = 1/|a| \geq 1$. Thus the range of the secant and cosecant functions consists of all real numbers less than or equal to -1 or greater than or equal to 1. The range of both the tangent and cotangent functions consists of all real numbers.

Equations (2), (3), and (4) also reveal the reciprocal identities, given in Eqs. (9). Two other useful identities, given in Eqs. (10), also follow.

$$\csc\theta = 1/\sin\theta \quad \sec\theta = 1/\cos\theta \quad \cot\theta = 1/\tan\theta \qquad (9)$$

$$\tan\theta = \sin\theta/\cos\theta \quad \cot\theta = \cos\theta/\sin\theta \qquad (10)$$

Since (a,b) is on the unit circle, $a^2 + b^2 = 1$, and so $(\sin\theta)^2 + (\cos\theta)^2 = 1$. This is called a pythagorean

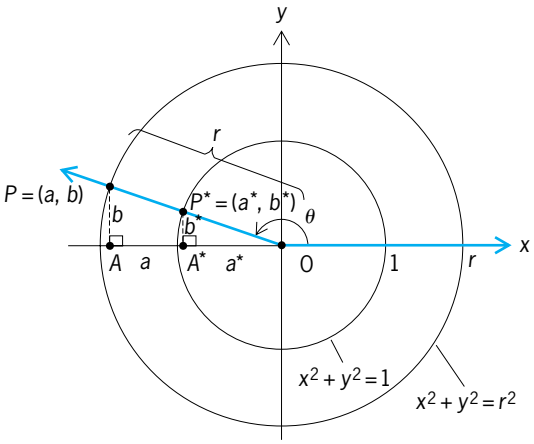


Fig. 3. Point *P* on a circle of radius *r* corresponding to angle θ, and corresponding point P^* on the unit circle.

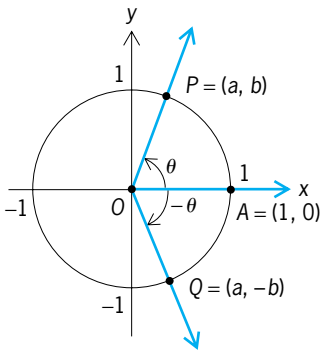


Fig. 4. Relationship between angles θ and $-\theta$, and coordinates of corresponding points on the unit circle.

TABLE 3. Signs of the trigonometric functions

Quadrant of point P	$\sin\theta$, $\csc\theta$	$\cos\theta$, $\sec\theta$	$\tan\theta$, $\cot\theta$
I	Positive	Positive	Positive
II	Positive	Negative	Negative
III	Negative	Negative	Positive
IV	Negative	Positive	Negative

identity and is written as Eq. (11).

$$\sin^2\theta + \cos^2\theta = 1 \tag{11}$$

As discussed above, for a given angle θ, measured in radians, there is a corresponding point $P = (a,b)$ on the unit circle. If 2π is added to θ (or any multiple k of 2π to θ), the point on the unit circle corresponding to the angle $\theta + 2\pi k$ is identical to P, and therefore the trigonometric functions of this angle are also the same. A function f is called periodic if there is a positive number p so that $f(\theta + p) = f(\theta)$ for all θ. The smallest number p for which this equation holds is called the period of f. The sine, cosine, secant, and cosecant functions are periodic with period 2π, whereas the tangent and cotangent functions have period π.

Since $\sin\theta = b$, $\cos\theta = a$, $\sin(-\theta) = -b$, and $\cos(-\theta) = a$ (**Fig. 4**), the even-odd properties given in Eqs. (12) and (13) follow.

$$\sin(-\theta) = -\sin\theta \quad \cos(-\theta) = \cos\theta$$
$$\tan(-\theta) = -\tan\theta \tag{12}$$

$$\csc(-\theta) = -\csc\theta \quad \sec(-\theta) = \sec\theta$$
$$\cot(-\theta) = -\cot\theta \tag{13}$$

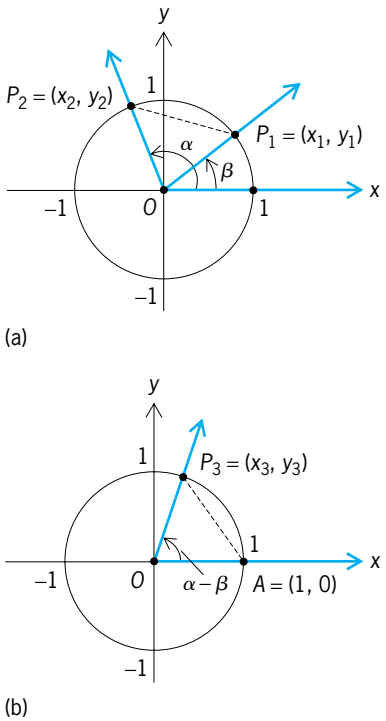


Fig. 5. Geometric construction used to prove the difference formula for cosines. (*a*) Distance between P_1 and P_2 equals (*b*) distance between A and P_3.

Signs of trigonometric functions. Once again, for an angle θ there is a corresponding point $P = (a,b)$ on the unit circle. If the quadrant in which the point P lies is known, then the signs of the trigonometric functions of θ can be determined (**Table 3**).

TABLE 4. Plot points for graph of the sine function

x	$y = \sin x$	(x,y)
0	0	(0,0)
$\pi/6$	1/2	$(\pi/6, 1/2)$
$\pi/2$	1	$(\pi/2, 1)$
$5\pi/6$	1/2	$(5\pi/6, 1/2)$
$\pi\gamma$	0	$(\pi, 0)$
$7\pi/6$	$-1/2$	$(7\pi/6, -1/2)$
$3\pi/2$	-1	$(3\pi/2, -1)$
$11\pi/6$	$-1/2$	$(11\pi/6, -1/2)$
2π	0	$(2\pi, 0)$

Sum and difference formulas. The sum and difference formulas for the cosine function are given in Eqs. (14) and (15). A geometric construction (**Fig. 5**)

$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta \tag{14}$$

$$\cos(\alpha - \beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta \tag{15}$$

is used to prove Eqs. (15). The points $A = (1,0)$, P_1, P_2, and P_3 lie on the unit circle, with P_1, P_2, and P_3 on the terminal sides of the angles β, α, and $\alpha - \beta$ respectively. Then the coordinates of P_1, P_2, and P_3 are given in Eqs. (16). The distance formula is used

$$P_1 = (\cos\beta, \sin\beta) \quad P_2 = (\cos\alpha, \sin\alpha)$$
$$P_3 = (\cos(\alpha - \beta), \sin(\alpha - \beta)) \tag{16}$$

to express the equality of the distances $|P_1, P_2|$ and $|AP_3|$. Then both sides of the resulting equation are squared, and Eq. (11) is used to arrive at the difference formula in Eq. (15). The fact that $\alpha + \beta = \alpha - (-\beta)$, and applying Eqs. (12) in Eq. (15), are used to obtain the sum formula in Eq. (14). *See* ANALYTIC GEOMETRY.

If $\alpha = \pi/2$ and $\beta = \theta$ are substituted in Eq. (15), the use of the values $\sin(\pi/2) = 1$ and $\cos(\pi/2) = 0$ (Table 1) yields Eq. (17). Likewise, if $\alpha = \theta - \pi/2$

$$\cos(\pi/2 - \theta) = \sin\theta \tag{17}$$

and $\beta = \pi/2$ are substituted in Eq. (14), then use of these same values and Eqs. (12) yields Eq. (18).

$$\sin(\pi/2 - \theta) = \cos\theta \tag{18}$$

If $\theta = \alpha + \beta$ is substituted in Eq. (17), then Eqs. (15), (17), and (18) can beused to obtain Eq. (19). The fact that $\alpha + \beta = \alpha - (-\beta)$ and Eqs. (12)

$$\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta \tag{19}$$

can then be used to obtain Eq. (20).

$$\sin(\alpha - \beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta \tag{20}$$

Product-to-sum formulas. The product-to-sum formulas are Eqs. (21), (22), and (23). Equation (21)

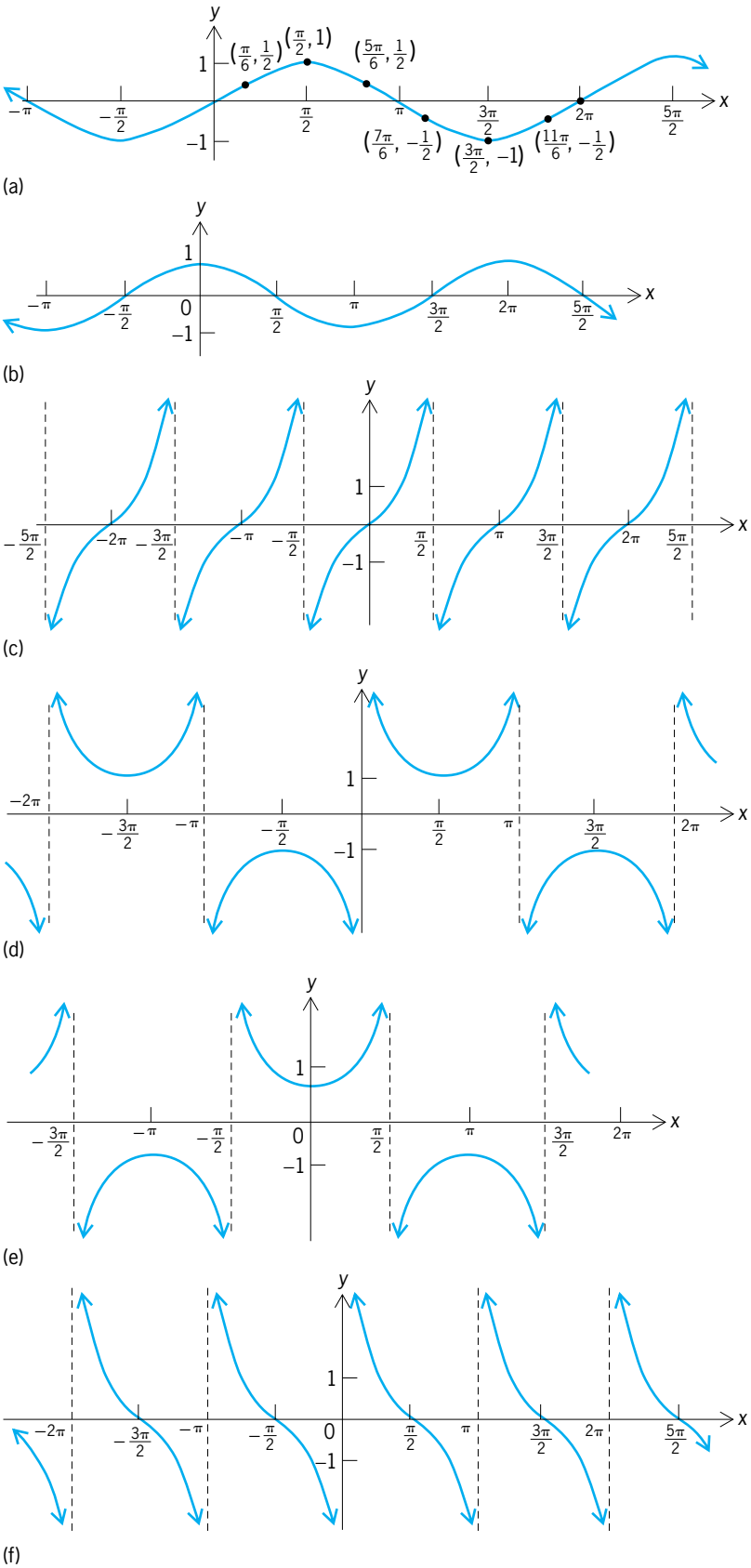


Fig. 6. Graphs of trigonometric functions. (*a*) $y = \sin x$. (*b*) $y = \cos x$. (*c*) $y = \tan x$, x not equal to odd multiples of $\pi/2$. (*d*) $y = \csc x$, x not equal to integral multiples of π. (*e*) $y = \sec x$, x not equal to odd multiples of $\pi/2$. (*f*) $y = \cot x$, x not equal to integral multiples of π.

TABLE 5. Plot points for graph of the tangent function

x	$y = \tan x$	(x,y)
$-\pi/3$	$-\sqrt{3} \approx -1.73$	$(-\pi/3, -\sqrt{3})$
$-\pi/4$	-1	$(-\pi/4, -1)$
$-\pi/6$	$-\sqrt{3}/3 \approx -0.58$	$(-\pi/6, -\sqrt{3}/3)$
0	0	(0,0)
$\pi/6$	$\sqrt{3}/3 \approx 0.58$	$(\pi/6, \sqrt{3}/3)$
$\pi/4$	1	$(\pi/4, 1)$
$\pi/3$	$\sqrt{3} \approx 1.73$	$(\pi/3, \sqrt{3})$

$$\sin\alpha \sin\beta = 1/2[\cos(\alpha - \beta) - \cos(\alpha + \beta)] \quad (21)$$

$$\cos\alpha \cos\beta = 1/2[\cos(\alpha - \beta) + \cos(\alpha + \beta)] \quad (22)$$

$$\sin\alpha \cos\beta = 1/2[\sin(\alpha + \beta) + \sin(\alpha - \beta)] \quad (23)$$

is obtained by subtracting Eq. (14) from Eq. (15); Eq. (22) is obtained by adding Eqs. (14) and (15); and Eq. (23) is obtained by adding Eqs. (19) and (20).

Graphs of trigonometric functions. Because the period of the sine function is 2π, once the graph of $y = \sin x$ is found on the interval $[0,2\pi]$, the remainder of the graph will consist of repetitions of this portion. A series of plot points (x, y), with the angle x between 0 and 2π can easily be compiled (**Table 4**) from the values of the sine function already obtained (Tables 1 and 2). These points can be plotted and connected with a smooth curve. The complete graph of $y = \sin x$ is obtained by repetition (**Fig. 6***a*). From Eq. (19) it follows that $\sin(x + \pi/2) = \cos x$, and therefore the graph of $y = \cos x$ can be obtained by shifting the graph of $y = \sin x$ to the left by $\pi/2$ units (Fig. 6*b*).

Since the tangent function has period π, once the graph of $y = \tan x$ is found on the interval $(-\pi/2, \pi/2)$, the rest of the graph will consist of repetitions of that graph. Again, a series of plot points can easily be compiled (**Table 5**). If x is close to $\pi/2$ but remains less than $\pi/2$, then $\sin x$ will be close to 1, $\cos x$ will be close to 0, and $\sin x/\cos x = \tan x$ will be positive and large, approaching ∞. Similarly, if x is close to $-\pi/2$, but remains greater than $-\pi/2$, $\tan x$ will approach $-\infty$. The graph of $y = \tan x$ (Fig. 6*c*) has vertical asymptotes at $x = \ldots, -\pi/2, \pi/2, 3\pi/2, \ldots$. The graphs of $y = \csc x$, $y = \sec x$, and $y = \cot x$ are constructed in a similar fashion (Fig. 6*d, e, f*).

The graphs of $y = A \sin \omega x$ and $y = A \cos \omega x$ are called sinusoidal graphs. Since $|\sin x| \leq 1$, it follows

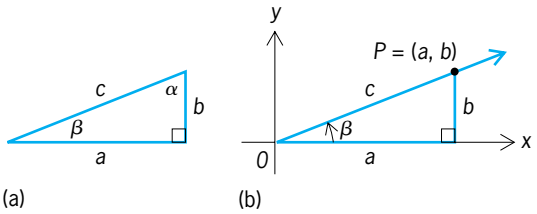


Fig. 7. Right triangle. (*a*) Labeling of sides and angles. (*b*) Relationship of sides to coordinates defining trigonometric functions.

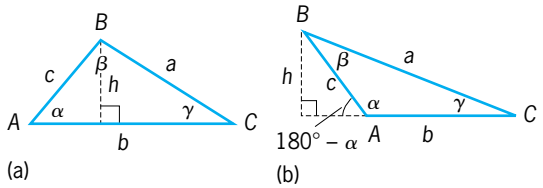


Fig. 8. Labeling of sides, angles, and vertices of oblique triangle, with altitude *h* used in proving the law of sines. (*a*) Angle α is acute. (*b*) Angle α is obtuse.

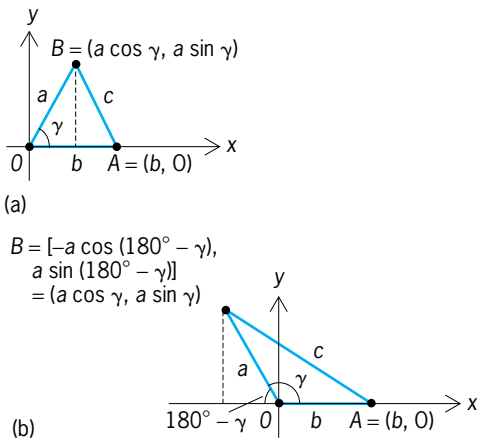


Fig. 9. Triangle and rectangular coordinates used in proving the laws of cosines. (*a*) Angle γ is acute. (*b*) Angle γ is obtuse.

that $-|A| \le A \sin \omega x \le |A|$, so the graph of $y = A \sin \omega x$ lies between $-|A|$ and $|A|$. The number $|A|$ is called the amplitude. The period of $y = A \sin \omega x$ is $2\pi/\omega$, so the graph of $y = A \sin \omega x$ repeats over any interval of length $2\pi/\omega$.

Inverse trigonometric functions. In the equation $x = \sin y$, if y is restricted so that $-\pi/2 \le y \le \pi/2$, then the solution of the equation for y is unique and is denoted by $y = \sin^{-1} x$ (read "y is the inverse sine of x"). Sometimes $y = \sin^{-1} x$ is written as $y = \text{Arcsin } x$. Thus, $y = \sin^{-1} x$ is a function whose domain is $-1 \le x \le 1$ and whose range is $-\pi/2 \le y \le \pi/2$. For example, $\sin^{-1} 1/2 = \pi/6$ and $\sin^{-1}(-1) = -\pi/2$.

Likewise in the equation $x = \cos y$, if y is restricted so that $0 \le y \le \pi$, then the solution of the equation for y is unique and is denoted by $y = \cos^{-1} x$ (read "y is the inverse cosine of x"). Thus, $y = \cos^{-1} x$ is a function whose domain is $-1 \le x \le 1$ and whose range is $0 \le y \le \pi$. Finally, in the equation $x = \tan y$, if y is restricted so that $-\pi/2 < y < \pi/2$, then the solution of the equation for y is unique and is denoted by $y = \tan^{-1} x$ (read "y is the inverse tangent of x"). Thus, $y = \tan^{-1} x$ is a function whose domain is $-\infty < x < \infty$ and whose range is $-\pi/2 < y < \pi/2$.

Solution of right triangles. The trigonometric functions can be expressed as ratios of the sides of a right triangle. Indeed, by Eqs. (6), (7), and (8), it follows that $\sin \beta = b/c$, $\cos \beta = a/c$, $\tan \beta = b/a$, and so on, where a and b are the sides adjacent to the right angle, c is the hypotenuse, and α and β are the angles opposite a and b respectively (**Fig. 7**).

If an angle and a side or else two sides of a right triangle are known, then the remaining angles and sides can be found. For any right triangle (Fig. 7*a*), Eqs. (24) hold.

$$c^2 = a^2 + b^2 \quad \alpha + \beta = 90^\circ \tag{24}$$

For example, to measure the height of a radio antenna, a surveyor walks 300 m from its base, and determines the angle of elevation to be 40°. To find the height b it is noted that $\tan 40^\circ = b/300$, so that $b = 300 \tan 40^\circ = 251.73$ m.

Other plane figures with lines as sides can often be solved by drawing perpendiculars to sides of the figure to divide it into right triangles and then solving the right triangles.

Solution of oblique triangles. If none of the angles of a right triangle is a right angle, the triangle is oblique. To solve such triangles, there are four possibilities to consider: (1) one side and two angles are given; (2) two sides and the angle opposite one of them are given; (3) two sides and the included angle are given; and (4) three sides are given. In all of the following discussion, the sides are labeled a, b, and c; the angles opposite these sides are α, β, and γ respectively; and the corresponding vertices are A, B, and C (**Fig. 8**).

Law of sines. The law of sines, Eq. (25), is used

$$\frac{\sin \alpha}{a} = \frac{\sin \beta}{b} = \frac{\sin \gamma}{c} \tag{25}$$

to solve possibilities (1) and (2). To prove the law of sines, an altitude h is drawn from vertex B (Fig. 8). The attitude forms two right triangles yielding Eqs. (26). Solving for h gives part of Eq. (25).

$$\sin \alpha = h/c \quad \sin \gamma = h/a \tag{26}$$

By drawing the altitude h', from vertex A, the rest of Eq. (25) is obtained.

Law of cosines. The law of cosines, used to solve possibilities (3) or (4), may be stated by three equivalent formulas, Eqs. (27), (28), and (29). To prove Eq. (27),

$$c^2 = a^2 + b^2 - 2ab \cos \gamma \tag{27}$$

$$b^2 = a^2 + c^2 - 2ac \cos \beta \tag{28}$$

$$a^2 = b^2 + c^2 - 2bc \cos \alpha \tag{29}$$

a rectangular coordinate system is used in which vertex C lies at the origin and side b lies on the x axis (**Fig. 9**). The coordinates of vertex B are $(a \cos \gamma, a \sin \gamma)$. The distance formula is used to compute c^2, the square of the distance from $A = (b, 0)$ to B.

Infinite series representation. With the definition $n! = 1 \cdot 2 \cdot 3 \cdots n$, the sine and cosine functions may be represented by the two infinite series, given in Eqs. (30) and (31). These series converge for all x.

$$\sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \tag{30}$$

$$\cos x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \tag{31}$$

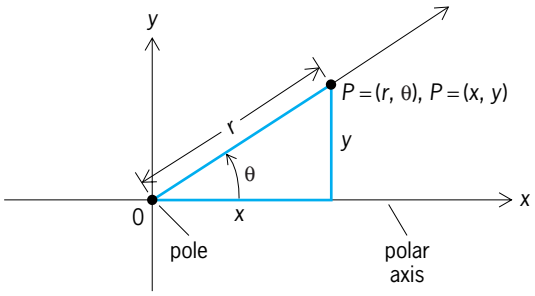


Fig. 10. Relationship between rectangular coordinates (*x*, *y*) and polar coordinates (*r*, θ).

Thus, to find a value of sin x or cos x, only as many terms of the series need to be used, as required to ensure required accuracy. The error that results will be less than the numerical value of the first term not used. For example, only the first two terms of the series in Eq. (30) are used to obtain $\sin 0.1 = 0.1 - (0.1)^3/6 = 0.099833$ (nearly). The first unused term, $(0.1)^5/120$ is less than 0.0000001, so the error in sin 0.1 due to only using the first two terms is less than 0.0000001. For five-decimal-place accuracy, only the first four terms of Eqs. (30) and (31) are required. *See* SERIES.

Polar coordinates. In a polar coordinate system, a point, called the pole, is selected, followed by a ray with vertex at the pole, called the polar axis. The pole coincides with the origin of a rectangular coordinate system, and the polar axis coincides with the positive x axis (**Fig. 10**). A point P may be represented by the ordered pair of numbers (r, θ), the polar coordinates of P. The relationship between rectangular coordinates (x, y) and polar coordinates (r, θ) is given by Eqs. (32). Many interesting curves are

$$x = r\cos\theta \quad y = r\sin\theta \tag{32}$$

defined by polar equations including limacons (with and without inner loops), lemniscates, roses, and spirals. *See* LEMNISCATE OF BERNOULLI; PLANE CURVE; ROSE CURVE; SPIRAL.

Complex numbers. A complex number is of the form $z = x + yi$, where i, the imaginary unit, is the number for which $i^2 = -1$. If (x, y) are the rectangular coordinates of a point, then the complex number $z = x + yi$ is in rectangular form. If (r, θ) are the polar coordinates of this point, then the polar form of z is given by Eq. (33). If $z_1 = r_1(\cos\theta_1 + i\sin\theta_1)$

$$\begin{aligned} z = x + yi &= r\cos\theta + ir\sin\theta i \\ &= r(\cos\theta + i\sin\theta) \end{aligned} \tag{33}$$

and $z_2 = r_2(\cos\theta_2 + i\sin\theta_2)$, then using Eqs. (33), (14), and (19), Eq. (34) follows.Using mathematical

$$z_1 z_2 = r_1 r_2[\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)] \tag{34}$$

induction, DeMoivre's theorem follows: If z is given by Eq. (33), then Eq. (35) is valid for $n \geq 1$, an integer.

$$z^n = r^n(\cos n\theta + i\sin n\theta) \tag{35}$$

Equation (35) also holds in n is a fraction, allowing the complex roots of a complex number to be found. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

Complex functions. The complex exponential function is defined by Eq. (36). If $z = i\theta$, then Eq. (37)

$$f(z) = e^z = e^x(\cos y + i\sin y) \tag{36}$$

$$e^{i\theta} = \cos\theta + i\sin\theta \tag{37}$$

holds. By using laws of exponents, DeMoivre's theorem, Eq. (38), follows. By substituting various values

$$e^{in\theta} = \cos n\theta + i\sin n\theta = (\cos\theta + i\sin\theta)^n \tag{38}$$

of n, this equation can be used to prove all the laws of trigonometry. *See* E (MATHEMATICS).

Equation (37) can be used to obtain a new trigonometry, in which the sine and cosine functions are defined by Eqs. (39). This theory has many applications in the theory of electricity. *See* ALTERNATING-CURRENT CIRCUIT THEORY.

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} \qquad \sin z = \frac{e^{iz} - e^{-iz}}{2i} \tag{39}$$

Michael Sullivan

Bibliography. J. D. Baley and G. Sarell, *Trigonometry*, 3d ed., 1996; R. V. Churchill and J. W. Brown, *Complex Variables and Applications*, 6th ed., 1996; J. H. Clough-Smith, *An Introduction to Spherical Trigonometry*, 1987; L. Drooyan, W. Hadel, and C. C. Carico, *Trigonometry: An Analytical Approach*, 6th ed., 1990; E. R. Heineman and J. D. Tarwater, *Plane Trigonometry*, 7th ed., 1993; K. J. Smith, *Trigonometry for College Students*, 7th ed., 1998; M. Sullivan, *Trigonometry*, 4th ed., 1996.

Trihedron

A geometric figure bounded by three noncoplanar rays called edges that emanate from a common point called the vertex, and by the plane sectors called faces that are formed by each pair of edges (**Fig. 1**). A trihedron has three dihedrons formed by pairs of face planes, and three face angles formed by pairs of edges. A plane intersecting the edge of a dihedron cuts it into two trihedrons, whose trihedral angles are measures whose sum is the dihedral angle of the

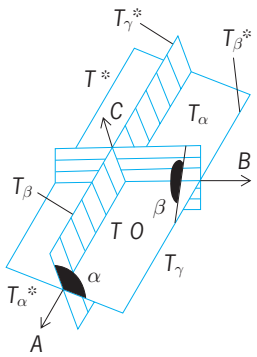


Fig. 1. Trihedron and trihedral angles.

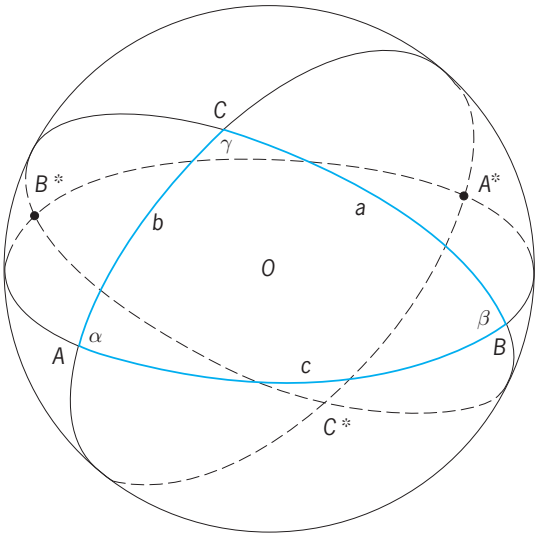


Fig. 2. Spherical triangle *ABC* formed by intersection of sphere with a trihedron. The center of the sphere is at the vertex of the trihedron.

dihedron. Three planes having a common point but not a common line cut space into eight associated trihedrons, of which opposite ones are congruent but not necessarily superposable. (One is the mirror image of the other.) If one of these eight trihedrons has dihedral angles α, β, and γ and trihedral angle σ, its three neighbors that each share one of its faces will have trihedral angles $\alpha - \sigma$, $\beta - \sigma$, and $\gamma - \sigma$. The sum of the four trihedral angles is 180°, so $2\sigma = \alpha + \beta + \gamma - 180^\circ$.

A sphere with its center at the vertex of a trihedron cuts the trihedron in a spherical triangle whose angles α, β, and γ are the dihedral angles of the trihedron and whose sides are measured by the face angles of the trihedron (**Fig. 2**). For a discussion of the relations between these angles and sides *see* TRIGONOMETRY

J. Sutherland Frame

Trilobita

A class of extinct Paleozoic arthropods, occurring in marine rocks of Early Cambrian through late Permian age. Their closest living relatives are the chelicerates, including spiders, mites, and horseshoe crabs (Xiphosura). About 3000 described genera make trilobites one of the most diverse and best-known fossil groups (**Fig. 1**). Species diversity peaked during the Late Cambrian and then declined more or less steadily until the Late Devonian mass extinction. Only four families survived to the Mississippian, and only one lasted until the group's Permian demise. Their dominance in most Cambrian marine settings is essential to biostratigraphic correlation of that system. *See* CAMBRIAN; CHELICERATA; DEVONIAN; PERMIAN.

Trilobites are typically represented in the fossil record by the mineralized portion of their exoskeleton, either as carcass or molt remains. The mineralized exoskeleton (**Fig. 2**) was confined mostly to the dorsal surface, curved under as a rimlike doublure (Figs. 1*b* and 2*b*); a single mineralized ventral plate, the hypostome, was suspended beneath the median region of the head (Fig. 2*b*). The mineralized exoskeleton was composed of low magnesian calcite and a minor component of organic material. Most of the ventral exoskeleton, including the appendages, was unmineralized.

Morphology of exoskeleton. The term "trilobite" refers to the longitudinal division of the body into an axial lobe and two lateral pleural regions (Fig. 2*c*); axial furrows separate the three divisions. The head shield, consisting of up to six fused segments and an anterior presegmental region, is called the cephalon. Its median (axial) lobe contains the glabella, typically convex and indented by a transverse occipital furrow and several pairs of lateral glabellar furrows (Fig. 2*a*). In primitive trilobites, segments of the palpebro-ocular (eye) lobes can be traced across the anterior region of the glabella (Fig. 1*a*).

Trilobites preserve the oldest known visual system in the history of life. Most had rigid compound eyes analogous to those of a housefly. The eyes are situated on the pleural field (genae, or cheeks). Most trilobites had a large number of small eye lenses that shared a single corneal covering (holochroal eye) [Fig. 1*e* and *f*.] The suborder Phacopina, a major Ordovician-through-Devonian group, had large separated lenses (schizochroal eye) [Fig. 1*i*].

In most trilobites, a facial suture, used for molting the exoskeleton, is developed on the dorsal side of the cephalon; it passes from the ventral side usually in front of the glabella, separates the visual surface of the eye from the palpebral lobe, and exits the cephalon in front of, through, or behind the genal angle. These different configurations of the suture are termed proparian (Fig. 1*e–h*), gonatoparian (Fig. 2*a*), and opisthoparian (Figs. 1*j* and 2*c*), respectively. The area between the axial furrow and facial suture is the fixigena (fixed cheek); together with the axial region of the cephalon (including the glabella), this single skeletal part (or sclerite) is the cranidium (Fig. 1*g*). The librigenae (free cheeks) represent the pleural areas outside the facial suture. Most trilobites had the anterior branches of the facial suture separated on the doublure by a rostral plate (Fig. 1*b* and *f*), although some had a median suture and others lost the ventral sutures and fused the doublure medially. The hypostome was rigidly sutured to the roof of the doublure in some groups (the conterminant condition; Fig. 2*b*), but in others it was free and supported by soft tissue (the natant condition).

Anterior wings on the lateral part of the hypostome bear processes that permitted its attachment to a stalk in the cephalic axial furrow (Fig. 2*b*). The thorax is composed of from 2 to more than 60 articulated segments (although typically 6 to 16), each consisting of an axial ring and pleural band. Articulation of the thorax, via processes and sockets on adjacent pleurae, allowed flexibility for enrollment (Fig. 1*i*). The pygidium is a posterior sclerite composed of one or more fused segments. Primitively, it is much smaller than the cephalon, but is enlarged in many

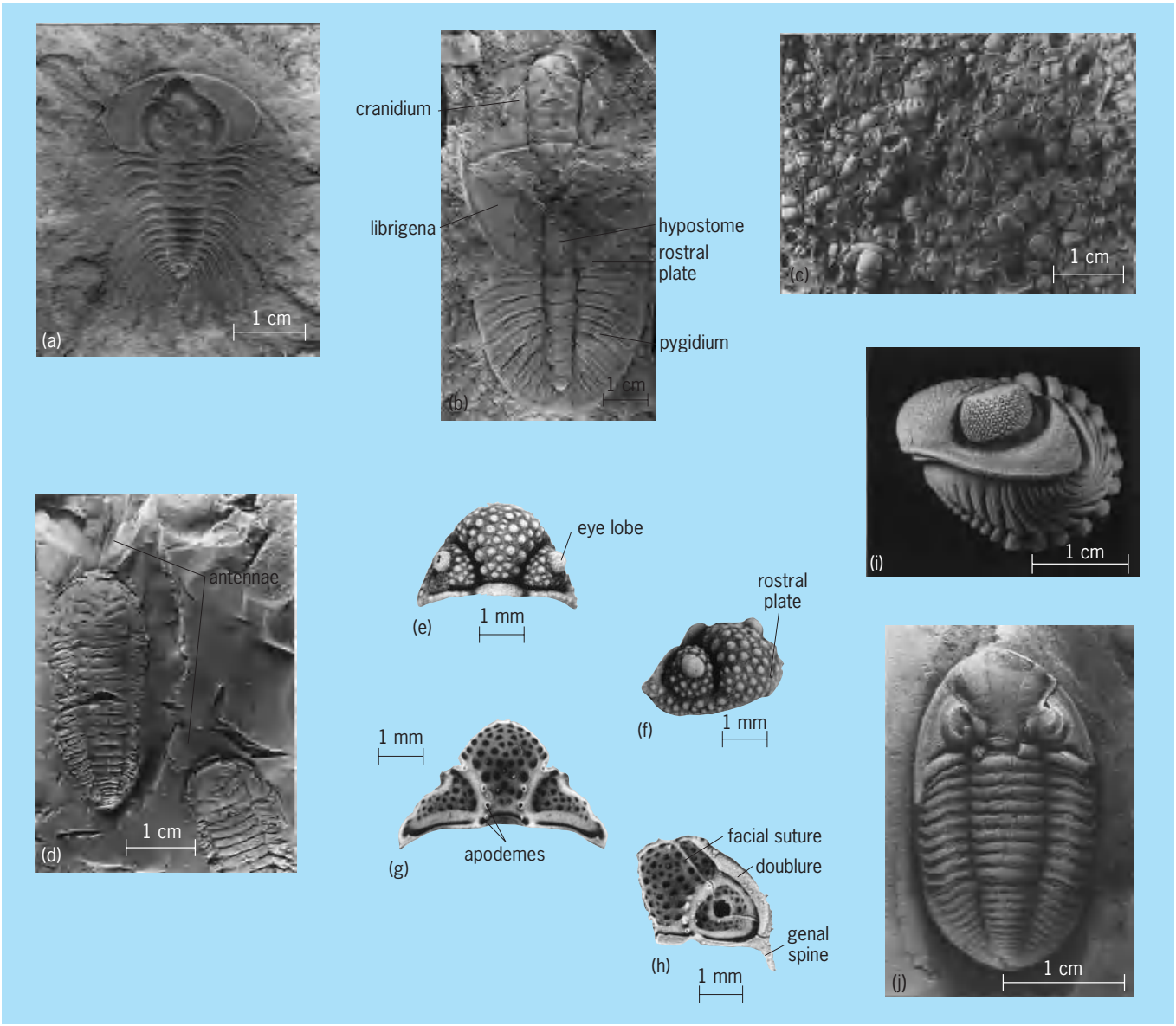


Fig. 1. Trilobite diversity and preservation. (*a*) Olenellid, Lower Cambrian (British Columbia). External mold of exoskeleton in shale. (*b*) *Ogygopsis*, Middle Cambrian (British Columbia). Molt assemblage, with cranidium and pygidium aligned, but librigenae, hypostome, and rostral plate inverted and rotated backward. Thorax is missing. Internal mold. (*c*) Slab of Upper Cambrian limestone (Sweden) with abundant disarticulated trilobite sclerites, mostly cranidia of *Olenus*. (*d*) *Triarthrus*, Upper Ordovician (New York). Exoskeletons replaced by pyrite, preserving antennae. (*e–h*) *Struszia*, Silurian (Northwest Territories), dorsal and anterolateral views of a cephalon, and ventral views of a cranidium and partial cranidium with attached librigena. Exoskeletons replaced by quartz; silicified fossils freed by dissolving limestone in acid. (*i*) *Phacops*, Devonian (Ohio). Enrolled exoskeleton, showing large lenses of schizochroal eye. (*j*) *Griffithides*, Mississippian (Indiana). Dorsal view of exoskeleton.

groups (Fig. 1*b*), and may bear spines along its margin. The cephalic doublure sometimes has notches or furrows that accommodated the pygidium and thoracic tips when the trilobite enrolled (Fig. 2*b*).

Appendages, preserved by pyrite or phosphate replacement or as films on shale, are well known for only a few trilobite species. A single pair of long, jointed antennae (Fig. 1*d*) projected forward from beneath the hypostome. Known Cambrian and Ordovician species have three pairs of postantennal cephalic appendages, while a Devonian example has four. In most cases, these show little structural differentiation from each other, or from postcephalic appendages on each segment along the length of the body (**Fig. 3**). The appendages are biramous, consisting of a jointed walking leg, or telopodite, and a filamentous exite, which attach toward the body axis to a spine-bearing coxa. Appendage-related musculature attached to the ventral exoskeleton at knoblike apodemes (Fig. 1*e* and *h*), just inward of the axial furrow. Enrollment and outstretching were achieved by flexor and extensor muscles; longitudinal, dorsoventral, and horizontal muscles have been observed, as well as a system of intersegmental bars. The exite (gill branch) functioned as a respiratory organ. The mouth opening was positioned above the rear margin of the hypostome and was directed posteriorly. The gut looped backward beneath the glabella, with

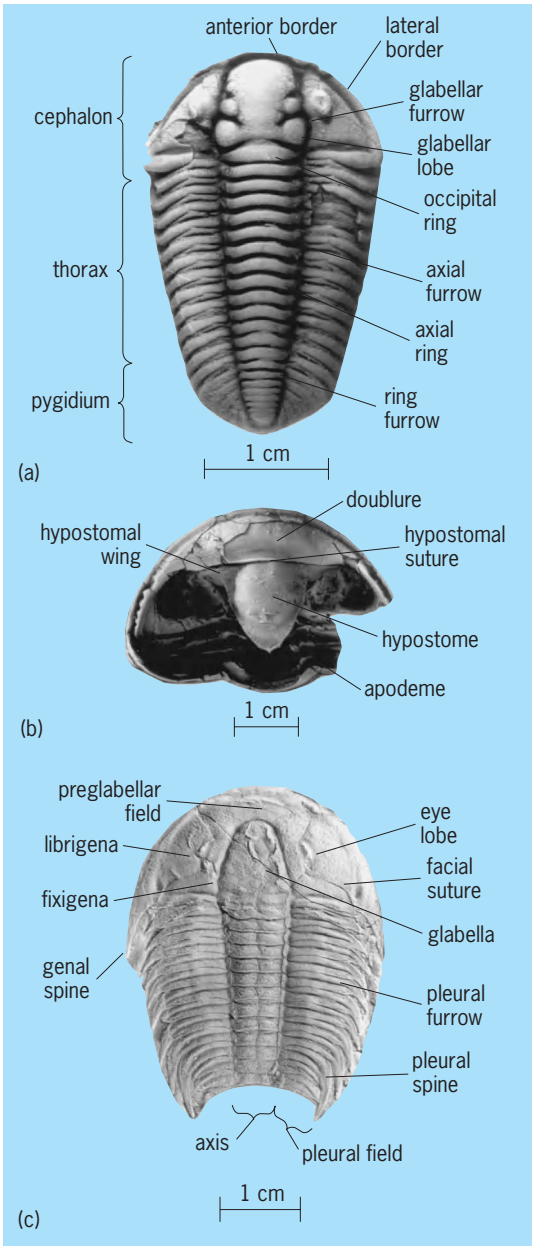


Fig. 2. Morphological features of Trilobita. (*a*) *Calymene*, Silurian (New York). Dorsal view of mineralized exoskeleton. (*b*) *Phacops*, Devonian (Ohio). Ventral (bottom) view of cephalon and anterior segments of thorax, with hypostome attached. (*c*) *Modocia*, Middle Cambrian (Utah). Dorsal view of cephalon and thorax preserved in shale; pygidium and last thoracic segment missing.

the digestive tract extending along the axis to a posterior anus.

Development and molting. Embryonic development of trilobites is unknown. Phosphatized arthropod eggs, which may be those of trilobites, have been discovered in Cambrian rocks. The term "protaspis" is applied to the earliest calcified larval stages, in which the cephalon and protopygidium are fused as an unjointed dorsal shield (**Fig. 4**). Several molts may occur within the protaspid period. The meraspid period is defined by articulation of the cephalon and transitory pygidium as separate sclerites; successive degrees are marked by the release of segments from the anterior part of the transitory pygidium to form the thorax. The holaspid has the complete adult complement of thoracic segments; development in this period is marked by continued increase in size and by changes in shape, but without further addition of segments to the thorax. Adult size ranges from 1.5 mm to 70 cm (0.06 to 28 in.); 2–5 cm (0.8–2 in.) is typical.

Trilobites show the typical arthropod solution to the problem of increasing size with a stiffened exoskeleton: they molted at regular intervals throughout the life cycle. In most species, this was effected by shedding the librigenae along the facial suture and shedding the hypostome. The soft-bodied animal emerged from the resulting gap. Several different molt strategies were employed by different trilobite groups, however, including shedding the entire cephalon, and inverting and rotating various skeletal elements (Fig. 1*b*). Molting results in the typical preservation of trilobite remains as disarticulated sclerites (Fig. 1*c*).

Ecology and macroevolution. Most trilobites were benthic deposit feeders or scavengers, living on the sediment-water interface or shallow-burrowing just beneath it. Some were evidently carnivores,

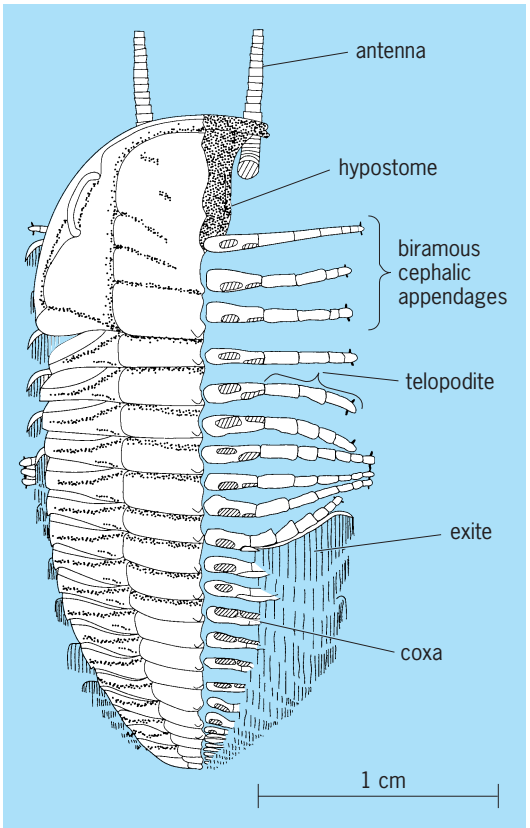


Fig. 3. *Triarthrus eatoni*, Upper Ordovician (New York). Reconstruction with dorsal exoskeleton removed on right side to show appendages. Antennae are incomplete (compare Fig. 1*d*). Exites of first nine postantennal appendages are removed to show structure of telopodite. The mouth was positioned above the posterior margin of the hypostome. (*After H. B. Whittington and J. E. Almond, Appendages and habits of the Upper Ordovician trilobite Triarthrus eatoni, Phil. Trans. Roy. Soc. Lond., B317:28, 1987*)

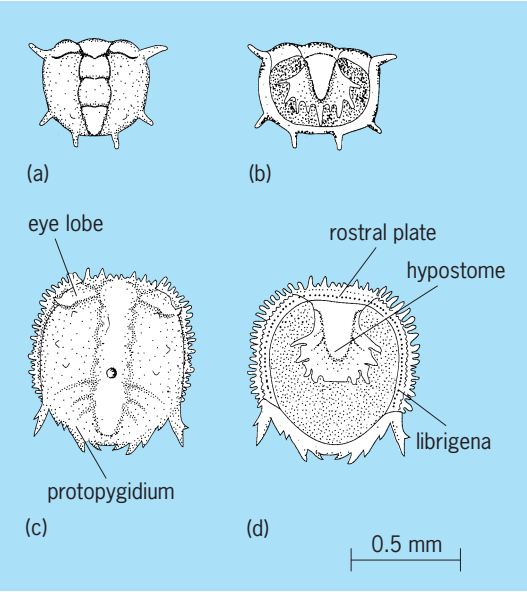


Fig. 4. ***Flexicalymene senaria*, Middle Ordovician (Virginia). Complete exoskeletons of protaspid larvae obtained from silicified residues. (*a*) Dorsal view and (*b*) ventral view of second of four protaspid instars for this species. (*c*) Dorsal view and (*d*) ventral view of fourth protaspid instar. Holaspides closely resemble the related genus *Calymene* (Fig. 2*a*). (*After B. D. E. Chatterton et al., Larvae and relationships of the Calymenina (Trilobita), J. Paleontol., 64:259, 1990*)**

equipped with sharp spines and processes projecting ventrally from their appendages. A few Cambrian and Ordovician groups acquired giant eyes coupled with narrow, streamlined bodies. The morphology and broad geographic and environmental ranges of these groups suggest they were active swimmers. Through their history, trilobites became adapted to all marine environments, from shallow high-energy shorefaces to deep-water, disaerobic habitats.

Trilobites are the most common marine fossils of the Cambrian Period, and their remains typically account for more than 90% of preserved Cambrian fossil assemblages. They were important through the Early Ordovician, but their numerical contribution to onshore communities was much reduced as a result of the Ordovician Radiation of marine life. This event saw filter-feeding organisms (for example, articulate brachiopods, bryozoans, crinoids) proliferate and rapidly evolve to dominate marine communities, a pattern that would last through the Paleozoic Era. Trilobites remained major components of deeper-water communities through the Silurian. Within-habitat, species diversity was generally constant in all environments from the Cambrian through the Silurian, despite their increasingly reduced relative abundance. This indicates that trilobites were largely unaffected by the major events of the Early Paleozoic and that their decline in importance was largely a function of increases in other groups.

Global trilobite diversity increased rapidly following the acquision of hard parts during the Cambrian Explosion, and it peaked during the Late Cambrian. Overall diversity gradually declined during the Ordovician, although a major subset of trilobite groups experienced an evolutionary burst during the Ordovician Radiation. The end-Ordovician mass extinction decimated the group, cutting their global diversity by about half. Surviving families were mainly those that had radiated during the Middle Ordovician. Global diversity continued to decline during the Silurian, although the most speciose trilobite faunas ever found occurred in this period. By the Devonian, trilobites were a relatively minor group, absent from many marine faunas, although still sometimes locally abundant. The Late Devonian mass extinction all but obliterated the trilobites, as only a handful of lineages survived. During the Late Paleozoic, trilobites were typically rare and confined to a limited number of facies. The last trilobites became extinct during the great end-Permian mass extinction. *See* CAMBRIAN; ORDOVICIAN; PALEOZOIC; SILURIAN.

Classification. Trilobita is usually assigned the ranking of class within Arthropoda. Affinities with Chelicerata are expressed by their grouping as Arachnata. Older classifications recognized a phylum or subphylum Trilobitomorpha, grouping Trilobita with an unnatural assortment of trilobite-, chelicerate-, or crustacean-like taxa lumped as Trilobitoidea. The soft-bodied Early-Middle Cambrian order Nectaspida is the closest relative (sister group) of the calcified Trilobita.

The high-level classification of trilobites remains controversial. Post-Cambrian groups (for example, orders Phacopida, Odontopleurida, Lichida, Proetida, Aulacopleurida) are well understood and are grouped into orders or suborders based on distinctive adult and larval morphologies. Cambrian trilobites are generally less well known (despite their abundance as fossils) and have tended to be classified in a small number of large unnatural orders such as Ptychopariida. A particular problem is a lack of understanding of the origins of post-Cambrian orders among Cambrian taxa, a phenomenon termed cryptogenesis. The result is that relationships between named orders of trilobites are essentially unknown. Recent progress has resulted from study of silicified life histories (Fig. 4), but inferring the high-level phylogeny of trilobites remains the cardinal problem in the paleobiology of the group.

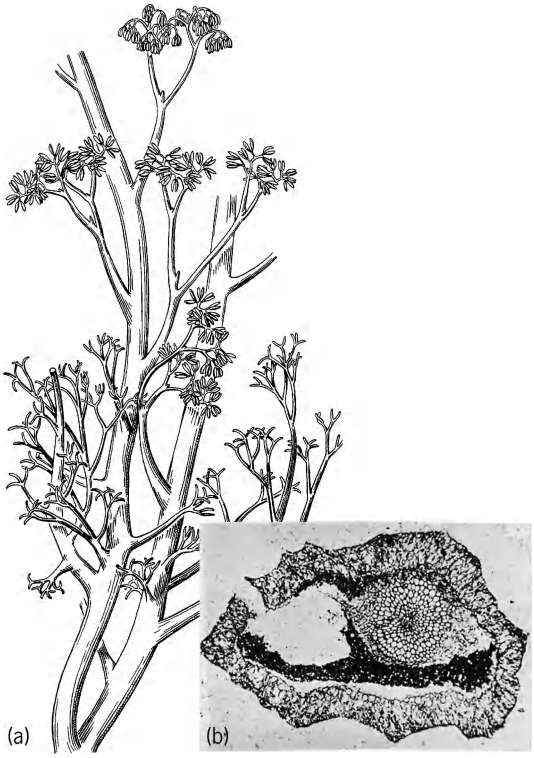
A group of blind marine arthropods, the Agnostida, has traditionally been recognized as an order of trilobites. Agnostids share a calcified dorsal exoskeleton with Trilobita, but otherwise lack most diagnostic trilobite features, including a calcified protaspid stage, facial sutures, articulating thoracic segments, and a true transitory pygidium. The appendages of agnostids are also fundamentally unlike those of trilobites. Their affinities are currently debated, with some workers defending their position as ingroup trilobites and others considering the agnostids to be stem group Crustacea. *See* ARTHROPODA; TAXONOMY. Gregory D. Edgecombe; Jonathan Adrain

Bibliography. R. A. Fortey, Ontogeny, hypostome attachment and trilobite classification, *Palaeontology*, 33:529–576, 1990; R. A. Fortey, *Trilobite!*

Eyewitness to Evolution, 2000; R. L. Kaesler (ed.), *Treatise on Invertebrate Paleontology*, pt. O (rev.), vol. 1, 1997; H. B. Whittington, *Trilobites*, 1990.

Trimerophytopsida

Mid-Early-Devonian into Mid-Devonian vascular plants at a higher evolutionary level than Rhyniopsida. Branching was profuse and varied, dichotomous, pseudomonopodial, helical to subopposite and almost whorled, and often trifurcate (see **illus.**).



***Psilophyton dawsonii*. (*a*) Reconstruction of known portion of a plant showing leafless stem, lateral branch systems terminating in large clusters of fusiform sporangia that dehisce longitudinally, and vegetative branching (*after H. P. Banks, H. P. S. Leclercq, and F. M. Hueber, Anatomy and morphology of Psilophyton dawsonii sp. nov., from the Late Lower Devonian of Quebec (Gaspé), and Ontario, Canada, Paleontol. Amer., 8:73–77, 1975*). (*b*) Photomicrograph of a transverse section of one stem. Outer cortex and xylem only are petrified. Centrally located, smaller cells indicate the maturation of xylem was centrarch.**

Vegetative branches were often in a tight helix, terminated by tiny recurved branchlets simulating leaf precursors. The axes were leafless and glabrous or spiny. Fertile branches were trifurcate or dichotomized several times in three planes and terminated in loose or tight clusters of 32–256 fusiform sporangia that dehisced longitudinally. The number of sporangia varied, in part, by abortions or failure of dichotomies. Spores were uniform in *Apiculi retusispora* or *Retusotriletes*, depending on preservation. Xylem is known only in *Psilophyton*, and was centrarch. Tracheids were helical, circular-bordered and scalariform-bordered pitted, and peculiarly multiaperturate. *Trimerophyton* and *Pertica* are two other natural genera. Form genera include some species of *Dawsonites*, *Hostinella*, and possibly *Psilophytites* and *Psilodendrion*. Progymnosperms, ferns, and articulates are derived from this group. *See* EMBRYOBIONTA; PSILOTOPHYTA; RHYNIOPHYTA; RHYNIOPSIDA.

Harlan P. Banks

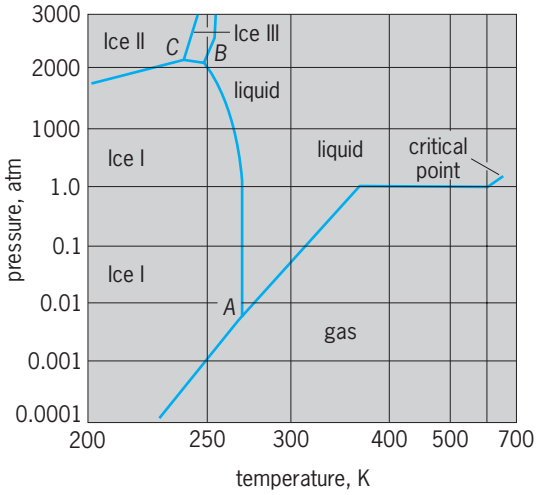
Triple point

A particular temperature and pressure at which three different phases of one substance can coexist in equilibrium. In common usage these three phases are normally solid, liquid, and gas, although triple points can also occur with two solid phases and one liquid phase, with two solid phases and one gas phase, or with three solid phases.

According to the Gibbs phase rule, a three-phase situation in a one-component system has no degrees of freedom (that is, it is invariant). Consequently, a triple point occurs at a unique temperature and pressure, because any change in either variable will result in the disappearance of at least one of the three phases. *See* PHASE EQUILIBRIUM.

Triple points are shown in the **illustration** of part of the phase diagram for water. Point *A* is the well-known triple point for Ice I (the ordinary low-pressure solid form) + liquid + water + water vapor at 0.01°C (273.16 K) and a pressure of 0.00603 atm (4.58 mmHg or 611 pascals). In 1954 the thermodynamic temperature scale (the absolute or Kelvin scale) was redefined by setting this triple-point temperature for water equal to exactly 273.16 K. Thus, the kelvin (K), the unit of thermodynamic temperature, is defined to be 1/273.16 of the thermodynamic temperature of this triple point.

Point *B*, at 251.1 K (−7.6°F) and 2047 atm (207.4 megapascals) pressure, is the triple point for liquid



Phase diagram for water, showing gas, liquid, and several solid (ice) phases; triple points at *A, B,* and C. The pressure scale changes at 1 atm from logarithmic scale at low pressure to linear at high pressure. 1 atm = 100 kPa; °F = (K × 1.8) − 459.67.

water + Ice I + Ice III; and point *C*, at 238.4 K (−31°F) and 2100 atm (212.8 MPa) pressure, is the triple point for Ice I + Ice II + Ice III. At least four other triple points are known at higher pressures, involving other crystalline forms of ice.

For most substances the solid-liquid-vapor triple point has a pressure less than 1 atm (about 100 kilopascals); such substances then have a liquid-vapor transition at 1 atm (normal boiling point). However, if this triple point has a pressure above 1 atm, the substance passes directly from solid to vapor at 1 atm. *See* SUBLIMATION.

For a two-component system, the invariant point in a phase diagram is a quadruple point at which four phases coexist. The three-phase situation is then represented by a line in the three-dimensional pressure-temperature-composition diagram. *See* BOILING POINT; ICE POINT; MELTING POINT; TRANSITION POINT; VAPOR PRESSURE; WATER.

Robert L. Scott

Triplet state

A molecule exists in this electronic state when its total spin angular momentum quantum number S is equal to one. The triplet state is an important intermediate of organic chemistry. In addition to the wide range of triplet molecules available through photochemical excitation techniques, numerous molecules exist in stable triplet ground states, for example, oxygen molecules. Theoretical calculations, furthermore, make predictions concerning the spin multiplicities of the ground states of many prototype organic molecules such as cyclobutadiene, trimethylene methane, and methylene, and indicate that they will be triplets. *See* ATOMIC STRUCTURE AND SPECTRA; REACTIVE INTERMEDIATES; SPIN (QUANTUM MECHANICS).

Practical definition. A good working definition of a triplet state for the chemist is the following: A triplet is a paramagnetic even-electron species which possesses three distinct but energetically similar electronic states as a result of the magnetic interaction of two unpaired electron spins. The several important terms of this definition allow some insight as to the essential features of a triplet. First of all, a triplet is paramagnetic, and should thus display this property in a magnetic field. This paramagnetism serves as the basis for experimental magnetic susceptibility and electron spin resonance studies of the triplet state. However, one can imagine many paramagnetic odd-electron species which are not triplets, for example, nitric oxide. Thus, the criterion that a triplet must also be an even-electron species is apparent.

However, one can imagine paramagnetic, even-electron species which possess (1) only two distinct electronic states or (2) five or more electronic states. The former occurs when the paramagnetism results from two electrons which act as two independent odd electrons. For example, two carbon radicals separated by a long saturated chain will behave as two

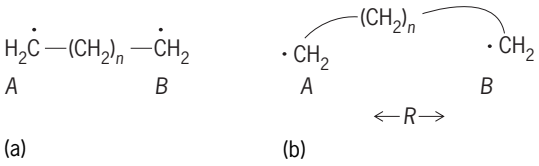


Fig. 1. Two carbon radicals connected by a long methylene chain. (*a*) Biradical state. (*b*) Triplet state.

doublet states if there is sufficient separation to prevent spin interactions. Five or more electronic states result when four or six parallel electronic spins interact (to yield quintet and septet states, respectively). *See* ELECTRON SPIN; PARAMAGNETISM.

One can now see that conceptual difficulties may arise in differentiating a biradical state (that is, a species possessing two independent odd-electron sites) from a triplet. Suppose two carbon radicals are separated by a long methylene chain as in **Fig. 1***a*. If the methylene chain is sufficiently long and the odd-electron centers are so far removed from one another that they do not interact (magnetically and electronically) with one another, then the system is a doublet of doublets, that is, two independent odd electrons or a true biradical. If the methylene chain should be folded (Fig. 1*b*) so that the odd electrons begin to interact (magnetically and electronically) with one another, then at some distance R between the —CH_2 groups the doublet or doublets will become a triplet state. This state will result from the fact that the spin of the electron on carbon A is no longer independent of the spin on carbon B. Since the spins are quantized, selection rule (1) applies, where S is

$$\text{Number of spin states} = 2|S| + 1 \qquad (1)$$

the sum of the spin quantum numbers for the two electrons. This means that either three spin states (if $S = 1$ or -1, that is, spins of both electrons on C_A and C_B are the same) or one spin state (if $S = 0$, that is, spin of the electron of C_A is paired with that of C_B) will result. The former describes a triplet state, and the latter a singlet.

This leads to a difficulty in terminology: The "triplet state" is not one state but three states even in the absence of an external magnetic field. Indeed, under favorable conditions transitions may be observed between triplet levels at zero external magnetic field. The effect of an external magnetic field is to further split the triplet levels and allow transitions between them to be more easily detected.

Properties. A triplet may result whenever a molecule possesses two electrons which are both orbitally unpaired and spin unpaired. As shown in **Fig. 2**, orbital unpairing of electrons results when a molecule absorbs a photon of visible or ultraviolet light. Direct formation of a triplet as a result of this photon absorption is a very improbable process since both the orbit and spin of the electron would have to change simultaneously. Thus, a singlet state is generally formed by absorption of light. However, quite often the lifetime of this singlet state is

sufficiently long to allow the spin of one of the two electrons to invert, thereby producing a triplet. The following discussion considers the ways in which such a species is unambiguously characteristic. *See* MOLECULAR ORBITAL THEORY.

The question to be answered is: What are the general properties to be expected of a molecule in the triplet state? Some of the more important physical properties are (1) paramagnetism; (2) absorption between triplet sublevels; (3) electronic absorption from the lowest triplet to upper triplets; (4) electronic emission from the lowest triplet to a lower singlet ground state (if the triplet level is not the ground state). The paramagnetism of the triplet results from the interaction of unpaired spins and the fact that an unpaired spin shows a paramagnetic effect (is attracted) in a magnetic field.

Absorption between triplet sublevels may be observed directly by the use of an electron spin resonance spectrometer. *See* ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY.

The triplet, like any other electronic state, may be excited to upper electronic states of the same spin as the result of light absorption. In favorable cases this may be observed by the method of flash spectroscopy. *See* PHOTOCHEMISTRY.

For most organic molecules the lowest triplet state is an excited electronic state and may emit light and pass to the ground singlet state. Since light absorption to form a triplet from a singlet is improbable, the symmetrically related emission of light from a triplet returning to a ground state is likewise improbable. Indeed, it takes the triplet states of some aromatic molecules an average of about 30 s to emit light. This phenomenon is known as phosphorescence and is to be contrasted with fluorescence, the emission of light from an excited singlet state returning to a singlet ground state, a process which often occurs in nanoseconds. *See* FLUORESCENCE; PHOSPHORESCENCE.

Although phosphorescence (long-lived emission) was the first method employed to study triplets, it is not a specific device for establishing whether a long-lived emission occurs from a triplet. For instance, examples are known for which the slow combination of positive and negative sites will generate excited molecules which emit light. In this case the combination reaction may be rate-determining for light emission.

Similarly, absorption from one triplet to another is not a specific method since the precise triplet-triplet absorption characteristics cannot be predicted accurately. It would thus remain to be proven that the absorbing species is indeed a triplet and not some other transient species.

Even paramagnetism is not an infallible probe for a triplet state since free radicals which are also paramagnetic are often produced by the absorption of light.

It appears that electron spin resonance (ESR) is probably the most powerful single method for establishing that a molecule is in its triplet state. The nature of the ESR signals may be predicted and fitted to theoretical relation (2), which describes the

$$H = g_0 H \cdot S + DS_z^2 + E(S_x^2 - S_y^2) \qquad (2)$$

magnetic spin interactions and expected absorptions. Here g_0 is the Landé g factor, D the dielectric constant, and E the electric field strength of the molecule. This particular equation is derived for the special case of molecules with a plane of symmetry and a symmetry axis perpendicular to that plane. However, the important general features of this equation are (1) the term $g_0 H \cdot S$ which describes the interaction of the external magnetic field H with the unpaired electron spin S; (2) the term $DS^2{}_z + E(S^2{}_x - S^2{}_y)$ which describes the spin-spin dipolar interactions along the x, y, and z axes of the molecule. These interactions are indicated in **Fig. 3**. Thus, from a study of the behavior of a triplet in a magnetic field, information on the electronic distribution in this excited state is obtained. In favorable cases, the nuclear geometry of the triplet may be derived.

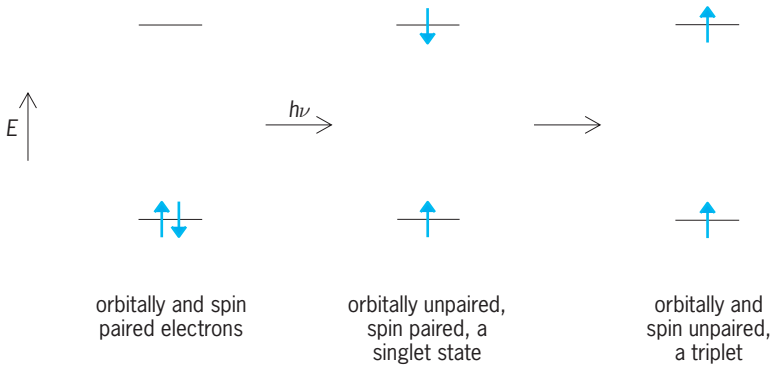


Fig. 2. Simple molecular orbital description of singlets and triplets.

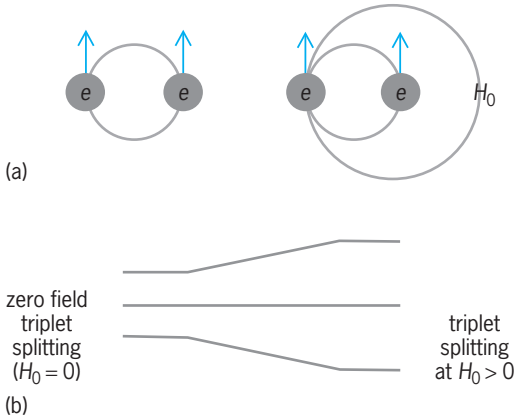


Fig. 3. Triplet state resulting from (*a*) internal spin-spin dipolar interaction ($D + E$) and (*b*) external interaction of electronic spin with magnetic field H_0.

Nicholas J. Turro

Bibliography. A. Devaquet et al., *Triplet States One*, 1975; *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., vol. 8, 1993; N. J. Turro, *Modern Molecular Photochemistry*, 1981; P. J. Wagner et al.,

Triplet States: No. 3, 1976; U. P. Wild et al., *Triplet States Two*, 1975.

Tripylida

An order of nematodes in which the cephalic cuticle is simple and not duplicated; there is no helmet. The body cuticle is smooth or sometimes superficially annulated. Cepahalic sensilla follow the typical pattern in which one whorl is circumoral and the second whorl is often the combination of circlets two and three. The pouchlike amphids have apertures that are inconspicuous or transversally oval. The stoma is variable, being simple, collapsed, funnel shaped or cylindrical, and armed or unarmed. In most taxa the stoma is surrounded by esophageal tissue; that is, it is entirely esophastome. When the stoma is expanded, both the cheilostome and esophastome are evident. Esophagi are cylindrical-conoid. Esophageal glands open anterior to the nerve ring. Males generally have three supplementary organs, more in some taxa. A gubernaculum accompanies the spicules. Caudal glands are generally present.

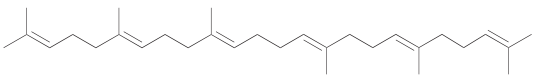
There are two tripylid superfamilies, Tripyloidea and Ironoidea. The characteristically well-developed cuticular annulation of the Tripyloidea is only rarely seen in other Enoplida. These nematodes are commonly found in fresh water or very moist soils; however, some are found in brackish water and marine habitats. Intestinal contents indicate that their food consists primarily of small microfauna that often include nematodes and rotifers. The Ironoidea contains species (presumably carnivorous) occurring in both fresh-water and soil habitats. *See* NEMATA (NEMATODA).

Armand R. Maggenti

Triterpene

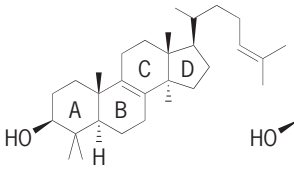
A hydrocarbon or its oxygenated analog containing 30 carbon atoms and composed of six isoprene units. Triterpenes form the largest group of terpenoids, but are classified into only a few major categories. Resins and saps contain triterpenes in the free state as well as in the form of esters and glycosides.

Biogenetically triterpenes arise by the cyclization of squalene (**1**) and subsequent skeletal rearrange-

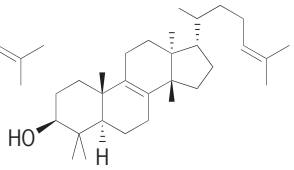


(**1**)

ments. Squalene can cyclize in five ways, leading to different stereochemical arrangements in the final triterpenoid structure. The conformation of a triterpenoid nucleus is determined in the initial folding of squalene into several chair or boat configurations. Apart from the linear squalene itself and some bicyclic, highly substituted skeletons, most triterpenes are either tetracyclic or pentacyclic compounds. The various structural classes are designated by the names of representative members. The lanosterol and euphol series comprise tetracyclic structures differing only in the stereochemical arrangements around the D ring. Lanesterol (**2**) occurs in sheep wool, and euphol (**3**) is obtained from Euphorbium

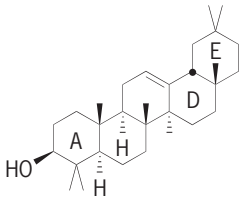


(**2**)

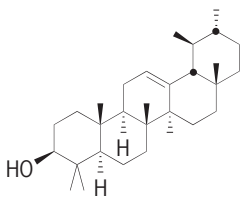


(**3**)

resin. The oleanane (β-amyrin, **4**, from grape seeds) and urasane (α-amyrin **5**, from Manila elemi resin)

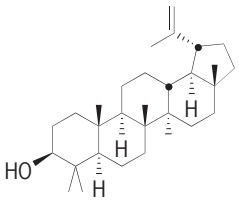


(**4**)

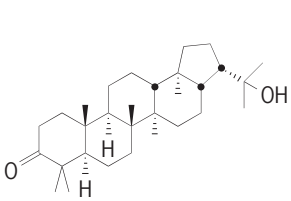


(**5**)

series of triterpenes are all pentacycles differing in the substitution pattern of methyl groups on ring E. Lupeol (**6**) from lupin seeds and hydroxyhopanone (**7**) from dammar resin are typical of



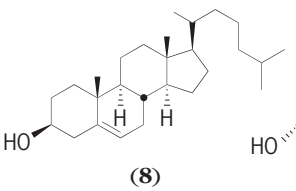
(**6**)



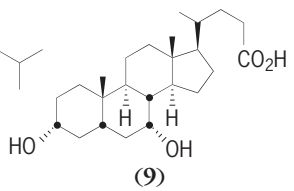
(**7**)

the lupane and hopane series. Hydroxyhopanone is one of the few triterpenoids that results from the cyclization of squalene without subsequent rearrangements.

Steroids and sterols are related to triterpenes. These important compounds are classed as nortriterpenes to indicate that they lack some of the 30 carbons of the triterpene skeleton. They have 27 to 29 carbons, and in most cases lack the geminal dimethyl group in ring A and one of the angular methyl groups in ring D of the lanosterol tetracyclic skeleton. Steroids are believed to arise from squalene via lanosterol, followed by the oxidative loss of methyl groups. Because of their significance in mammalian metabolism, steroids are usually treated as a separate class although their terpenoid origin is well understood. Cholesterol (**8**) and cholic acid (**9**)



(**8**)



(**9**)

are representative examples of steroid structures. *See* SQUALENE; STEROID; TERPENE.

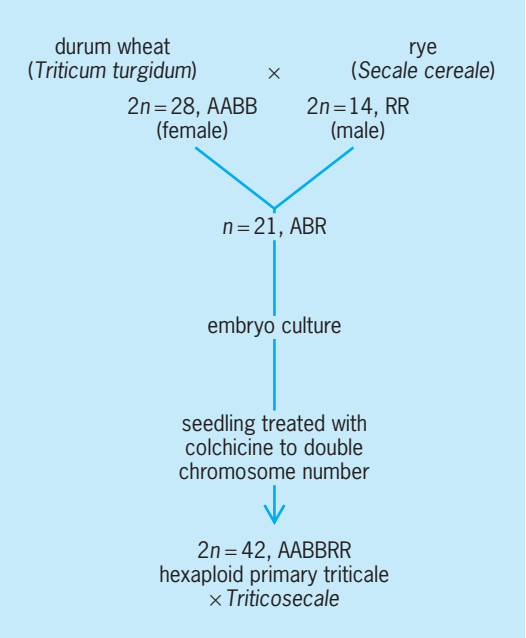
Tomas Hudlicky

Triticale

A cereal grass plant (× *Triticosecale*) obtained from hybridization of wheat (*Triticum*) with rye (*Secale cereale*). It is a crop plant with a small-seeded cereal grain that is used for human food and livestock feed. Worldwide, triticale is slowly gaining importance as a cereal grain. In 1998 it was estimated that about 2.9 million hectares (7.2 million acres) were planted each year. The European continent dominates triticale production with 70% of the total area. Countries with the greatest production (more than 100,000 ha, or 247,000 acres) are Poland, Russia, Germany, France, Australia, Brazil, and the United States.

Origin and types. Triticale was first developed in 1876, but not until the 1960s were types developed that were suitable for cultivation. Modern varieties are called secondary triticales because they were selected after interbreeding of various triticales, including primary types. In some triticale varieties, one or more rye chromosomes have been replaced by wheat chromosomes, giving secondary-substituted triticales, as contrasted to complete triticale having all seven rye chromosomes.

Triticale is produced by deliberate hybridization of either bread wheat [*Triticum aestivum*; diploid number of chromosomes ($2n$) = 42] or durum wheat (*T. turgidum* var. *durum*; $2n = 28$) with rye ($2n = 14$), followed by the doubling of the chromosome number of the hybrid plant (see **illus.**). The drug colchicine is commonly applied at a very low concentration to the seedling hybrid plants to cause the chromosome number to double. Hexaploid triticale (durum wheat × rye; $2n = 42$) is a more successful crop plant than octoploid triticale. The octoploid form ($2n = 56$) is produced by hybridization of bread wheat ($2n = 42$) with rye ($2n = 14$). A third type ($2n = 28$) from hybridizing diploid wheat (*T. monococcum*; $2n = 14$) with rye is of no economic consequence.



Flow diagram for hexaploid triticale development showing chromosome numbers and genome identifications. The octoploid form ($2n = 56$) is produced in a similar way, but starting with bread wheat (*Triticum aestivum*, $2n = 42$, AABBDD) as the female parent.

Adaptation. Triticale is grown from seeds sown in soil by using cultivation practices similar to those of wheat or rye. Both winter-hardy and nonhardy types exist, the latter used where winters are mild or for spring sowing. Triticale tends to have a greater ability than wheat to grow in adverse environments, such as saline or acid soils or under droughty conditions. While some triticales are resistant to attack by disease pathogens, there is great variation in resistance, just as in wheat. Ergot (caused by the fungus *Claviceps purpurea*) can infect triticale in some localities, probably to a higher degree than wheat. *See* ERGOT AND ERGOTISM.

Usage. Being a cereal grain, triticale can be used in food products made from wheat flour. Varieties tend to have large, somewhat irregularly shaped grains that produce a lower yield of milled flour than wheat. Bread and pastry products can be made very well with triticale flour. Experimental trials have shown triticale grain also to be acceptable as a starch source for beermaking.

As a livestock feed, triticale grain is a good source of carbohydrate and protein. Its protein is richer than wheat protein in lysine, one of the amino acids, making it especially desirable for the diets of monogastric animals, such as swine and birds, since these animals cannot synthesize lysine and so must obtain it in their diet. Triticale is also becoming more widely used as a fodder (forage) source for livestock feeding. Whole plants are harvested while still green, before the grain is mature. The fodder is fed mainly to cattle in one of three ways: immediately after harvest (green chop), after fermentation as silage, or after drying as hay. Triticale is also a pasture crop; sheep and cattle are allowed to graze on triticale during the early growth phase of the crop. However, triticale, or any other types of grain, infested with ergot bodies should not be eaten by humans or livestock.

Intense breeding and selection have made very rapid genetic improvements in triticale seed quality. Some triticale varieties now have seeds with hardness and shape similar to wheat. The gluten proteins of wheat are being incorporated into triticale so that triticale flour may be substituted for wheat in certain baked products. The agronomic advantages and improved end-use properties of the grain of triticale over wheat make triticale an attractive option for increasing global food production. *See* BREEDING (PLANT); RYE; WHEAT.

Calvin O. Qualset

Bibliography. M. Bernard and S. Bernard (eds.), *Genetics and Breeding of Triticale*, Institut National de la Recherche Agronomique, Paris, 1985; H. Guedes-Pinto, N. Darvey, and V. P. Carnide (eds.), *Triticale: Today and Tomorrow*, Kluwer Academic Publishers, Dordrecht, 1996; P. Juskiw (ed.), *4th International Triticale Symposium*, International Triticale Association, Red Deer, Alberta, Canada, 1998; A. Müntzing,

Advances in Plant Breeding, no. 10, supplement to *J. Plant Breed*. (Berlin), 1979; National Research Council, Board on Science and Technology for International Development, *Triticale: A Promising Addition to the World's Cereal Grains*, 1989.

Tritium

The heaviest isotope of the element hydrogen and the only one which is radioactive. Tritium occurs in very small amounts in nature but is generally prepared artificially by processes known as nuclear transmutations. It is widely used as a tracer in chemical and biological research and is a component of the so-called thermonuclear or hydrogen bomb. It is commonly represented by the symbol ^3_1H, indicating that it has an atomic number of 1 and an atomic mass of 3, or by the special symbol T. For information about the other hydrogen isotopes. *See* DEUTERIUM; HYDROGEN.

Properties. Both molecular tritium, T_2, and its counterpart hydrogen, H_2, are gases under ordinary conditions. Because of the great difference in mass, many of the properties of tritium differ substantially from those of ordinary hydrogen.

Chemically, tritium behaves quite similarly to hydrogen. However, because of its larger mass, many of its reactions take place more slowly than do those of hydrogen. The ratio of reaction rates may be as large as 64:1. These differences in reactivity can give rise to serious errors of interpretation when tritium is used as a tracer for hydrogen.

The nucleus of the tritium atom, often called a triton and symbolized *t*, consists of a proton and two neutrons. It has a mass of 3.01700 atomic mass units (amu), a nuclear spin of 1/2, and a magnetic moment of 2.9788 nuclear magnetons. It undergoes radioactive decay by emission of a beta particle to leave a helium nucleus of mass 3. No gamma rays are emitted in this process. The half-life for the decay is 12.26 years. The most energetic of the beta particles emitted by tritium have the comparatively low energy of 18.6 keV; beta particles are completely stopped by 7 mm of air or by 0.01 mm of paper or similar material. The average energy of the beta particles is 5.69 keV.

When tritium is bombarded with deuterons of sufficient energy, a nuclear reaction known as fusion occurs and energy considerably greater than that of the bombarding particle is released. The reaction may be written as (1). This reaction is one of those which

$$^3_1H + ^2_1H \rightarrow ^4_2He + ^1_0n + 18\ MeV \qquad (1)$$

supply the energy of the thermonuclear bomb. It is also of major importance in the development of controlled thermonuclear reactors. Enormous quantities of tritium will be required if such reactors are perfected and brought into use as electric power generators.

Compounds. Very few compounds of pure tritium have been prepared and studied. Such compounds would undergo decomposition quite rapidly under the action of the tritium beta radiation. Tritium oxide, T_2O, has been prepared by oxidation of tritium gas with hot copper oxide or by passing an electric spark through a mixture of tritium and oxygen. Its melting point is 4.49°C (40.08°F), compared with 0°C (32°F) for ordinary water. Of much greater importance are compounds, especially organic compounds, in which a small fraction of the hydrogen atoms have been replaced by tritium. Such labeled compounds are employed in tracer studies, such as those indicated above. Tritium-labeled compounds may be prepared by ordinary synthetic chemical methods, such as the catalytic addition of tritium-hydrogen mixtures to unsaturated compounds. Tritium may be exchanged for hydrogen in the presence of a catalyst such as platinum or a strong acid. In recoil labeling, a mixture of an organic compound and a lithium salt are irradiated with neutrons in a nuclear reactor; some of the energetic tritons produced are incorporated into the organic compound.

Another important labeling procedure consists of the exposure of an organic compound to tritium gas in a sealed vessel; the tritium beta radiation facilitates the exchange of hydrogen in the compound with tritium in the gas. Some compounds of biological interest have been prepared by growing organisms in tritiated water.

Analysis. Because of its weak beta radiation, tritium is not readily measured by the ordinary Geiger-Müller counter. More efficacious is the introduction of tritium as a gas inside the counting tube. Alternatively, the ionization of a gas caused by the beta radiation may be measured in an ionization chamber, or the tritium compound may be dissolved in a suitable solvent containing a phosphor and the light pulses excited by the beta particles then may be counted with a scintillation counter. Tritium gas containing only small amounts of ordinary hydrogen may be analyzed with a mass spectrometer or by measuring the density of the gas. Because of the very short range of the tritium beta particle, autoradiography, the exposure of radioactive material to a photographic plate, is often used to locate precisely the position of tritium in biological material.

Natural occurrence. Before the start of thermonuclear weapons testing in 1954, rainwater contained approximately 1–10 atoms of tritium per 10^{18} atoms of hydrogen. Such tritium originates largely from the bombardment of nitrogen in the upper atmosphere with neutrons and protons from cosmic rays, as in reaction (2). Because the half-life of tritium is short

$$^{14}_7N + ^1_0n \rightarrow ^3_1H + ^{12}_6C \qquad (2)$$

in comparison with the time required for mixing of the ocean waters, the concentration of tritium in the ocean is much lower than in rainwater. Before 1954 the total amount of tritium on the Earth's surface was estimated at 1800 g (63 oz), of which about 11 g (0.39 oz) was in the atmosphere and 13 g (0.46 oz) in ground waters. Testing of thermonuclear weapons resulted in sharp rises in the tritium content of

rainwater to values as high as 500 atoms per 10^{18} atoms of hydrogen.

Preparation. Tritium was first produced in the laboratory by bombarding compounds of deuterium with high-energy deuterons, as in reaction (3). A

$$^{2}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} \rightarrow {}^{3}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H} \quad (3)$$

number of other nuclear reactions also give rise to tritium. The most important of these is the absorption of slow neutrons by the lithium isotope of mass 6, according to reaction (4). By irradiating enriched

$$^{6}_{3}\mathrm{Li} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{3}_{1}\mathrm{H} + {}^{4}_{2}\mathrm{He} \quad (4)$$

lithium-6, in the form of an alloy with magnesium or aluminum, with neutrons from a nuclear reactor, tritium may be prepared on a large scale.

Uses. As a result of its production for use in nuclear weapons, tritium became available in large quantities at very low cost. It is used in admixture with zinc sulfide in the production of luminous paints, which have largely replaced the radium formerly used on watch dials; such mixtures are also used to produce small, permanent light sources. Tritium adsorbed on metals is used in targets for the production of fast neutrons by bombardment with deuterons. Tritium has been much used in hydrological studies, since it is an ideal tracer for water movement. Some studies depend on natural tritium or that introduced by weapons testing; in order cases large amounts of tritium are deliberately added. Investigations include the distribution of groundwater in oil fields; the tracing of springs, rivers, and lakes; water seepage and loss from reservoirs; and the movement of glaciers.

Tritium has also been used as a tracer for hydrogen in the study of chemical reactions. The most widespread use of tritium has probably been in biological research, where it has been used both as a hydrogen tracer and as a molecular label in studies of metabolism, biosynthesis, and cytology. In particular, tritiated thymidine and other nucleotides and nucleosides have been extensively used in studies of the formation of DNA and RNA. *See* HEAVY WATER; NUCLEAR FUSION; RADIOACTIVE TRACER; RADIOCHEMISTRY; TRITON.

Louis Kaplan

Bibliography. E. Buncel, *Tritium in Organic Chemistry*, vol. 4, 1978; F. Mannone, *Safety in Tritium Handling Technology*, 1993; G. Vasaru, *Tritium Isotope Separation*, 1993.

Triton

The nucleus of $_{1}\mathrm{H}^{3}$ (tritium); it is the only known radioactive nuclide belonging to hydrogen. The triton is produced in nuclear reactors by neutron absorption in deuterium ($_{1}\mathrm{H}^{2} + {}_{0}n^{1} \rightarrow {}_{1}\mathrm{H}^{3} + \gamma$), and decays by β^{-} emission to $_{2}\mathrm{He}^{3}$ with a half-life of 12.4 years. The spin of the triton is $^{1}/_{2}$, its magnetic moment is 2.9788 nuclear magnetons, and its mass is 3.01700 atomic mass units. Much of the interest in producing $_{1}\mathrm{H}^{3}$ arises from the fact that the fusion reaction $_{1}\mathrm{H}^{3} + {}_{1}\mathrm{H}^{1} \rightarrow {}_{2}\mathrm{He}^{4}$ releases about 20 MeV of energy. Tritons are also used as projectiles in nuclear bombardment experiments. *See* NUCLEAR REACTION; TRITIUM.

Henry E. Duckworth

Trochodendrales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the Eudicotyledon. The order consists of two families, the Trochodendraceae and Tetracentraceae, each with only a single species. The two species are often united into the single family Trochodendraceae. The group is of considerable botanical and evolutionary interest, as it is situated near the base of the advanced Eudicotyledon and links this larger group with more primitive flowering plants. Trochodendrales comprise trees of eastern and southeastern Asia with primitive (without vessels) wood. The flowers have a much reduced perianth with scarcely sealed carpels that are only slightly fused to each other. *See* PLANT KINGDOM; PLANT TAXONOMY.

K. J. Sytsma

Trogoniformes

A small order of birds that contains only the family Trogonidae. Thirty-seven species are found throughout the tropics; two species reach the southern border of the United States. The trogons and quetzals are jay-sized birds with large heads, and tails that vary from medium length and squared to very elongated and tapered. The dorsal plumage of trogons and quetzals is predominantly metallic green, with blue, violet, red, black, or gray in a few. The ventral feathers are bright red, yellow, or orange. Despite their vivid coloration, the birds are inconspicuous when sitting quietly in the forest. Quetzals possess long tapered tails of upper covert feathers, not tail feathers. Trogon plumage is soft with lax feathers. Sexes are dissimilar in appearance, with the males being more brightly colored. The head is large and rounded, and the bill is small and weak. Legs are short and feet are weak, with the toes arranged in a heterodactyl fashion, with the first and second toes reverted, opposing the third and fourth toes. Flight is rapid, undulating, and brief; trogons rarely walk. The diet consists of fruit and small invertebrates, as well as insects caught in flight as the bird darts out from a perch.

Trogons are nonmigratory, arboreal, and sedentary, and they can remain on a perch for hours. The monogamous pairs nest in solitude in a hollow tree or termite nest. After the eggs have been incubated by both parents, the naked hatchlings remain in the nest and are cared for by both parents.

A few fossil trogons are known from the Oligocene and Miocene of Europe. The Eocene to Oligocene fossils have been assigned to the family Archaeotrogonidae (*Archaeotrogon*) of enigmatic affinities. They may be ancestral to the trogons, but they may also belong to a totally different group of birds.

The relationship of the trogons to other birds is still unresolved. Some authors argue for a close relationship to the Coraciiformes, but the evidence is inconclusive. Also difficult to explain is their current pantropical distribution. They may have dispersed over the Eocene land connection across the North Atlantic between Europe and eastern North America during a warmer climatic period. *See* CORACIIFORMES.

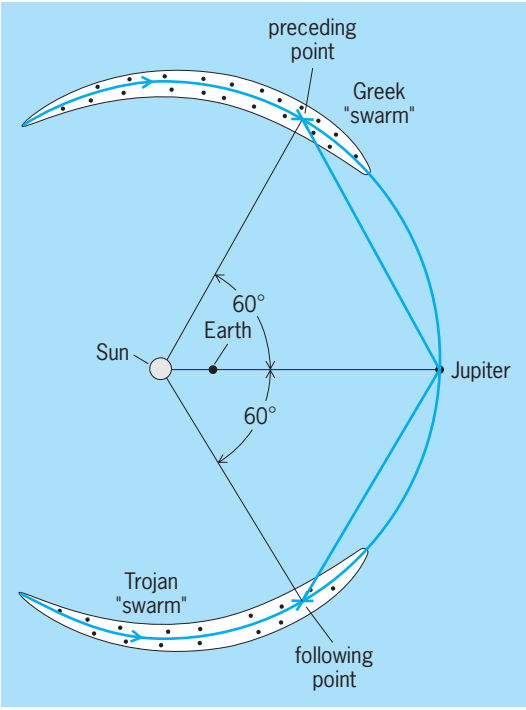
The quetzals are the most spectacular members of this brightly colored group of birds. The resplendent quetzal (*Pharomachrus mocinno*) of Central America is the national bird of Guatemala. *See* AVES.

Walter J. Bock

Bibliography. N. J. Collar, Order Trogoniformes, in J. del Hoyo et al. (eds.), *Handbook of the Birds of the World*, vol. 6, pp. 80–127, Lynx Edicions, 2001; P. A. Johnsgard, *Trogons and Quetzals of the World*, Smithsonian, 2000.

Trojan asteroids

Asteroids located near the equilateral lagrangian stability points of a Sun-planet system (see **illus.**). As shown by J. L. Lagrange in 1772, these are two of the five stable points in the circular, restricted, three-body system, the other three points being located along a line through the two most massive bodies in the system. In 1906 Max Wolf discovered an asteroid located near the lagrangian point preceding Jupiter in its orbit. Within a year, two more were found, one of which was located near the following lagrangian point. It was quickly decided to name these asteroids after participants in the Trojan War as given in Homer's *Iliad*. Hence the term "Trojan asteroid," originally referring to asteroids orbiting the Sun near one of Jupiter's equilateral lagrangian points. With the exception of (624) Hektor in the preceding "swarm" and (617) Patroclus in the following "swarm" (named before this convention was adopted), asteroids in Jupiter's preceding and following lagrangian swarms are named after Greek and Trojan participants, respectively.



Lagrangian points and Trojan asteroids.

The term "Trojans" is sometimes used in a generic sense to refer to objects occupying the equilateral lagrangian points of other pairs of bodies. Small bodies have been found at the equilateral lagrangian points of Saturn's satellites Tethys and Dione: The satellites Calypso and Telesto orbit near the preceding and following points, respectively, of the Saturn-Tethys system, and the satellite Helene orbits near the preceding point of the Saturn-Dione system. Unsuccessful searches have been made for Trojans of the Earth, Saturn, and Uranus, as well as for the Earth-Moon system. Due to its eccentric orbit, and close proximity to Venus, it is considered unlikely that there are any stable regions for Mercury. It is believed that stable regions near the equilateral lagrangian points of Earth, Venus, Saturn, and Uranus exist but are considerably smaller than those of Jupiter. Indeed most of Jupiter's Trojans do not move in the plane of its orbit, but in orbits inclined by as much as 25° and at longitudes differing by up to 40° from the longitudes of the theoretical lagrangian points. *See* SATURN.

On June 20, 1990, D. H. Levy and H. E. Holt at Palomar Observatory discovered an asteroid, later named (5261) Eureka, occupying the following lagrangian point of the planet Mars. In 2001 the first Trojan of Neptune (2001 QR322) was discovered in the course of the Deep Ecliptic Survey, and as of September 2006 three additional Neptunian Trojans had been discovered. These Martian and Neptunian Trojans are the only confirmed non-Jupiter Trojans.

Trojan asteroids are classified using their osculating orbital elements, as opposed to their proper orbital elements. (Osculating orbital elements are the instantaneous orbital elements of the ellipse that the asteroid would follow if all bodies other than the Sun suddenly ceased to exist. Proper orbital elements are the orbital elements averaged over a long period of time according to a precisely specified procedure.) This classification technique works very well for Jupiter Trojans and for the four Neptunian Trojans discovered to date, but not so well for Martian Trojans. Thus, long-term integration of high-quality orbits will be necessary to eventually decide which of the six potential Martian Trojans that had been discovered as of December 2003 are actually Trojans. *See* ORBITAL MOTION.

Jupiter Trojans are dark objects reflecting only between 3 to 9% of the visual light they receive. The majority are compositionally similar to the most common type of outer main-belt asteroid, but some, perhaps as many as one-third, have no known analog among the asteroids or meteorites. *See* ASTEROID; METEORITE.

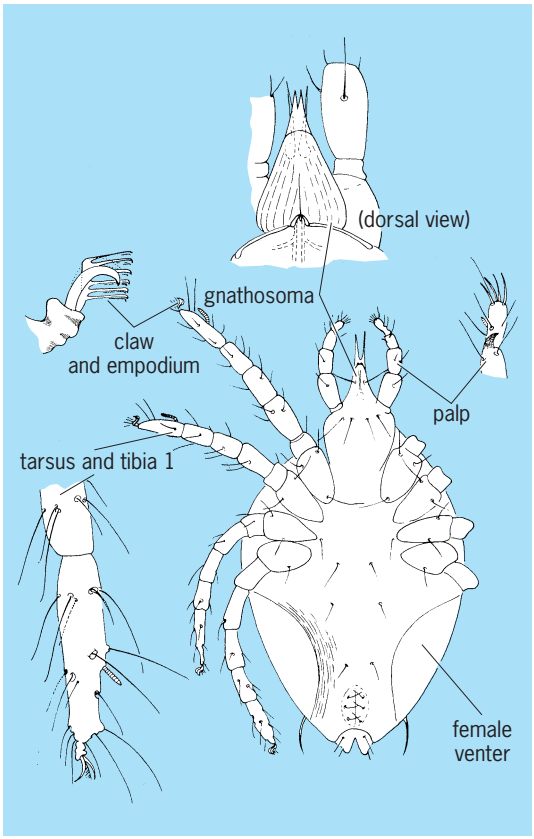
In September 2006 there were 2059 known Jupiter Trojans. Of these, 1139 were in the preceding swarm and 1920 in the following. Albedos (percent of visual light reflected) and diameters have been measured for 71 Jupiter Trojans; the mean albedo for objects in the two swarms is the same, the overall mean albedo being 5.5 ± 1.6%. Analysis of these data indicates that there are about 20 Trojan asteroids with diameters greater than 100 km (60 mi) and about 1750 with diameters exceeding 15 km (9 mi). Of these, 1000 ± 200 are in the preceding swarm and 750 ± 200 are in the following swarm. Approximately 20% of short-period comets are thought to originate from collisions among the Trojan asteroids. *See* ALBEDO; COMET.

Edward F. Tedesco

Bibliography. B. Bottke et al. (eds.), *Asteroids III*, 2002; F. Marzari, P. Tricarico, and H. Scholl, Saturn Trojans: Stability regions in the phase space, *Astrophys. J.* 579:905–913, 2002; S. Tabachnik and N. W. Evans, Cartography for Martian Trojans. *Astrophys. J. Lett.*, 517:L63–L66, 1999; S. A. Tabachnik and N.W. Evans, Asteroids in the inner solar system—I. Existence, *Mon. Not. Roy. Astron. Soc.*, 319:63–79, (2000).

Trombidiformes

A suborder of the Acarina (also known as Prostigmata) commonly called the trombidiform mites, more closely related to the Sarcoptiformes than to the other suborders. They are usually distinguished by presence of a respiratory system opening at or near the base of the chelicerae. Other distinguishing characters are to be found in the tarsi, chelicerae, and genitalia (see **illus.**). These, although variable within the suborder, are distinct from those in all other groups. The Trombidiformes are probably the most heterogeneous group of mites, both morphologically and ecologically, varying from baglike forms with degenerate legs to the highly evolved, fully developed, parasitic forms. There are wormlike forms found in pores of their hosts and flattened types found under scales of lizards; some are parasites in the respiratory tracts of birds, others are free-living predators of other arthropods, and some are plant feeders. They are also variable in their life histories. Some are held within the brood sac of the mother until the siblings, or offspring of the same parents, have had an opportunity to copulate, while others hatch as larvae and pass through a series of molts before becoming sexually mature. Economically, this group contains two families of plant-feeding mites of great importance to agriculture: the Tetranychidae (spider mites) and the Eriophyidae (bud mites or gall mites). The Tarsonemidae, Eupodidae, and Tenuipalpidae are of less importance. The Tetranychidae cause damage by feeding on the leaves and weakening the tree, thus decreasing fruit production and even causing complete defoliation at times. Humans, through commerce, have disseminated some of the most important species throughout the world. The Eriophyidae, by feeding, cause weakening of



Trombidiform mite. (*Institute of Acarology, University of Maryland*)

trees and distortion of fruit and tree growth. A few are vectors of virus diseases of plants. In the United States one transmits streak-mosaic virus of wheat in the Middle West; another, peach mosaic virus in the Far West.

While some of the Tarsonemini feed on plants, others feed on insects. Two species are of particular importance. The hay itch mite, *Pyemotes ventricosus*, normally lives on insect larvae, but when these are destroyed in the process of harvesting grain, the mites may cause a serious dermatitis on humans. The second species, *Acarapis woodi*, causes a disease of honeybees in many parts of the world, which so far has not been reported in the United States.

Medically, the Trombiculidae (chiggers, or red bugs) are important because the larval forms, which are parasites of vertebrates, can cause intense irritation to their hosts by their feeding. More seriously, some transmit a rickettsial disease, scrub typhus, to humans in the Far East and South Pacific regions. The nymphs and adults of the trombiculids are free-living and prey upon eggs of other arthropods. A few families related to the Trombiculidae are predators as nymphs and adults, but the larvae are parasites of arthropods.

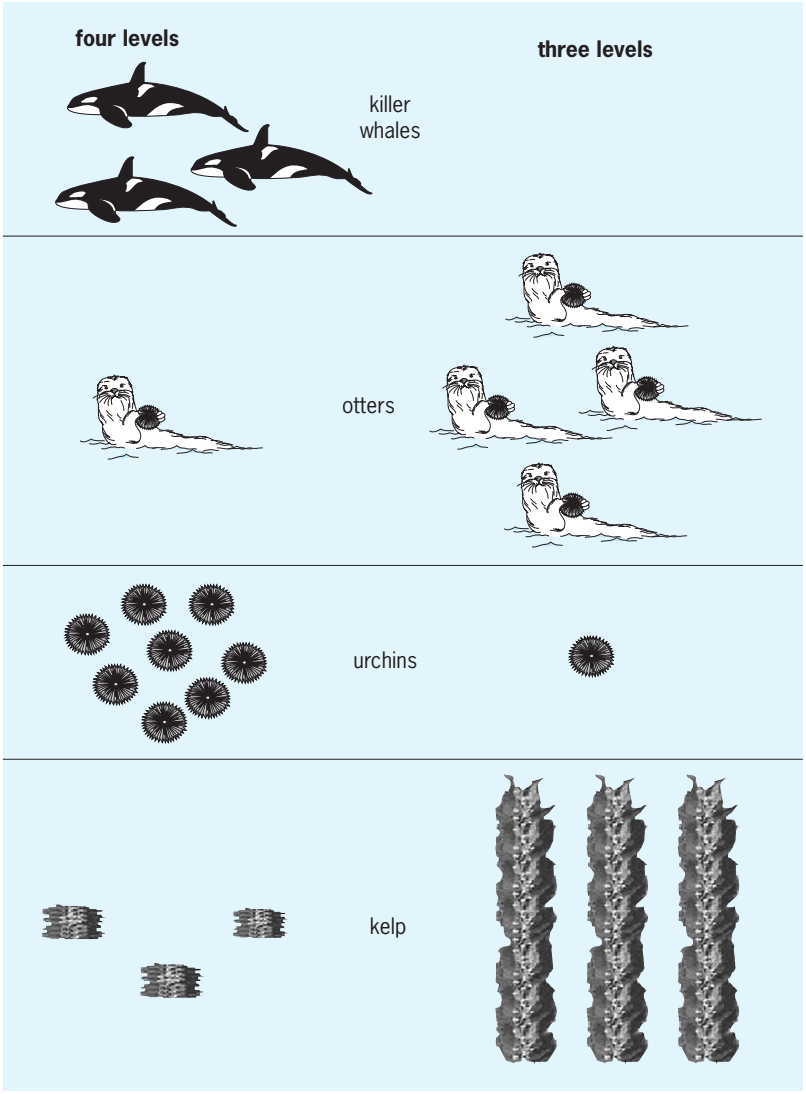
The other families are either free-living and usually considered to be predacious, or are minor parasites of birds, reptiles, and mammals. Of these, the most important are the Demodicidae, or pore mites. *Demodex folliculorum* is frequently present in pores of humans, but it is seldom noticed and is of little

medical importance. Species on domesticated animals produce more apparent symptoms, and *Demodex canis* is at times fatal for dogs. A large and interesting group are the colorful Hydrachnellae which, with few exceptions, are found living in fresh water. A few are parasites of fresh-water mussels, and the larvae of others parasitize aquatic insects. Adults are predacious on small aquatic animals. *See* PLANT PATHOLOGY; PLANT VIRUSES AND VIROIDS.

Edward W. Baker

Trophic ecology

The study of the structure of feeding relationships among organisms in an ecosystem. Researchers focus on the interplay between feeding relationships and ecosystem attributes such as nutrient cycling, physical disturbance, or the rate of tissue production by plants and the accrual of detritus (dead organic material). Feeding or trophic relationships can be represented as a food web or as a food chain. Food webs depict trophic links between all species sampled in a habitat, whereas food chains simplify this complexity into linear arrays of interactions among trophic levels. Thus, trophic levels (for example, plants, herbivores, detritivores, and carnivores) are amalgamations of species that have similar feeding habits. (However, not all species consume prey on a single trophic level. Omnivores are species that feed on more than one trophic level.) *See* ECOLOGY; ECOSYSTEM; FOOD WEB.



Food chain dynamics in subtidal kelp forests. Four-level system: in open coastal areas, killer whales decimate sea otter populations, releasing urchins which are capable of regulating macroalgae. Three-level system: in the absence of killer whales, otter populations increase and prevent urchins from razing kelp forests.

The three fundamental questions in the field of trophic ecology are: (1) What is the relationship between the length of food chains and plant biomass (the total amount of plants at the bottom of the food chain)? (2) How do resource supply to producers (plants) and resource demand by predators determine the relative abundance of organisms at each trophic level in a food chain? (3) How long are real food chains, and what factors limit food chain length?

Effect of food chain length on plant biomass. A central theory in ecology is that "the world is green" because carnivores prevent herbivores from grazing green plant biomass to very low levels. Trophic structure (the number of trophic levels) determines trophic dynamics (as measured by the impact of herbivores on the abundance of plants). Indirect control of plant biomass by a top predator is called a trophic cascade. Cascades have been demonstrated to varying degrees in a wide variety of systems, including lakes, streams, subtidal kelp forests, coastal shrub habitats, old fields, grassland savannas, arctic tundra, shrublands, and below-ground soil communities. In many of these systems, the removal of a top predator has been shown to precipitate dramatic reductions in the abundance (or biomass) of species at lower trophic levels. Food chain theory predicts a green world when food chains have odd numbers of trophic levels, but a barren world (plants suppressed by herbivores) in systems with even numbers of trophic levels. The reduction and subsequent return of sea otters in coastal ecosystems provides a lucid example of cascading trophic effects in marine food chains and alternation between plant-dominated and plant-depleted subtidal habitats.

Sea otters once were abundant in coastal regions from northern Japan to central Baja California, Mexico, but were reduced to a number of widely scattered remnant populations by hunting. The reduction of sea otters likely resulted in dramatic changes in prey population dynamics in ecosystems previously occupied by sea otters. For example, in rocky subtidal habitats the abundance of sea urchins, a preferred prey item of otters, was much higher in habitats with depleted otter populations. In otter-free habitats (two trophic levels), urchin abundance was high enough to overgraze large benthic algae, including kelps. By contrast, in areas supporting remnant populations of otters (three trophic levels), urchin populations were limited to small individuals and

were often restricted to cryptic habitats. As a result, kelp biomass was much higher in these areas. Thus, odd and even food chains lead to green and barren subtidal worlds, respectively (see **illus.**). The result of the presence of otters in nearshore habitats is an increased abundance of kelps and other macroalgae which, in turn, provide habitat and food for a number of associated species. Otters thus play a critical role in structuring subtidal communities. Otters and other species whose effects on the abundance of other species in a food chain or food web are large compared with their relative abundance in the food web are called keystone species.

Recently, sea otters have declined in abundance in the Aleutian Islands for the first time since the regulation of commercial harvest (International Fur Seal Treaty). A convincing hypothesis for the cause of this decline is a shift in the foraging behavior of killer whales from the once abundant baleen whales and pinnipeds (seals, sea lions, and walruses) to sea otters. A comparison of otter abundance in coastal areas of Adak Island, Alaska, where otters were exposed to and protected from killer whale predation, revealed strong effects of these top predators on near-shore ecosystems. Comparison of urchin abundance, grazing rates, and kelp biomass during time periods before and after observed increases in killer whale predation on otters has revealed cascading effects of killer whales in this system. Prior to increased predation by killer whales, coastal subtidal habitats were characterized by low urchin and high kelp abundance. By contrast, these same habitats had much higher urchin densities and more barren substrates immediately following shifts in killer whale foraging and concomitant otter declines. The addition of killer whales (four trophic levels) led to a shift from green to barren worlds as a result of the otters' diminished control of urchin grazing. *See* MARINE ECOLOGY.

Supply and demand in food chain dynamics. Although predators often have strong indirect effects on plant biomass as a result of trophic cascades, both predation (a top-down force) and resource supply to producers (a bottom-up force) play strong roles in the regulation of plant biomass. The supply of inorganic nutrients (such as nitrogen and phosphorus) at the bottom of a food chain is an important determinant of the rate at which the plant trophic level produces tissue (primary production, or productivity) and, in some cases, of the total biomass of this trophic level. However, the degree to which nutrient supply enhances plant biomass accrual depends on how many herbivores are present (which in turn depends on how many trophic levels there are in the system). The relative importance of top-down (demand) versus bottom-up (supply) forces is well illustrated by lake systems, in which the supply of phosphorus (bottom-up force) and the presence of piscivorous (fish-eating) fish (top-down force) have significant effects on the standing stock of phytoplankton, the plant trophic level in lake water columns.

In small lakes of the Canadian Shield of North America, phytoplankton production is strongly dependent on phosphorus. In a classic experiment, a single lake was divided in half with a plastic barrier, and phosphorus, nitrogen, and carbon were added to one side of the lake while just nitrogen and carbon were added to the other (control) side. Blooms of algae turned the water green on the side to which phosphorus was added but not on the control. Side. This result was instrumental in convincing local and national governments in Canada and the United States to regulate phosphate release in sewage to prevent noxious, oxygen-depleting blooms of algae in lakes and other sources of drinking water. *See* FRESHWATER ECOSYSTEM; LAKE; PHYTOPLANKTON; ZOOPLANKTON.

Herbivores may be capable of counteracting nutrient-driven algal blooms, especially in relatively nutrient-poor (oligotrophic) lakes. In these lakes, the effects of top predators, such as largemouth bass, on phytoplankton are analogous to those of killer whales on kelp, and the response of phytoplankton to nutrient loading depends on the number of fish trophic levels in the system. Small lakes have up to two distinct trophic levels of fish species—those that eat zooplankton (zooplanktivores) and those that eat other fish (piscivores). In oligotrophic lakes with only zooplanktivores (that is, which have only three trophic levels), these fish deplete zooplankton which would otherwise graze on phytoplankton, thereby allowing phytoplankton biomass to increase with nutrient (phosphorus) loading from the bottom of the food chain. By contrast, if piscivores such as largemouth bass are added to these same lakes (which then have four-trophic levels), they eat the zooplanktivores, zooplanktivore abundance declines, and zooplankton recover and graze the once green lake until it is barren. In these four-trophic-level systems, phosphorus additions may increase the productivity but not the biomass of the plant trophic level. Thus, top-down control of phytoplankton biomass by predators is possible in oligotrophic lakes, depending on food chain length.

In more fertile (mesotrophic) lakes, piscivores may control zooplanktivore abundance, but the zooplankton (herbivores) are still not capable of keeping pace with increasing phytoplankton production across gradients of increasing nutrient loading. Top-down control attenuates between the third (zooplanktivore) and second (herbivore) trophic levels, and nutrient loading has positive effects on plant productivity and biomass.

In lakes with moderate-to-high nutrient loading, high productivity by species at all trophic levels may preclude control by consumers at the top. In this case, increased nutrient levels lead to increases in the biomass of not only phytoplankton but zooplankton and fish as well. Thus, although plant productivity may increase fairly predictably with nutrient loading, the biomass of the plant trophic level depends both on nutrient supply and trophic structure. Top-down control of plant biomass by top predators appears to be more important in nutrient-poor systems.

As nutrient supply increases, the relative influence of bottom-up control on plant biomass is increasingly important. *See* BIOLOGICAL PRODUCTIVITY; BIOMASS.

Determinants of food chain length. There are three major hypotheses for what determines the length of food chains in anture, based on energy, resilence, and ecosystem size.

Productivity and efficiency. From an energetic perspective, food chain length is limited by two factors: the total rate of plant or bacterial production, and the efficiency at which members of each trophic level assimilate this energy as it moves up the food chain. Productivity determines the total energy supply for a system; however, not all of this energy is incorporated by successively higher trophic levels. Inefficiency in transfer, in either consumption or assimilation, reduces the fraction of total available energy propagated between each trophic level. Typical transfer efficiencies range 5–15% for herbivores and carnivores (for example, zooplankton and bass). In three transfers, the total energy base of the food chain is reduced by more than 99%, limiting the energy available to an additional trophic level. Thus, given fixed transfer transfer efficiencies, additional trophic levels may be added only with increases in productivity at the base of the food chain. Several large-scale syntheses (on lake and arctic tundra island systems) suggest that food chain length does increase with plant productivity. Nevertheless, long food chains (more than five trophic levels) are extremely rare in nature—rare enough to suggest to some ecologists that these food chains may be inherently incapable of persisting. *See* BIOLOGICAL PRODUCTIVITY; ECOLOGICAL ENERGETICS.

Resilience. The idea that the upper limit to food chain length may be determined by the inherent instability of long chains derives from theoretical studies of food chains. In these studies, theoretical ecosystems are constructed as coupled differential equations describing the population dynamics of single-species trophic levels. To analyze food chain stability, disturbance is introduced by changing (reducing or increasing) the abundance at one trophic level. Resilience, an ecological metric for stability, is measured as the inverse of the time required for all trophic levels to return to their previous abundance levels. Longer food chains have consistently longer return times (lower resilience) than short chains (four or less trophic levels), suggesting that longer chains should be rare in nature because return times may be longer than the recurrence interval of disturbance. Although results from model food chains are consistent with field tests in small aquatic systems, the empirical mechanisms behind decreased stability of longer food chains at larger spatial scales are not as clearly developed as the logic of the productivity hypothesis. Moreover, experiments in rivers suggest that disturbance in some cases may act to lengthen food chains, suggesting that the effects of disturbance on food chain length may vary between ecosystem types. *See* SYSTEMS ECOLOGY; THEORETICAL ECOLOGY.

Ecosystem size. One final factor that may set constraints on food chain length is ecosystem size. Size can be defined relatively easily in habitats with discrete boundaries (for example, lakes and oceanic islands). It has been hypothesized that ecosystem size and productivity would interact to determine food chain length. This theory holds that the coverage (in area or volume) of a given level of productivity would provide a more comprehensive measure of the energy supply at the base of food chains and, thus, their potential length. Food chains should increase in length with increasing "productive space" rather than with increasing productivity alone. However, the size of ecosystems alone may provide an equally robust prediction of food chain length. Both body size and home range size may increase with trophic position. Small habitats are simply not large enough to support the home range or provide ample habitat for larger carnivorous species and so may limit the length of food chains. In lake systems, for instance, ecosystem size alone predicts more variability in food chain length than either productivity or productive space. In these systems, large piscivores are often found only in deep waters—habitats found only in larger lakes.

Bioaccumulation of contaminants in top predators. One application of trophic ecology has been the recognition and prevention of bioaccumulation of pesticides and heavy metals in top predators. An example of bioaccumulation is the near-demise of birds of prey such as the bald eagle, osprey, and peregrine falcon in the United States as a result of the formerly unrestricted application of DDT and other organochlorine pesticides to crop fields, lakes, and ponds to kill insect pests (crop herbivores and mosquitoes).

Although DDT is not lethal to animals at higher trophic levels, it is not excreted by them. Thus, birds and fish that consume large quantities of insects or other herbivores exposed to DDT concentrate the pesticide in their tissues. Birds of prey, in turn, further concentrate DDT because they rely heavily on fish or other birds as prey species (two trophic transfers of DDT). Peregrine falcons may be especially vulnerable to DDT magnification because in some areas these birds consume fish-eating seabirds (three trophic transfers of DDT). Birds with high pesticide levels tend to lay eggs with abnormally thin shells that crack during incubation. As a result of impaired reproduction, peregrine falcon populations showed dramatic declines throughout the world. Recognition of this problem led to a ban on the use of DDT and other stable pesticides in many industrialized countries. DDT was outlawed in the United States in 1972; this ban allowed the eventual recovery of peregrine falcons to over 80 breeding pairs in the United States. Unfortunately, DDT is still used in many developing nations. Although the effect that this chemical may have on humans is still unknown, its effect on raptors suggests that humans could also be vulnerable.

Many fish species common in markets around the world are top predators and may be reservoirs for

other common toxins such as mercury. Interestingly, fish from ecosystems with long food chains may pass along more mercury to humans than those from food chains with fewer trophic levels and, thus, fewer trophic transfers of this toxin. *See* ECOLOGY, APPLIED; INSECTICIDE; PESTICIDE. John L. Sabo; Leah R. Gerber

Bibliography. S. R. Carpenter and J. F. Kitchell (eds.), *The Trophic Cascade in Lakes*, Cambridge University Press, 1993; R. Carson, *Silent Spring*, Fawcett Crest, New York, 1962; S. L. Pimm, *Food Webs*, Chapman & Hall, New York, 1982; G. A. Polis and K. O. Winemiller (eds.), *Food Webs: Integration of Patterns and Dynamics*, Chapman & Hall, New York, 1996.

Tropic of Cancer

The parallel of latitude about $23^1/_2^\circ$ (23.45°) north of the Equator. The importance of this line lies in the fact that its degree of angle from the Equator is the same as the inclination of the Earth's axis from the vertical to the plane of the ecliptic. Because of this inclination of the axis and the revolution of the Earth in its orbit, the vertical overhead rays of the Sun may progress as far north as $23^1/_2^\circ$. At no place north of the Tropic of Cancer will the Sun, at noon, be 90° overhead.

On June 21, the summer solstice (Northern Hemisphere), the Sun is vertical above the Tropic of Cancer. On this same day the Sun is 47° above the horizon at noon at the Arctic Circle, and at the Tropic of Capricorn, only 43° above the horizon. The Tropic of Cancer is the northern boundary of the equatorial zone called the tropics, which lies between the Tropic of Cancer and Tropic of Capricorn. *See* LATITUDE AND LONGITUDE; MATHEMATICAL GEOGRAPHY; SOLSTICE. Van H. English

Tropic of Capricorn

The parallel of latitude approximately $23^1/_2^\circ$ (23.45°) south of the Equator. It was named for the constellation Capricornus (the goat), for astronomical reasons which no longer prevail.

Because the Earth, in its revolution around the Sun, has its axis inclined $23^1/_2^\circ$ from the vertical to the plane of the ecliptic, the Tropic of Capricorn marks the southern limit of the zenithal position of the Sun. Thus, on December 22 (Southern Hemisphere summer, but northern winter solstice) the Sun, at noon, is 90° above the horizon. On this same day, at noon, the Sun is 47° above the horizon at the Antarctic Circle, $66^1/_2^\circ$ at the Equator, and 43° at the Tropic of Cancer. Sun rays will just reach the horizon tangentially at the Arctic Circle.

The Tropic of Capricorn is the southern boundary of the equatorial zone referred to as the tropics, which lies between the Tropic of Capricorn and the Tropic of Cancer. *See* LATITUDE AND LONGITUDE; MATHEMATICAL GEOGRAPHY; SOLSTICE. Van H. English

Tropical meteorology

The study of atmospheric structure and behavior in the areas astride the Equator, roughly between 30° north and south latitude. The weather and climate of the tropics involve phenomena such as trade winds, hurricanes, intertropical convergence zones, jet streams, monsoons, and the El Niño Southern Oscillation. More energy is received from the Sun over the tropical latitudes than is lost to outer space (infrared radiation). The reverse is true at higher latitudes, poleward of 30°. The excess energy from the tropics is transported by winds to the higher latitudes, largely by vertical circulations that span roughly 30° in latitudinal extent. These circulations are known as Hadley cells, after George Hadley who first drew attention to the phenomenon in 1735. This type of circulation is an important ingredient of the tropical general circulation.

For the most part, the oceanic tropics (the islands) experience very little change of day-to-day weather except when severe events occur. Tropical weather can be more adverse during the summer seasons of the respective hemispheres. The near equatorial belt between 5°S and 5°N is nearly always free from hurricanes and typhoons: the active belt lies outside this region over the tropics. The land areas experience considerable heating of the Earth's surface, and the summer-to-winter contrasts are somewhat larger there. For instance, the land areas of northern India experience air temperatures as high as 108°F (42°C) in the summer (near the Earth's surface), while in the winter season the temperatures remain 72°F (22°C) for many days. The diurnal range of temperature is also quite large over land areas on clear days during the summer (32 °F or 18 °C) as compared to winter (18 °F or 10 °C).

Weather observations. Vast areas of the tropics are oceanic, and there is a general lack of surface and upper-air observations of temperature, wind, pressure, and humidity over most regions. The network for weather observations over the land areas of Africa and South America are rather limited in comparison to the rest of the inhabited tropical areas. Tropical meteorology has greatly benefited from the space-based observations by meteorological satellites, which provide day and night images of cloud cover from visible and infrared sensors and can resolve as high as a few kilometers over the Earth's surface. There are two types of satellites—the polar orbiter and the geostationary; the latter are over the Equator at a height of roughly 22,500 mi (36,000 km). Five geostationary satellites are in orbit for complete global tropical coverage, providing images and tropical air motions at the lower tropical troposphere (about 0.6 mi or 1 km above the Earth's surface) and at the upper troposphere (about 7 mi or 12 km above the Earth's surface). The air motions are estimated from cloud tracking from adjacent frames (some 30 min apart) of cloud images at high resolution. With satellite observations, it is possible to monitor tropical weather phenomena on a daily basis around the global belt.

See SATELLITE METEOROLOGY; WEATHER OBSERVATIONS.

Trade winds. The steady northeast surface winds over the oceans of the Northern Hemisphere between 5° and 20°N and southeast winds over the corresponding latitudes of the southern oceans constitute the trade winds. Trade winds have intensities of around 5–10 knots (2.5–5 m/s). They are the equatorial branches of the anticyclonic circulation (known as the subtropical high pressure). The steadiness of wind direction is quite high in the trades. On the equatorial side of the strongest trade winds (known as the cyclonic shear side), tropical depressions usually form over warm oceans. The trade winds of the Southern Hemisphere (Atlantic, Pacific, and Southern Indian oceans) exhibit more disturbance activity. Trade winds are not present north of the Equator over the Indian Ocean; this region experiences southwesterly monsoonal flows. Near the Equator, the trade winds carry moisture to rain bands known as the intertropical convergence zones. Air-sea interaction in the Atlantic is strongly affected by the steady and nonsteady components of the trade wind systems of the two hemispheres. In addition, the air-sea interaction varies strongly during passage of disturbances, such as hurricanes. The Bowen ratio (sensible heat flux/latent heat flux) over this region is generally much less than 1; the dominant transfer is that of the latent heat. In strong trades, fluxes of latent heat of the order of 275 watts/m^2 are of common occurrence. If a hurricanelike disturbance is present, then on the mesoscale, fluxes of the order of 500 watts/m^2 are noted from aircraft and model-based estimates. The more important area is that of air-sea coupling during the passage of hurricanes. A cold wake with ocean temperature anomalies of the order of 3–5°F (2–3°C) is often generated by the action of strong winds. The lifting of the thermocline and the resulting upwelling of colder waters gives rise to these cold wakes. Passage of one hurricane often leaves such a wake that affects the intensity of a second hurricane that follows it. *See* WIND.

Hurricanes. Hurricanes are also known as typhoons in the west Pacific and tropical cyclones in the Indian Ocean and south Pacific. If the wind speed exceeds 65 knots (33 m/s) in a tropical storm, the storm is labeled a hurricane. A hurricane usually forms over the tropical oceans, north or south of 5° latitude from the Equator. The strongest winds are found in the lowest levels above the ocean; however, the winds weaken with height very slowly in the troposphere and, therefore, a sizable strength of the vortex can be seen all the way up to 6 mi (10 km). The vertical extent of a hurricane extends up to 7 mi (12 km); around that level the cyclonic circulation (clockwise in the Southern Hemisphere and counterclockwise in the Northern Hemisphere) tapers off with an anticyclonic circulation, with outflowing air that extends to an altitude of 9 mi (15 km). The inflowing air is found in the planetary boundary layer, that is, mostly in the lowest mile over the ocean, and spirals into the storm in a few marked channels known as spiral rain bands that converge toward an eye wall of the hurricane, whose diameter is of the order of 6–12 mi (10–20 km). Hurricanes form in regions of warm tropical oceans (with temperatures generally above 81°F or 27°C), where the variation of wind with height over the troposphere is small. They possess a warm thermal core and usually form from incipient tropical weaker cyclonic weather systems. The source of energy that drives these storms is the evaporation from the oceans as a result of wind action. Great efforts have been placed on intense real-time monitoring of hurricanes using aircraft, radar, and satellites. A mix of these platforms has provided dropwindsonde winds, dual Doppler radar–based hydrometeor structure, and satellite-based indirect estimates of precipitation using onboard radar and microwave instrumentation. Numerical models have, on the mesoscale, improved the short-range predictive capability of tracks, intensity, and precipitation amounts. Global models, on the medium-range time frame, up to 6 days into the future, have developed sophisticated data assimilation systems to absorb all these diverse data sets. The global models have also shown marked improvement of skills in the above areas. *See* HURRICANE; TROPOSPHERE.

Intertropical convergence zones. These zones are located usually between 5 and 10°N latitude. They are usually oriented west to east and contain cloud clusters with rainfall of the order of 1.2–2 in. (30–50 mm) per day. The trade winds of the two hemispheres supply moisture to this precipitating system. Embedded within this line of cloud clusters are westward propagating tropical waves that usually move at a speed of roughly 300 mi (500 km) per day. Occasionally these waves and associated cloud systems move away from the equatorial latitudes, amplify, and form tropical depressions and eventually hurricanes. The intertropical convergence zone tends to form over the warm oceans and move slowly poleward to almost 10° from the Equator during the warmest summer months over the respective hemisphere. *See* CLOUD PHYSICS; PRECIPITATION (METEOROLOGY).

Jet streams. A number of fast-moving air currents, known as jets, are important elements of the tropical general circulation. With speeds in excess of 30 knots (15 m/s), they are found over several regions of the troposphere. The most prominent of these, the tropical easterly jet, is found near 5–10°N latitude over the Asian regions and extends to the west African Atlantic coast. The core of maximum winds is located roughly 8 mi (14 km) above sea level. It forms on the equatorward side of the Asian and west African summer monsoon. During this monsoon period, there also occurs a somewhat weaker but important lower tropospheric easterly jet stream over west Africa. The core of this jet is located generally close to 14°N latitude at around 2.4 mi (4 km) above the Sahara desert. Many air motions form on the equatorward side of this jet. These waves, known as African waves, have east-west scales of the order of 1800 mi (3000 km); they move westward out of Africa into the Atlantic Ocean. Some of these waves amplify and form hurricanes over the tropical Atlantic Ocean.

Another important jet, occurring during the winter seasons of the two hemispheres, is known as the subtropical westerly jet stream. It is most prominent during December, January, and February over the Northern Hemisphere and is located at an altitude of about 7 mi (12 km) above the Earth's surface, near 27.5°N latitude. This jet stream encircles the globe in a basic three-wave pattern that is relatively stationary. Strong winds of the order of 70 knots (35 m/s) are found near the southeastern United States coast and over the north African Mediterranean coast. The third wave is located off southern Japan, where winds as strong as 50 knots (100 m/s) are of common occurrence. The subtropical jet of the Southern Hemisphere is strongest during July and August. Its latitude is close to 25°S, and its core is located at the tropopause level near 7.2 mi (12 km). *See* ATMOSPHERIC GENERAL CIRCULATION; JET STREAM.

Monsoons. Basically the entire landmass from the west coast of Africa to Asia and extending to the date line experiences a phenomenon known as the monsoon. Monsoon circulations are driven by differential heating between relatively cold oceans and relatively warm landmasses. The west African region near 5–10°N latitude, the Indian region between 10 and 20°N latitude, and the east Asian and Chinese region between 10 and 40°N latitude experience the northern summer monsoon. An annual cycle of monsoon reveals itself as a north-south oscillation of the precipitating systems. These systems are located near the Indonesia-Borneo-northern Australian regions during the northern winter months, and they move northward toward the eastern foothills of the Himalaya by late June. Rainfall amounts in the active monsoon can exceed 120 in. (3000 mm) per month. The heaviest monsoon rainfall occurs at the foothills of the mountains, where rainfall amounts have been known to be as large as 100 in. (2500 mm) per month. It is not uncommon for single rainfall events, during the passage of a monsoon disturbance, to bring in amounts of order of 8 in. (200 mm) per day. Warm oceans are the source of moisture carried by monsoonal lower tropospheric flows. In the tropics the troposphere extends from the surface to about 8.4 mi (14 km). Over the northern summer monsoon this is a southwesterly flow, and in the northern winter it is generally a northeasterly flow. *See* MONSOON METEOROLOGY.

El Niño Southern Oscillation. Every 2–6 years the eastern equatorial Pacific Ocean experiences a rise in sea surface temperature of about 5–9°F (3–5°C). This phenomenon is known as El Niño, which is part of a larger cycle referred to as the El Niño Southern Oscillation (ENSO). The other extreme in the cycle is referred to as La Niña. El Niño has been known to affect global-scale weather.

A complete understanding of El Niño begins with how the ocean and the atmosphere operate in the equatorial Pacific region under normal conditions. Fluctuations in the atmosphere pressure occur between the west Pacific and Indonesia, and constitute the Southern Oscillation. The state of the pressure system of the Southern Oscillation is characterized by the Southern Oscillation Index, which is a measure of the pressure difference anomaly between Papeete, Tahiti, and Darwin, Australia. The difference between high pressure in the west Pacific and low pressure in Indonesia helps to drive the easterly trade winds along the Equator. Strong trade winds pile up the water in the west and actually cause the sea level to be an average of 16 in. (40 cm) higher in the west Pacific than in the east Pacific. In a simplified model, the Pacific Ocean can be considered as having two layers separated by a thermocline. The upper layer is warm and well mixed, while the lower layer is cold and well stratified. In the west Pacific basin the upper mixed layer is thicker than in the east Pacific. The difference in the thickness of the upper mixed layer leads to differences in the sea surface temperatures across the ocean. The water in the east Pacific is generally colder because of upwelling. Upwelling is a process caused by wind stress on the eastern boundary of the Pacific Ocean. The easterly winds tend to pull the water away from the coast. This water is replaced by cooler water from beneath the thermocline. Hence, the coastal waters of Ecuador and Peru are generally colder than the west Pacific. *See* WIND STRESS.

El Niño, considered the warm phase of the Southern Oscillation, is marked by a few prevalent conditions. First the Southern Oscillation Index decreases, inducing a westerly wind anomaly, that is, a weakening of the easterly trade winds. This forcing by the atmosphere causes a response by the ocean in the form of internal Kelvin waves. In the west Pacific Ocean a Kelvin wave is generated that propagates eastward. It takes 2–3 months for this wave to reach the coastline of South America. The Kelvin wave increases the thickness of the upper mixed layer in the central and eastern equatorial Pacific. The upwelling process now takes place in the warm mixed layer instead of bringing up cold water from below the thermocline. This causes the sea surface temperature to rise in the central and east Pacific. Tropical convective clouds and rain form over the anomalously warm oceanic region. Large areas of rising air are found in these regions of heavy rain, while sinking air is found surrounding the regions. As this warm pool of water moves eastward, it brings convection. These changes in the ocean and atmosphere in the equatorial Pacific have been shown to affect the global climate. *See* CLIMATE HISTORY; DROUGHT; EL NIÑO; MARITIME METEOROLOGY; METEOROLOGY; UPWELLING.

Biennial oscillation. The quasi-biennial oscillation (QBO) is a reversal of the west-to-east winds in the equatorial latitudes in the stratosphere. This reversal has a time scale of roughly 26 months. Vertical propagation of energy (and downward phase propagation) is a characteristic feature of this oscillation. A slow downward phase propagation with alternating westerlies and easterlies characterizes this system. It has its strongest amplitude near the 19-mi (30-km) level. It is generally seen clearly to about 9 mi (15 km). A tropospheric biennial oscillation (TBO) on the time scale of roughly 2 years has also drawn considerable

interest. It has been noted in the Asian monsoon longitude and the west Pacific Ocean. This time scale has been noted in the air-sea fluxes, the monsoon precipitation, and the elevation of pressure surfaces. Both the QBO and TBO are being studied intensely, and possible relationships among them are being investigated.

Weather prediction. Barotropic forecasts based on the principle of conservation of absolute vorticity was the centerpiece for tropical modeling in the early 1960s. Since then the progress has been steady. Multilevel complete physical/dynamical models at very high resolution are currently being used for prediction. Much thought has been given to the issue of how the effect of cumulus convection has to be included within the large-scale models. This is the area of cumulus parametrization. The coverage of data is a major issue. Large amount of satellite-based data sets provide cloud-tracked winds, water vapor–tracked winds, precipitation rates, and sea surface temperatures. Assimilation of these data sets has provided improvement in hurricane track and intensity forecasts and tropical precipitation forecasts. One area that has contributed to forecast skill improvement is ensemble forecasts. Here a large number of forecasts from the same start date provide a robust ensemble. Combining statistics and the ensemble of forecasts, it has been possible to make some major improvements in tropical forecasts. T. N. Krishnamurti

Bibliography. S. Ackerman and J. A. Knox, *Meteorology: Understanding the Atmosphere*, 2d ed., 2006; G. D. Atkinson, *Forecasters' Guide to Tropical Meteorology*, 2002; G. R. McGregor and S. Nieuwolt, *Tropical Climatology: An Introduction to the Climates of the Low Latitudes*, 2d ed., 1998; B. Wang, *The Asian Monsoon*, 2006.

Tropopause

The boundary between the troposphere and the stratosphere in the atmosphere. The tropopause is broadly defined as the lowest level above which the lapse rate (decrease) of temperature with height becomes less than 6°F/mi (2°C/km). In low latitudes the tropical tropopause is at a height of 9–10.5 mi or 15–17 km (∼−135°F or 180 K), and the polar tropopause between tropics and poles is at about 6 mi or 10 km (∼−63°F or 220 K). There is a well-marked "tropopause gap" or break where the tropical and polar tropopause overlap at 30–40° latitude. The break is in the region of the subtropical jet stream and is of major importance for the transfer of air and tracers (humidity, ozone, radioactivity) between stratosphere and troposphere. Tropopause breaks also occur in the neighborhood of polar jet streams. The height of the tropopause varies seasonally and also daily with the weather systems, being higher and colder over anticyclones than over depressions. The detailed vertical temperature structure is often complex, showing multiple or laminated tropopauses, and it is often difficult to decide on the precise height of the tropopause, particularly in winter at high latitudes. *See* AIR TEMPERATURE; ATMOSPHERE; RADIO-WAVE PROPAGATION; STRATOSPHERE; TROPOSPHERE. R. J. Murgatroyd

Troposphere

The lowest major layer of the atmosphere. The troposphere extends from the Earth's surface to a height of 6–10 mi (10–16 km), the base of the stratosphere. It contains about four-fifths of the mass of the whole atmosphere. *See* ATMOSPHERE.

On the average, the temperature decreases steadily with height throughout this layer, with a lapse rate of about 18°F/mi (6.5°C/km), although shallow inversions (temperature increases with height) and greater lapse rates occur, particularly in the boundary layer near the Earth's surface. Appreciable water-vapor contents and clouds are almost entirely confined to the troposphere. Hence it is the seat of all important weather processes and the region where interchange by evaporation and precipitation (rain, snow, and so forth) of water substance between the surface and the atmosphere takes place. *See* ATMOSPHERIC GENERAL CIRCULATION; CLIMATOLOGY; CLOUD PHYSICS; METEOROLOGY; WEATHER. R. J. Murgatroyd

Tropospheric scatter

A term applied to propagation of radio waves caused by irregularities in the refractive index of air. The phenomenon is predominant in the lower atmosphere; little or no scattering of importance occurs above the troposphere. Tropospheric scatter propagation provides very useful communication services but also causes harmful interference. For example, it limits the geographic separation required for frequency assignments to services such as television and frequency-modulation broadcasting, very high-frequency omnidirectional ranges, and microwave relays. It is used extensively throughout most of the world for long-distance point-to-point services, particularly where high information capacity and high reliability are required. Typical tropospheric scatter relay facilities (**Fig. 1**) are commonly 200–300 mi (320–480 km) apart. Some single hops in excess of 500 mi (800 km) are in regular use. High-capacity circuits carry 200–300 voice circuits simultaneously. *See* TROPOSPHERE.

Theory. To scatter is to spread at random over a surface or through a space or substance. Tropospheric scattering that tends to be coherent is more properly called forward scatter, reflection, refraction, focusing, diffraction by atmospheric inhomogeneities, or all of these. In the absence of the atmosphere, only diffraction by the earth would support propagation to distances well beyond the horizon. Atmospheric refractive index variations, which account for the remaining propagation mechanisms, are the result of variations in temperature, pressure, and gaseous constituents, the main variable being water vapor. The approximate formula for refractivity N, expressing changes in parts per million of the radio refractive

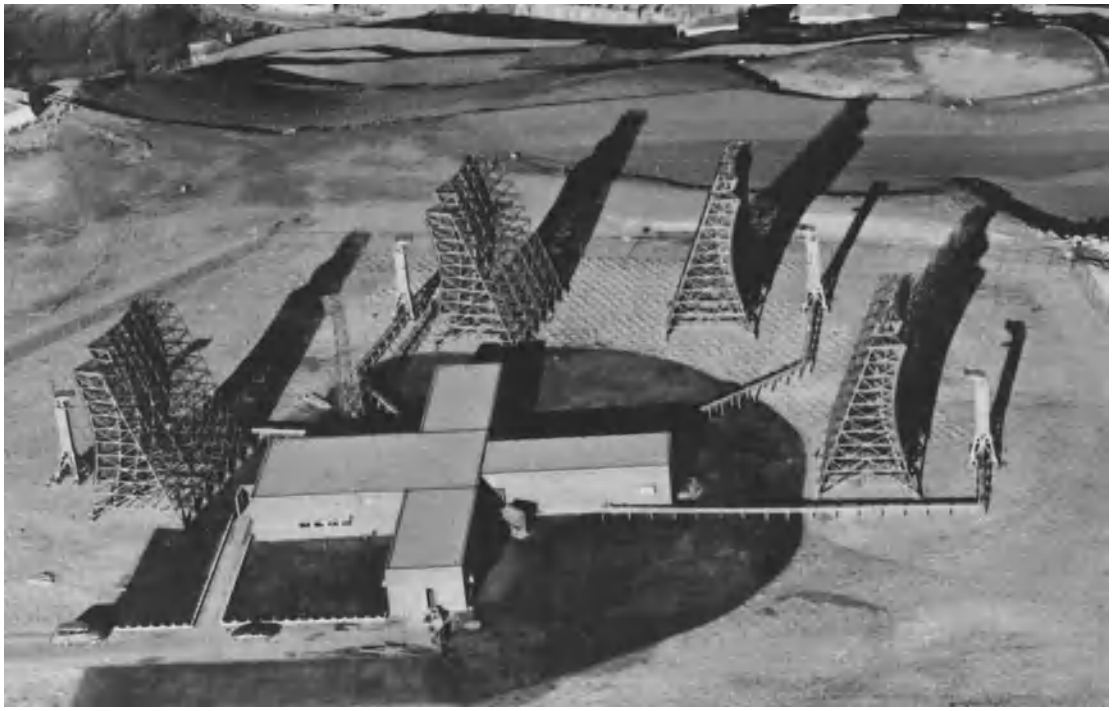


Fig. 1. Tropospheric scatter relay facility operated by the U.S. Air Force in Spain. High-powered ultra-high-frequency transmitters, sensitive receivers, and large parabolic antennas operating with frequency and space diversity make it possible to transmit many voice and teletype circuits simultaneously far beyond the horizon. (*Page Engineers, Inc.*)

index of air compared to that for a vacuum, is given by Eq. (1), where T is temperature in kelvins, P is total pressure in millibars, and e is partial water-vapor pressure in millibars.

$$N = \frac{77.6}{T}\left(P + \frac{4810e}{T}\right) \qquad (1)$$

Scattering from refractive index discontinuities, reflection from layers, and ducting from steep vertical gradients of refractive index account for nearly all of

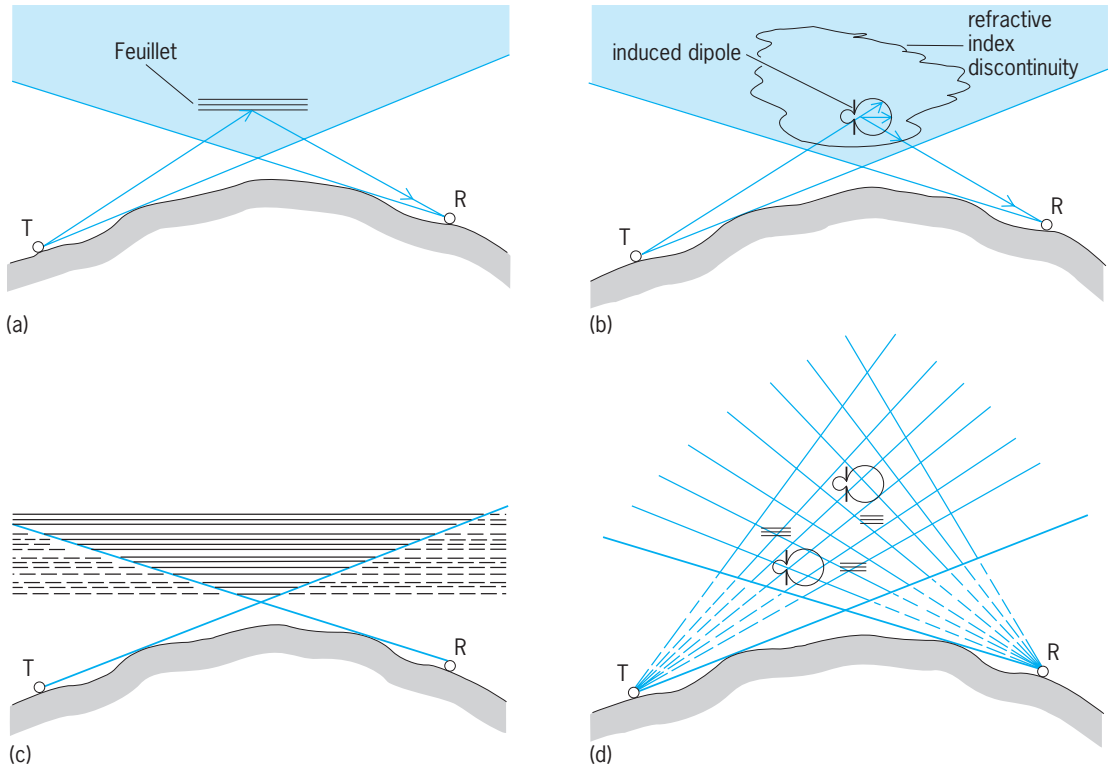


Fig. 2. Geometry for tropospheric forward scatter transmission from transmitter T to receiver R. (*a*) Feuillet. (*b*) Dipole induced in tropospheric layers. (*c*) Continuous scattering layers. (*d*) Limited scattering layers.

the radio energy at large distances beyond the horizon. These long-distance fields change with time. Rapid fading is caused by multipath components and is termed phase interference fading. The slower fading of 5-min or hourly medians from hour to hour and day to day throughout the year is called power fading and is caused by slowly changing atmospheric characteristics and by changes in the relative dominance of various propagation mechanisms.

The theory of tropospheric scattering is based on a combination of several models: The mechanism indicated by **Fig. 2***a* involves a tropospheric layer, or feuillet, which has a sufficiently abrupt change in refractive index, usually associated with fair weather conditions, to reflect a substantial amount of radio energy at the grazing angles and frequencies of interest. Layers can be either continuous as in Fig. 2*c*, or limited as in Fig. 2*d*. The most nearly specular reflections from atmospheric layers are usually observed between 30 and 200 MHz. At higher frequencies, where focusing, defocusing, and ducting are common and where extensive layers are not sufficiently abrupt or sufficiently numerous to provide strong reflections, a number of small, randomly oriented surfaces come into play. The radio wave is scattered forward by all the scattering subvolumes visible to both antennas, as indicated in Fig. 2*c* and *d*.

Data. Available long-term median forward scatter radio transmission loss data usually show attenuation in power inversely proportional to the wavelength cubed (λ^3). Long-term measurements on two or more frequencies rarely show power attenuation ratios outside the range λ^2 to λ^4. The λ^2 dependence is also characteristic of free-space propagation and of omnidirectional scattering by precipitation, and reflects the λ^2 dependence of the effective absorbing area of a receiving antenna. As a first approximation, Eq. (2) gives a simple formula for the long-term

$$L_{bm} = 46 + 30 \log_{10} \text{frequency (MHz)} + 20 \log_{10} \text{distance (miles)} + 0.1 \text{ distance (miles)} \quad (2)$$

median basic transmission loss L_{bm} for tropospheric scatter paths. L_{bm} is defined as the ratio, in decibels, of power delivered to an isotropic transmitting antenna to power from an isotropic receiving antenna, both antennas being lossless. *See* RADIO-WAVE PROPAGATION.

Robert S. Kirby

Bibliography. M. P. Hall, *Effects of the Troposphere on Radio Communication*, 1980; S. Shibuya, *A Basic Atlas of Radio-Wave Propagation*, 1986; M. Valkenburg, *Reference Data for Engineers: Radio, Electronic, Computer, and Communications*, 8th ed., 1996.

Truck

A motor vehicle ("lorry" in British English) carrying its load on its own wheels and primarily designed for the transportation of goods or cargo.

A truck is similar to a passenger car in many basic aspects, but truck construction is usually heavier throughout with strengthened chassis and suspension, and lower transmission and drive-axle ratios to cope with hilly terrain. Other common truck characteristics include cargo-carrying features such as rear doors or a tailgate, and a flat floor. However, there are many different kinds of trucks, often specially designed with unique features for performing a particular job, including catering trucks, cement trucks, dump trucks, fire trucks, flat-bed trucks, pickup trucks, refrigerated trucks, tank trucks, and walk-in van-bodied trucks. *See* AUTOMOBILE; BUS.

Types. A truck is rated by its gross vehicle weight (gvw), which is the combined weight of the vehicle and load. Trucks are classified as light-, medium-, or heavy-duty according to gross vehicle weight as follows:

Light-duty trucks
- Class 1: 0–6000 lb (0-2700 kg)
- Class 2: 6001-10,000 lb (2700-4500 kg)
- Class 3: 10,001-14,000 lb (4500-6300 kg)

Medium-duty trucks
- Class 4: 14,001-16,000 lb (6300-7200 kg)
- Class 5: 16,001-19,500 lb (7200-8775 kg)
- Class 6: 19,501-26,000 lb (8775-11,700 kg)

Heavy-duty trucks
- Class 7: 26,001-33,000 lb (11,700-14,850 kg)
- Class 8: 33,001 lb up (14,850 kg up)

Although a variety of models and designs are available in each category, there are two basic types of vehicles, the straight truck and the truck tractor. The straight truck has the engine and body mounted on the same chassis. The chassis includes the engine, frame, and other essential structural and mechanical parts, but not the body. The body is the structure or fixture especially provided to contain or support the goods or cargo to be transported.

The truck tractor is essentially a power unit that is the control and pulling vehicle for truck trailers such as full trailers or semitrailers. A full trailer has a front axle and one or more rear axles, and is constructed so that all its own weight and that of its load rests on its own wheels. A semitrailer has one or more axles at the rear, and is constructed so that the front end and a substantial part of its own weight and that of its load rests upon another vehicle. A retractable mechanism mounted on the front end of the semitrailer is lowered to support it when the pulling vehicle is disconnected. A full trailer may be drawn by a truck or behind a semitrailer.

Truck tractor. A truck tractor is a vehicle of short wheelbase for hauling semitrailers. It carries a swiveling mount, known as the fifth wheel, above the rear axle to support the front end of the semitrailer. If the tractor has two axles, the drive is through the rear axle. However, a three-axle tractor (one front and two rear) may drive through only one rear axle with one trailing, or through both rear axles.

The tractor-semitrailer combination permits the

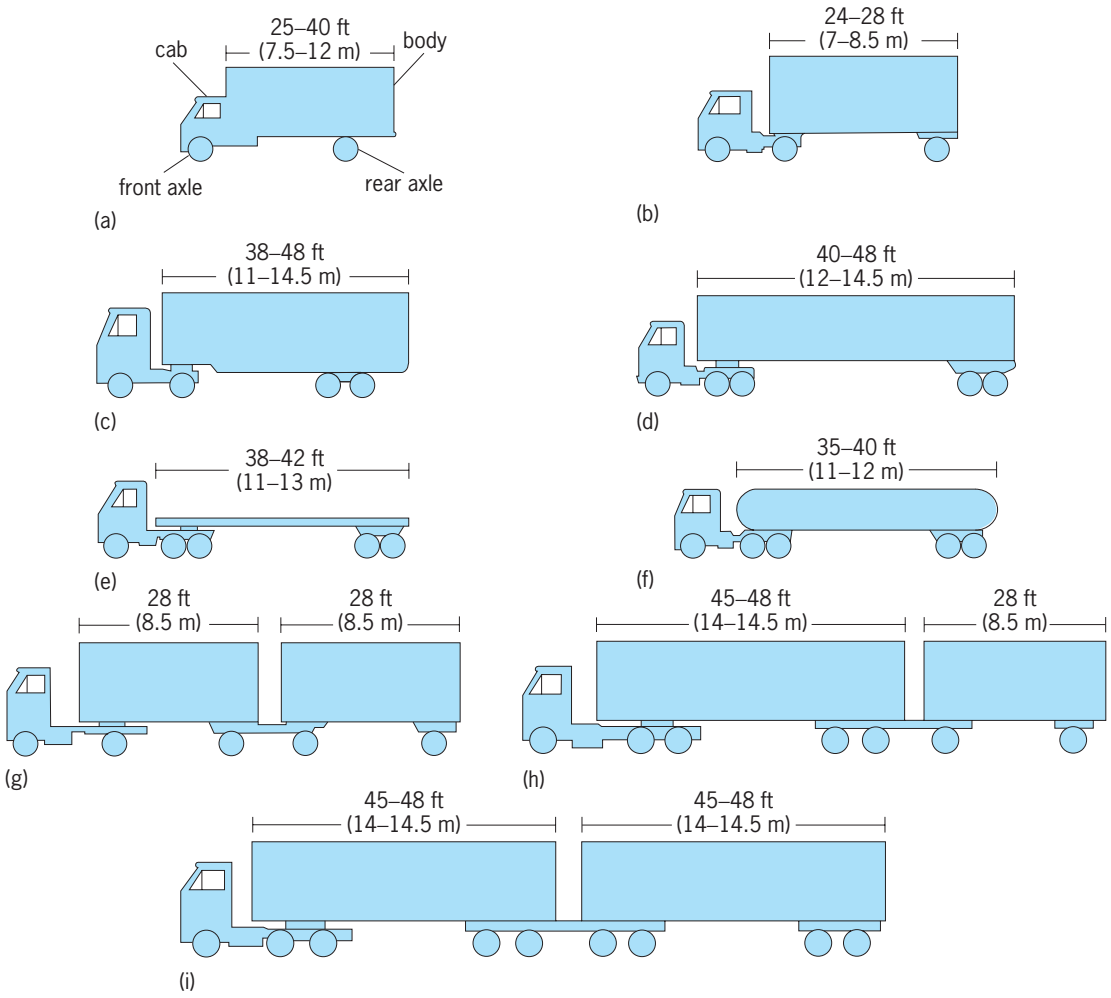


Fig. 1. Various types of truck, truck tractor, and trailer combinations. Lengths shown are typical. Other lengths are possible depending on the carrier's needs and state laws. For each type, the designation in brackets (see below) identifies the various combinations by axles: the first digit refers to the number of axles in the power unit; a second digit refers to a full trailer, while a second digit with prefix S refers to a semitrailer. For example, 3-S2-2 identifies a three-axle truck trailer pulling a two-axle semitrailer and a two-axle full trailer. (*a*) Straight truck [2]. (*b*) Three-axle tractor semitrailer [2-S1]. (*c*) Four-axle tractor semitrailer [2-S2]. (*d*) Five-axle tractor semitrailer [3-S2]. (*e*) Five-axle tractor flatbed trailer [3-S2]. (*f*) Five-axle tractor tank trailer [3-S2]. (*g*) Twin trailer or doubles [2-S1-2]. (*h*) Rocky Mountain doubles, operated only in certain states [3-S2-2]. (*i*) Turnpike doubles, operated only in certain states [3-S2-4].

use of longer bodies with greater carrying capacity and better maneuverability than is possible with a straight truck. The forward positioning of the cab, the short wheelbase of the tractor, and the multiplicity of axles provide maximum payloads and operating economy in the face of restriction on overall length imposed by some states, and regulations limiting the weight carried on a single axle. **Figure 1** shows various types of truck and of truck tractor and trailer combinations.

The standard Class 8 tractor-semitrailer combination, which has five axles and is often called an 18-wheeler, generally is limited to a gross combination weight (gcw) of 80,000 lb (36,300 kg) for operation on U.S. Interstate highways.

Cab and body. The cab is the part of the truck or tractor that encloses the driver and vehicle operating controls. It may be an integral part of the body, as in a van; or it may be a separate compartment alongside the engine, behind the engine, or over the engine. With the cab-over-engine design, the cab may be in fixed position or it may tilt forward for access to the engine. A cab with an interior or adjacent sleeping space is known as a sleeper or sleeper cab. The truck cab also serves many drivers as a mobile office. Its electronics may include a wireless computer, multifunction printer and fax machine, and global positioning system (GPS), which tracks vehicle location for security, navigation, and recommended routing that will be the most time-, cost-, and fuel-efficient. *See* SATELLITE NAVIGATION SYSTEMS.

At highway speeds, half the fuel burned in the engine is used to overcome the air resistance or aerodynamic drag of the cab, body, and trailer. Streamlined designs—for example, contoured windshields, hoods, fenders, and bumpers and devices such as air deflectors—reduce drag. This improves fuel economy and lowers the vehicle operating cost. *See* STREAMLINING.

Frame. The truck frame (**Fig. 2**) supports the load, power train, and steering mechanism while maintaining alignment of the components of the body

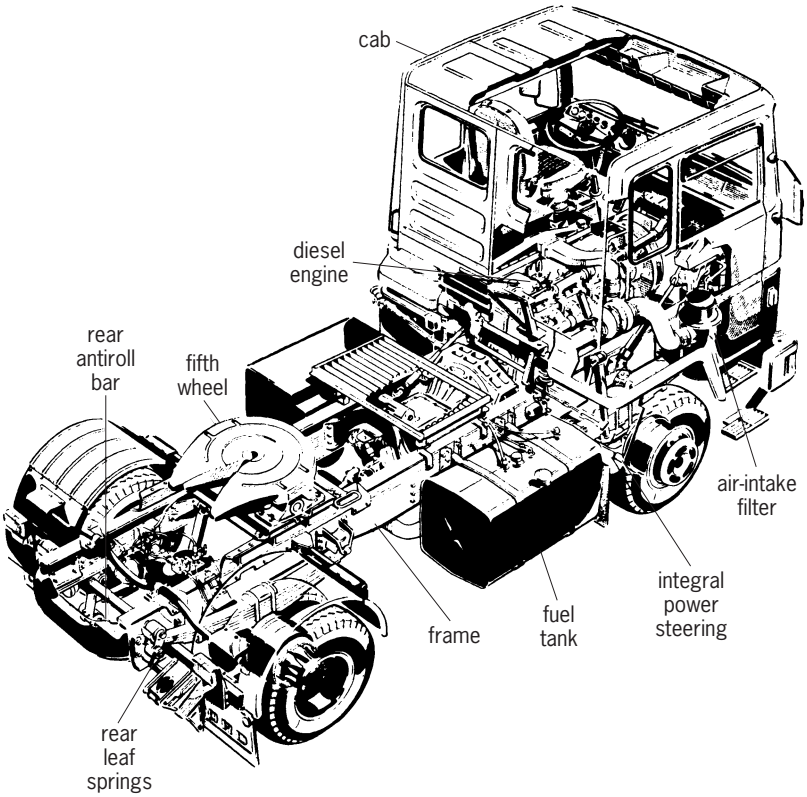


Fig. 2. Two-axle cab-over-engine truck tractor with fifth wheel mounted. (*Ford Motor Co.*)

and chassis. The load-bearing ability of a truck is determined by the strength of the frame, which is designed to handle the side, torsional, and vertical loads encountered in its load-rating category.

Frames are made of steel or aluminum alloy. Typical construction is of channel side rails held in place by cross-members, which resist buckling and frame twisting. In this ladder-type frame, the cross-members are usually riveted or welded to the side rails. Some cross-members are bolted in place to provide accessibility for service operations such as transmission removal. Most truck frames have a standard width of 34 in. (850 mm) between rails so specialty bodies from various body builders can be mounted. Integral body-and-frame construction is used in some light-duty trucks.

Engine. Trucks and truck tractors use inline, V-type, or pancake engines, which usually have 4, 6, 8, 10, or 12 cylinders, and use gasoline, compressed natural gas, liquid petroleum gas, or diesel fuel. The gasoline engines operate on the four-stroke cycle. The diesel engines are either two- or four-stroke. Electronic engine controls are used to help meet exhaust emission standards, reduce exhaust smoke, and improve fuel economy. *See* DIESEL FUEL; GASOLINE; LIQUEFIED NATURAL GAS (LNG); LIQUEFIED PETROLEUM GAS (LPG).

Most truck engines are liquid-cooled, and some are air-cooled. Brake horsepower for on-highway vehicles ranges from about 64 hp at 5500 revolutions/min for a gasoline engine in a light truck to 565 hp at 2000 revolutions/min for a diesel engine in a heavy-duty truck or tractor. Supercharging and turbocharging are used to develop more power from an engine of given size. *See* AUTOMOTIVE ENGINE; DIESEL ENGINE; ENGINE; INTERNAL COMBUSTION ENGINE; SUPERCHARGER; TURBOCHARGER.

Power train. The group of components that transmits power from the engine to the wheels is the drive train, or power train (**Fig. 3**). It includes the clutch, transmission, universal joints, drive shafts, and powered or drive axles. The clutch provides a means by which the driver can engage and disengage the transmission from the engine. A dry single-plate or two-plate friction clutch is normally used, with medium and heavy vehicles in severe start-stop operation having a wet clutch with two or more plates running in oil. *See* CLUTCH; UNIVERSAL JOINT.

Transmission. Manual, semiautomatic, and automatic transmissions are used in trucks and truck tractors. Because of the limited operating-speed range of most diesel engines, a relatively large number of forward gear ratios is required, especially in heavy-duty trucks and tractors. These may be provided, for example, by mounting a two-speed auxiliary transmission behind a five-speed manual transmission. This combination gives 10 forward speeds. When a two-speed rear drive axle is used, there are 20 possible forward speeds. *See* GEAR; GEAR TRAIN; PLANETARY GEAR TRAIN.

To reduce transmission length and weight, multiple countershaft designs are employed in some larger truck transmissions (**Fig. 4**). Two countershafts are spaced 180° apart on opposite sides of the mainshaft, or three countershafts are equally spaced around the mainshaft. Operation and function of the transmission are the same as with a single countershaft.

Two types of semiautomatic transmission are used, a straight mechanical one and a mechanical transmission with hydraulic torque converter. The semiautomatic allows the driver to choose any available gear ratio, while power shifting spares the driver the labor of gear changing. Declutching devices, which permit all forward shifts to be made without depressing the clutch pedal, also help the driver. With these, the clutch pedal is used only for standing starts. The automatics are similar to automobile automatic transmissions, with hydraulic torque converter and planetary gear trains providing four to six forward speeds.

Transfer case. Many transfer cases have two gear ratios and function as a combination two-speed auxiliary transmission and power divider. Transfer cases are used in military and other vehicles expected to engage in off-highway operation, and in on-highway vehicles requiring four-wheel or all-wheel drive. *See* AUTOMOTIVE TRANSMISSION.

Axle. Positioned transversely under the frame, an axle is a supporting member carrying the weight of a vehicle and its payload, and has mounted at either end the wheels on which the vehicle rolls. Drive axles transmit power from an input shaft to the wheels, forcing them to rotate. Nondriving or dead axles do not power the wheels but merely allow them to rotate freely. A steering function may be provided on either type by including means to pivot the wheels. Highway truck front axles (Fig. 3) are

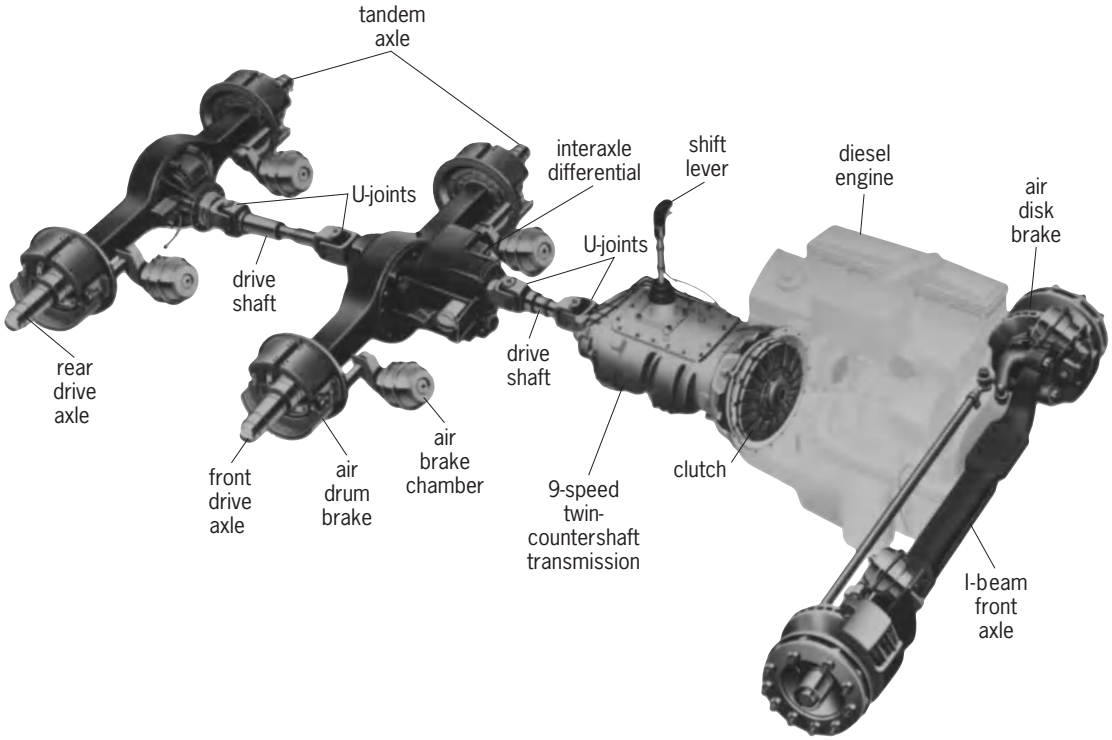


Fig. 3. Power train for an on-highway truck or truck tractor, with a nondriving front axle and a tandem axle at the rear. Both rear axles are driven. (*Rockwell International Corp.*)

typically the nondriving steer type, using a forged steel I-beam between pivot centers. Front axles for off-highway trucks are often steerable drive axles.

Drive axle. The truck or tractor may have a single driving axle at the rear for lighter loads, relatively short distances, and tight maneuvering. For heavier hauling, two rear axles (a tandem axle) or three rear axles (a triple or triaxle arrangement) may be used. In a tandem axle, the drive may be through one or both axles. When additional traction is needed, the vehicle may have all-wheel drive, in which the front axle and one or more rear axles are powered.

In a drive axle, power is transmitted from an input shaft at the center of the axle to a primary right-angle gear reduction, then to a differential mechanism integral with the gear reduction and through connecting axle shafts to the wheels. Drive axles may include a driver-controlled locking differential for additional traction during vehicle operation in mud, ice, and snow. Axles for off-highway trucks usually include additional reduction obtained with a planetary gear set at each wheel. Very large mining trucks may use a propulsion system adapted from the diesel-electric locomotive, with traction motors and planetary gear reduction at the driven wheels. *See* AUTOMOTIVE DRIVE AXLE; DIFFERENTIAL; LOCOMOTIVE.

Two-speed drive axles include an alternate second gear-reduction set, or power-flow path if of planetary design, thereby permitting operation in two speed ranges. Tandem- or triple-axle arrangements are powered by extending the input shaft through the first driving axle and coupling to the second (Fig. 3), and in like fashion from second to third driving axles.

An interaxle differential (Fig. 3) is used in tandem and triple arrangements to avoid internal torque buildup that could result from operation at slightly differing axle speeds, such as occurs with unmatched tires. Trucks used in off-highway service may operate satisfactorily without an interaxle differential, especially if operating in poor traction conditions. Axle lubricants are usually SAE-90 mineral oil modified with extreme-pressure additives because of

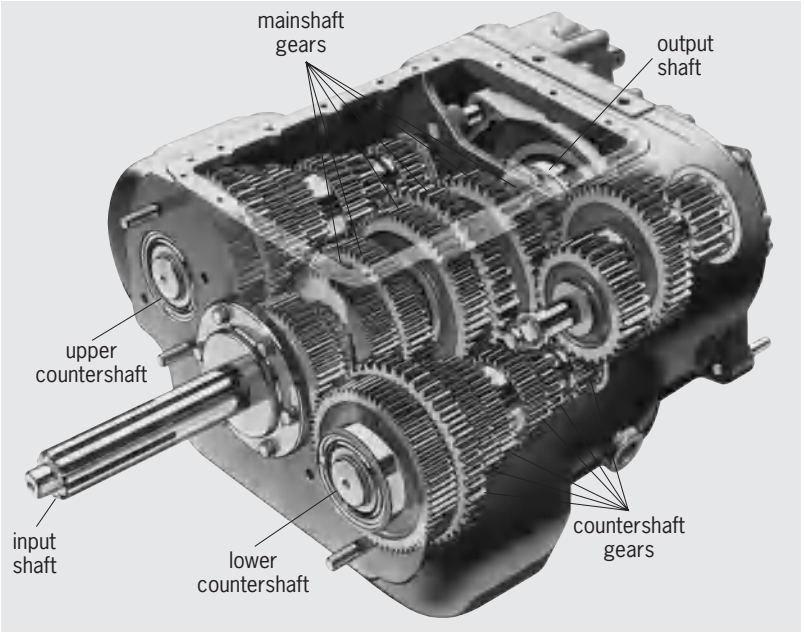


Fig. 4. Nine-speed twin-countershaft manual transmission. (*Rockwell International Corp.*)

the high specific loading and sliding tooth contact inherent in hypoid gear sets. *See* LUBRICANT.

Suspension. The suspension mounts the axle to the truck frame and reduces the shock and vibration transmitted to it. Leaf springs are widely used in truck suspension (Fig. 2). A front axle may be an I-beam suspended by leaf springs, or an independent front suspension using twin I-beams or control arms with coil springs. A variety of rear suspension systems are in use, including leaf, coil, rubber, and air springs. Air springs are employed especially when constant frame height and axle articulation, regardless of load, are important considerations. On tandem and triple-axle arrangements, the suspension includes the necessary beams and torque arms to distribute load to each axle. *See* AUTOMOTIVE SUSPENSION.

Steering. Trucks use Ackerman steering geometry, usually with a recirculating-ball steering gear. Power assist generally is provided by a hydraulic power-steering system or sometimes by air pressure in heavy-duty vehicles. Typical turning angle on nondrive steering axles is approximately 40°. Some heavy-duty trucks have a tandem front axle in which all four front wheels swing in and out for steering. For improved maneuverability, the rear axle on some trucks and trailers, such as fire and rescue vehicles, may also be a steering axle. *See* AUTOMOTIVE STEERING.

Brakes. Wheel brakes, commonly called service brakes or foundation brakes, and usually emergency or parking brakes are drum or disc brakes integral with or mounted on the axle (Fig. 3). Hydraulic brakes are used on light and medium-duty trucks, frequently with a hydraulic or vacuum brake booster for power assist. In heavy-duty trucks, air systems prevail because of their adaptability for tractor-trailer systems and their ability to operate at high temperatures that are generated in the brakes during severe service. Slack adjusters at each wheel brake automatically compensate for lining wear or are manually adjusted. Many newer trucks have an antilock braking system. *See* AIR BRAKE; AUTOMOTIVE BRAKE; BRAKE.

Truck brakes are not designed for continuous application. To prevent brake fade and possible failure during prolonged downhill braking, many heavy-duty trucks have an additional wear-free braking system called a retarder. It operates independently of the wheel brakes and may be a primary retarder located between the engine and transmission, or a secondary retarder located in the driveline between the transmission and the drive axle. Both types act to supplement the wheel brakes in dissipating the kinetic energy of the vehicle's motion.

Wheels and tires. Various combinations of wheels and tires are available for most trucks. Wheels and tires should be large enough for proper handling of the loads encountered. The total weight carried should not exceed the maximum rating of the tire and wheel. Trucks usually have steel or aluminum disk wheels or cast spoke wheels. Requirements for truck wheels include high fatigue strength and service life; lowest possible weight to increase payload; minimal unevenness, runout, and imbalance; and ease of assembly for tire mounting. A common truck and tractor wheel is a one-piece disk wheel with a rim having a 15° tapered-bead seat for use with a tubeless tire.

Tubeless radial tires are widely used to minimize rolling resistance and to improve fuel economy. Truck tires may be recapped to prolong carcass life. *See* TIRE.

Advances in truck technology. In light trucks, such as pickups, vans, and sport utility vehicles, many drivers expect the same comfort, convenience, and entertainment features as in an automobile, and most of these accessories are offered as factory- or dealer-installed options. In trucks of all sizes, electronics are incorporated into almost every system and component, including the transmission and brakes that previously could be controlled only by the driver.

To reduce engine fuel consumption and exhaust emissions, electronic devices and computer controls provide engine management and onboard diagnostics (**Fig. 5**). Rising fuel prices and the possibility of future fuel shortages are encouraging truck product development. Advances in truck technology include clean diesel technology, hybrid power sources, alternative fuels, lighter and stronger materials, and other methods of improving fuel efficiency, operation, and safety. *See* ELECTRIC VEHICLE.

Clean diesel technology. Most light-duty trucks have gasoline-fueled spark-ignition engines, although diesel engines in light-duty trucks have become more popular in recent years. Many medium-duty trucks and most heavy-duty trucks for highway use have diesel engines. However, diesel exhaust emissions of hydrocarbons (HC), carbon monoxide (CO), nitrogen oxides (NO_x), and soot or particulate matter (PM) are considered harmful to the

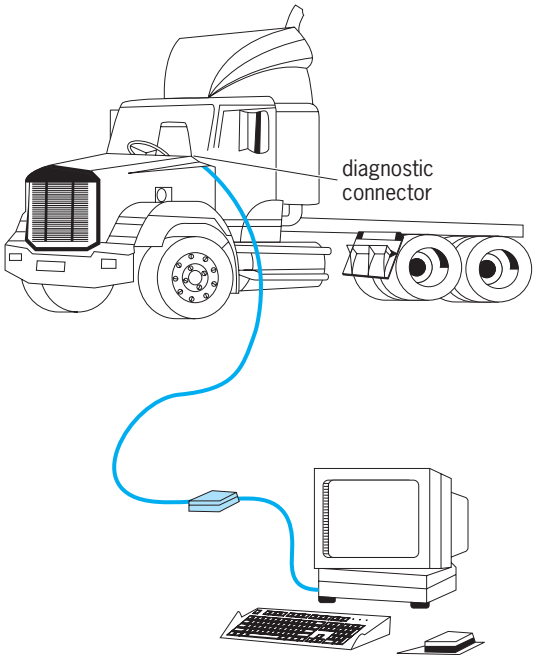


Fig. 5. Accessing the stored diagnostic trouble codes and other electronic data through the diagnostic connector in the cab of a heavy-duty truck.

environment and human health. In the presence of sunlight, nitrogen oxides react with hydrocarbons or carbon monoxide in the air to form ozone, smog, and acid rain. Small particles of particulate matter, if inhaled, may cause breathing and other health problems. *See* AIR POLLUTION.

The U.S. Environmental Protection Agency (EPA) issued new standards for the 2007 model year, requiring the petroleum industry to produce ultra-low-sulfur diesel fuel which is cleaner-burning and contains a maximum of 15 parts per million (ppm) sulfur by weight. This has been estimated as equivalent to one ounce of sulfur in a tank truck of diesel fuel. The new highway ultra-low-sulfur diesel fuel replaces most low-sulfur diesel fuel that contains up to a maximum of 500 ppm sulfur. By 2010, ultra-low-sulfur diesel fuel must also be used for off-highway vehicles.

Reducing the sulfur content of diesel fuel allows aftertreatment of the exhaust gas with particulate traps, which can be clogged by higher levels of sulfur. The combining of the three elements of (1) ultra-low-sulfur diesel fuel, (2) modifications of diesel engines to lower emissions of nitrogen oxides, and (3) advanced emission controls has gained the name "clean diesel technology." Application of clean diesel technology should result in diesel truck exhaust that has no smoke, no smell, and virtually no emissions.

Donald L. Anglin

Bibliography. *Bosch Automotive Handbook*, 2004; M. Ehsani et al., *Modern Electric, Hybrid Electric, and Fuel Cell Vehicles: Fundamentals, Theory, and Design*, CRC Press, 2005; Society of Automotive Engineers, *SAE Handbook*, 3 vols., annually; Society of Automotive Engineers, *Truck Systems Design Handbook*, vol. 2, 2002.

Truss

An assemblage of structural members joined at their ends to form a stable structural assembly. If all members lie in one plane, the truss is called a planar truss or a plane truss. If the members are located in three dimensions, the truss is called a space truss.

A plane truss is used like a beam, particularly for bridge and roof construction. A plane truss can support only weight or loads contained in the same plane as that containing the truss. A space truss is used like a plate or slab, particularly for long span roofs where the plan shape is square or rectangular, and is most efficient when the aspect ratio (the ratio of the length and width) does not vary above 1.5. A space truss can support weight and loads in any direction.

Because a truss can be made deeper than a beam with solid web and yet not weigh more, it is more economical for long spans and heavy loads, even though it costs more to fabricate. *See* BRIDGE; ROOF CONSTRUCTION.

The simplest truss is a triangle composed of three bars with ends pinned together. If small changes in the lengths of the bars are neglected, the relative positions of the joints do not change when loads are applied in the plane of the triangle at the apexes.

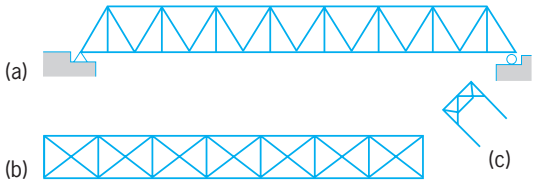


Fig. 1. Loads in through trusses are borne on the lower chord. (*a*) Through Warren truss with verticals. (*b*) Top-chord bracing as seen from above. (c) End-on view showing portal bracing.

Planar trusses. Such simple trusses as a triangle, perhaps with the addition of a vertical bar in the middle, are sometimes used to support peaked roofs of houses and other narrow structures. For longer spans, flat roofs, or bridges many triangles are combined to form a truss.

In metal trusses, connections may be riveted, bolted, welded, or pinned; in wood trusses, they may be bolted, nailed, or glued. Because of long spans, provision must be made to permit movement at one support due to loads and temperature changes; rollers, rockers, or sliding plates generally are used for this purpose.

The top members of a truss are called the upper chord; the bottom members, the lower chord; and the verticals and diagonals, the web members.

Framing to be carried by a truss usually is arranged so that it brings loads to bear on the truss at the intersections of a chord and web members. As a result, truss members are subjected only to direct axial stress—tension or compression—and can be made of less material than if they also had to resist bending stresses.

Roof trusses carry the weight of roof deck and framing and wind loads on the upper chord. They may also support a ceiling or other loads on the lower chord. On the other hand, bridge trusses may carry loads on either chord. Deck trusses support loads on the upper chord; through trusses, on the lower chord (**Fig. 1**).

To maintain stability of truss construction, bracing must be used normal to the planes of the trusses. Usually framing is inserted between the trusses. For roofs, trussed bracing should be placed in the plane of either the top chord or the bottom chord. For bridges, bracing must be inserted in the planes of both top and bottom chords, because of the greater need for stability under heavy moving loads. Vertical trusses are known as braced towers. Vertical trusses are used extensively in buildings to resist wind and earthquake loads.

Space trusses. Space trusses are essentially composed of two sets of plane trusses, with the top and bottom chords intersecting the edges at 90° (orthogonal; **Fig. 2**) or three sets of plane trusses with top and bottom chords intersecting the edges at acute angles (skewed; **Fig. 3**). Most space trusses have the individual plane trusses lying in an inclined plane so that the entire system is inherently stable. The inclination of the plane trusses results in

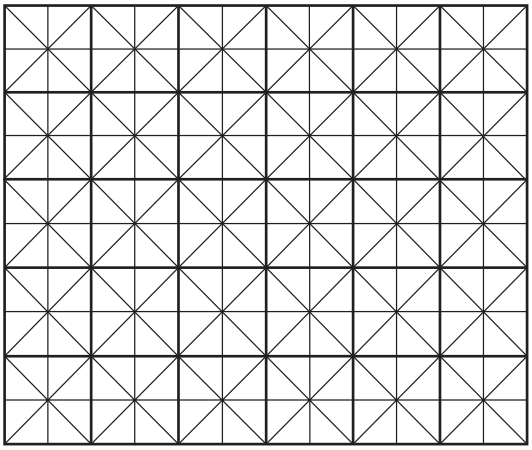


Fig. 2. Orthogonal truss.

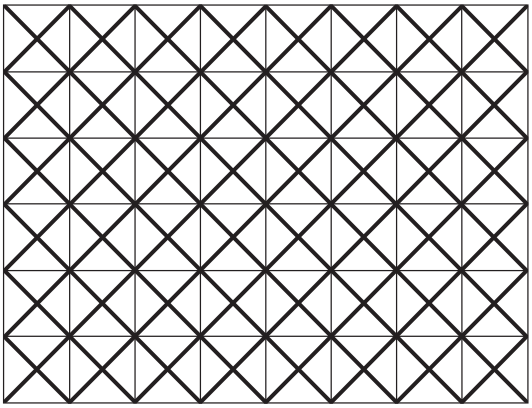


Fig. 3. Skewed space truss.

the basic element of the space truss being a pyramid (a tetrahedron). As a result, the top and bottom chord are offset one-half module. Curved space trusses are used to span long distances as part of an overall structural shape, in particular, a dome. One such space truss dome is the geodesic dome. *See* GEODESIC DOME.

Computing stresses. Primary axial stresses in truss members are computed on the assumption that connections at joints are made with frictionless pins. With loads applied at joints, each truss member or bar is subjected to pure tension or compression.

Since the bars change length under load, the angles of each triangle constituting the truss tend to change. But this change is resisted, since pins are not frictionless, and since rivets, bolts, or welds offer restraint. Consequently, members bend slightly, the bending moments creating secondary stresses.

Planar trusses of single-span configuration lend themselves to simple statical hand calculations. At a truss joint, the primary stresses and loads form a coplanar, concurrent force system in equilibrium. This force system satisfies two conditions: the sums of the horizontal and vertical components both equal zero. These equations are used in computing stresses by the method of joints: joints with two unknowns are selected in succession and the two equilibrium equations are applied to them to determine the stresses.

A section may be passed through the truss to cut three bars with unknown stresses. These, together with bars with known stresses that are cut and the loads on the part of the structure on either side of the section, constitute a coplanar, noncurrent force system in equilibrium. This system satisfies the two previous conditions; but in addition, the sum of the moments of the forces about any axis normal to the plane equals zero. With these three equations the three unknowns can be determined. However, the unknowns also can be found by the method of moments, in which two of the unknowns are eliminated by taking the moment axis at their point of intersection, and the third is found by equating the sum of the moments to zero. The method of shears is used to determine one force when the other two unknown forces are both normal to the shearing force, for example, for finding the stresses in the diagonals of parallel-chord Warren, Pratt, and Howe trusses.

If n is the number of joints in a truss, stresses can be found by the methods of joints, moments, or shears when the number of bars equals $2n - 3$. If a truss is composed of fewer bars, it is unstable; if of more bars, statically indeterminate.

Influence lines are useful in determining the stresses in bridge trusses, because the live load is a moving load. Influence lines can be drawn to show the variation in any function—stress, shear, moment, deflection—as a unit load moves along the truss.

Multiple-span plane trusses (defined as statically indeterminate or redundant) and space trusses require very complex and tedious hand calculations. Modern high-speed digital computers and readily available computer programs greatly facilitate the structural analysis and design of these structures. *See* COMPUTER; STRUCTURAL ANALYSIS.

Charles Thornton; I. Paul Lew

Trypanites

A simple cylindrical boring common in the fossil record for the past 540 million years. *Trypanites* is an ichnogenus, which is a formal name given to a distinct trace fossil (evidence of organism behavior in the fossil record as distinct from the remains or other representations of the organism's body). *Trypanites* is thus not a type of life but a structure built by a variety of wormlike animals over time. *See* TRACE FOSSILS.

Trypanites is formally defined as a cylindrical, unbranched boring in a hard substrate (such as a rock or shell) with a length up to 50 times its width (see **illustration**). Most *Trypanites* are only a few millimeters long, but some are known to be up to 12 cm (4.72 in.). They usually penetrate the substrate perpendicularly to its surface and remain straight, but they are sometimes found at oblique angles, and a

Trypanites **borings in a carbonate hardground from the Upper Ordovician of northern Kentucky. The borings descend from the top of the rock surface perpendicularly to the bedding. Each boring is filled with tiny dolomite crystals, making them stand out against the dark limestone. The longest boring is 4.7 cm (1.85 in.).**

few may curve in response to irregularities in the material they bored. Almost all *Trypanites* are found in calcareous substrates, which is evidence that the producing organisms may have used at least in part some sort of acid or other dissolving chemical to excavate the substrate. *Trypanites* is often confused with the similar boring *Palaeosabella*, which has a clavate (expanded) terminus rather than a simple rounded one. Since *Trypanites* have a circular cross section, they can sometimes be confused with predatory borings (*Oichnus*) when they pass through shells which are later liberated by erosion from their lithological (rock) context.

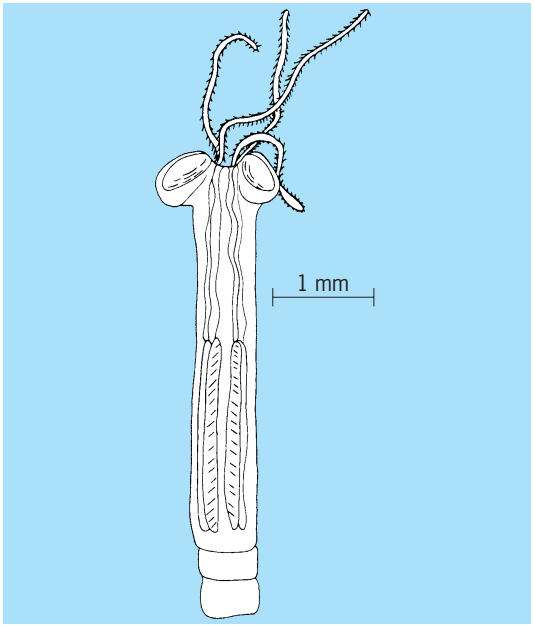
The organisms which produced *Trypanites* have varied over geological time. Most *Trypanites* were made by marine wormlike animals, with a few known from freshwater paleoenvironments. Very rarely remains are found in the fossil borings which can be attributed to particular groups, such as polychaetes. Similar borings are made today by sabellarid polychaete and sipunculid worms, among other groups. These borings are made as dwellings (domichnia) for filter-feeding the waters above, so they have nothing to do with predation or digestion of the substrate.

Trypanites is important to paleontologists as one of the earliest examples of bioerosion (the erosion of hard substrates by biological actions). It is the first known macroboring (a boring easily visible to the naked eye) and is found in Lower Cambrian hardgrounds about 540 million years old. *Trypanites* is surprisingly rare until the Middle and Late Ordovician, when it reappears in great abundance as part of the Ordovician Bioerosion Revolution. This was a time when many other macroborings enter the fossil record and bioerosion intensity greatly increased. *Trypanites* is the most common macroboring until the Jurassic (about 200 million years ago), when it was exceeded in abundance by the bivalve boring *Gastrochaenolites* and other more complex tubular borings. *Trypanites* is also one of the first borings to host nestling organisms, such as brachiopods, inside the empty cavities, thereby expanding ecological niche space on marine hard substrates. *See* ORDOVICIAN; PALEOECOLOGY. Mark A. Wilson

Bibliography. R. G. Bromley, On some ichnotaxa in hard substrates, with a redefinition of *Trypanites* Mägdefrau, *Paläontologische Zeitschrift*, 46:93–98, 1972; N. P. James, D. R. Kobluk, and S. G. Pemberton, The oldest macroborers: Lower Cambrian of Labrador, *Science*, 197:980–983, 1977; P. D. Taylor and M. A. Wilson, Palaeoecology and evolution of marine hard substrate communities, *Earth-Sci. Rev.*, 62:1–103, 2003; M. A. Wilson and T. J. Palmer, *Hardgrounds and Hardground Faunas*, University of Wales, Aberystwyth, Institute of Earth Studies Publ. 9, 1992.

Trypanorhyncha

An order of tapeworms of the subclass Cestoda, also known as the Tetrarhynchoidea. All are parasitic in the intestine of elasmobranch fishes. They are distinguished from all other tapeworm groups by having spiny, eversible proboscides on the head. An elongated head stalk contains the proboscis apparatus made up of a proboscis sheath and a muscular bulb. The head also bears two or four shallow, weakly muscular suckers (see **illus.**). Segment anatomy resembles that of Proteocephaloidea, except that the yolk glands are scattered. A complete life history is not known for any trypanorhynchid, although larval forms have been found in the tissues of various



Scolex of tapeworm *Eutetrarhynchus*.

marine invertebrates and teleost fishes. *See* EUCESTODA; PROTEOCEPHALOIDEA.

Clark P. Read

Trypanosomatidae

A family of Protozoa, order Kinetoplastida, containing flagellated parasites which change their morphology; that is, they exhibit polymorphism during their life cycles. The life cycles of the organisms may involve only an invertebrate host, or an invertebrate and a vertebrate host, or an invertebrate and a plant host. Several distinct morphological forms are recognized: trypanosomal, crithidial, leptomonad, and leishmanial. Differentiation into genera is dependent upon the host infected as well as the morphologic types involved. None of the stages possesses a mouth opening, and nutritive elements are absorbed through the surface of the body; that is, the organisms are saprozoic. **Figure 1** illustrates the morphologic stages and the hosts of the genera.

Morphology. All the stages possess a single nucleus and a kinetoplast composed of a rod-shaped parabasal body and a minute basal granule called the blepharoplast. In the elongated stages of the parasite, the kinetoplast may be located at the anterior end of the leptomonads, near the center in the crithidials, or at the posterior end in the trypanosomal forms. With the exception of the leishmanial stage, either an axoneme or fiberlike structure arises from the basal granule and extends immediately out of the body as a free flagellum in the leptomonad or along the edge of an undulating membrane in the crithidial and trypanosomal forms. The axoneme may terminate at the end of the undulating membrane or continue as a free flagellum. The organism moves in the direction of the protruding flagellum. The undulating membrane is a finlike structure composed of two folds of the outer pellicle of the body, with the axoneme supporting the outer edge.

Reproduction. Multiplication of the Trypanosomatidae is usually by longitudinal binary fission. For example, in the trypanosomal stage the process is as follows: The kinetoplast divides first and a second axoneme develops from the new basal granule; as the axoneme increases in length toward the anterior end, the nucleus divides; finally, when all the structures have been duplicated, the body itself splits longitudinally beginning at the anterior end (**Fig. 2**).

Taxonomy. The six generally recognized genera of the family Trypanosomatidae are *Trypanosoma*, *Leishmania*, *Leptomonas*, *Phytomonas*, *Crithidia*, and *Herpetomonas*. The first two, *Trypanosoma* and *Leishmania*, are of medical and veterinary importance. Since they are found in the blood of vertebrates, they are referred to as hemoflagellates. The other four, *Leptomonas*, *Phytomonas*, *Crithidia*, and *Herpetomonas*, occur only in invertebrates and plants.

Trypanosoma. This is the most important genus of the family Trypanosomatidae from a number of standpoints. It contains the largest number of species infecting a wide variety of hosts such as mammals, birds, fishes, amphibians, and reptiles. Although most of the species cause no damage to the hosts, there are several which produce serious diseases in humans, domesticated animals, and wild animals. The pathogenic species, prevalent in Africa, have

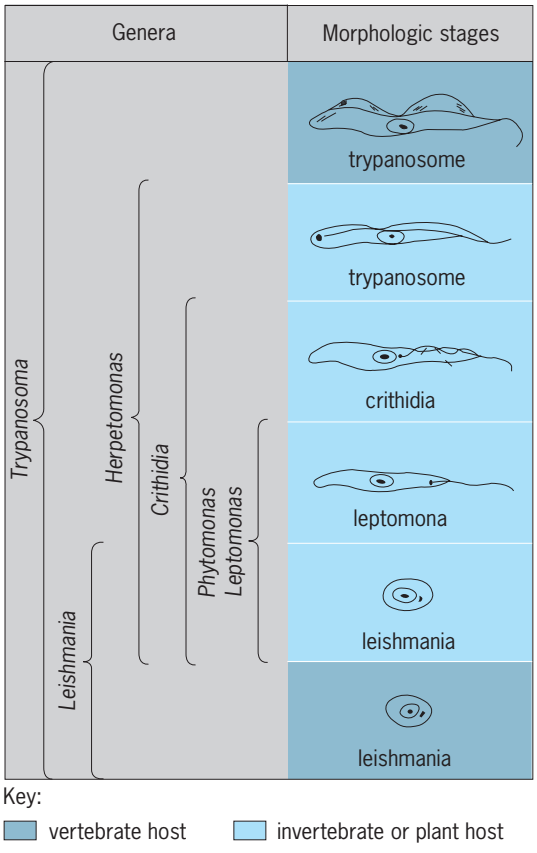


Fig. 1. Morphologic stages and the hosts of genera of Trypanosomatidae.

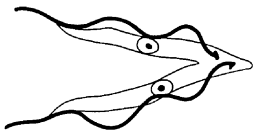


Fig. 2. A dividing trypanosome.

been responsible to a great extent for the slow development of civilization in many parts of that continent.

Developmental stages. Certain species of *Trypanosoma* possess all four developmental stages in their life cycles—trypanosomal, crithidial, leptomonad, and leishmanial forms; however, the trypanosomal stage is the most important, being found in the circulating blood of all of the vertebrate hosts. Differentiation of the species is based upon the morphology of the blood-form trypanosomes, the particular vertebrates and invertebrates serving as the hosts, the stages present in the hosts, and the course of development and location of the various stages in the hosts.

Morphology. The trypanosomal stages of the different species differ in size and shape, location of the nucleus, presence and position of the kinetoplast, and development of the undulating membrane. Trypanosomal forms range in length from less than 15 to over 80 micrometers, and may be either slender or broad. In general the nucleus is located near the center, but the kinetoplast may be at the posterior tip or some distance from it. The undulating membrane may be very prominent with many convolutions, or more or less flat and inconspicuous. There may or may not be a free flagellum at the anterior end.

Reproduction. Reproduction of the *Trypanosoma* may take place in several stages of the parasite and at different locations in the vertebrate and invertebrate hosts. Typically it is by longitudinal binary fission, but in some instances the division of the cytoplasm is delayed and multiple fission occurs.

Life history. With only rare exceptions, an invertebrate host such as a fly, bug, or leech is involved in the transmission of the *Trypanosoma* from vertebrate to vertebrate. In cyclical transmission it is necessary for the parasite to undergo developmental changes in the vector before infective, or metacyclic, trypanosomes occur. In the insect, these infective stages, depending upon the species, are located either in the salivary gland or proboscis and are transmitted by the bite, as in the case of *T. gambiense* and *T. rhodesiense*, or are located in the hindgut and are deposited on the surface of the vertebrate when the vector defecates, as in *T. cruzi*. In the latter case, the infective stage enters the body of the vertebrate host through a break in the skin, sometimes through the hole produced by the bite of the insect. Mechanical transmission can occur with the various species of *Trypanosoma* perhaps without exception, and with some it is the principal means of transmission. The trypanosomes thus obtained during a blood meal merely survive on the mouthparts of the insect and are introduced into another vertebrate when the insect feeds again within a few minutes. In *T. equiperdum* infections of horses and donkeys, contact transmission occurs during the sexual act and thus no intermediate invertebrate host is required in the life cycle.

Nonpathogenic species. Although the life cycles of the numerous species of *Trypanosoma* may differ in various respects, the one for *T. lewisi* serves to illustrate the complexity of the process for a nonpathogenic form. This parasite is worldwide in its distribution and is found in various species of rats. Under ordinary conditions it causes no apparent harm in the vertebrate host. The flea *Nosopsyllus fasciatus* transmits the organism from rat to rat. After the flea ingests the trypanosome stage in the rat's blood, a cyclical development of the parasite takes place in the intestinal tract of the invertebrate host. The long blood-form trypanosomes enter the epithelial cells lining the stomach, and reproduce by multiple fission. The modified trypanosomes which emerge from the cells migrate to the hindgut, where they attach to the lining by their anterior ends. They transform into the crithidial form and divide by binary fission. The infective, short metacyclic stages finally develop and accumulate in the rectum, from which they pass out with the feces. The rat becomes infected through ingestion of the flea's feces contaminating its body or by ingestion of the entire infected flea. For 8 or more days after reaching the bloodstream of the rat, the parasites reproduce by multiple fission in the crithidial stage and in other bizarre shapes. During this process, the individual organism divides several times without complete fission of the cytoplasm. Finally, the progeny break away from one another and develop into separate trypanosomes. At the end of this reproductive period, only typical trypanosomal forms remain. In a month or more they are destroyed by the immune response of the host.

Other interesting nonpathogenic species are *T. rotatorium* in the frog, which is transmitted by leeches; *T. granulosum* in the eel; and *T. danilewskyi* in goldfish.

Information on a number of the important pathogenic *Trypanosoma* is presented in the **table**.

Pathogenic species. Some authorities believe that the species which cause sleeping sickness, *T. gambiense* and *T. rhodesiense*, represent the wild animal species, *T. brucei*, which have become adapted to the human body; ordinarily the serum of humans is trypanocidal for *T. brucei*. In any event, the species are very similar morphologically and exhibit comparable development in the tsetse fly, *Glossina*. After the African sleeping sickness parasites enter the human body during the bite of the tsetse fly, they first multiply in the lymph and blood. Later they may invade the nervous system. Only the trypanosome stage occurs in humans. In the tsetse flies *G. palpalis* and *G. morsitans*, multiplication takes place first in the midgut while in the trypanosome stage. Later the parasites migrate to the salivary glands, where they transform into crithidial forms and multiply. Eventually metacyclic trypanosomes develop in this site. *See* DIPTERA.

Trypanosoma cruzi, the cause of Chagas' disease,

Epidemiology of *Trypanosoma* species

Species of *Trypanosoma*	Principal vertebrate hosts	Disease	Geographical distribution	Insect vectors	Mode of transmission
T. gambiense	Humans and domestic animals	Sleeping sickness	West Equatorial Africa	Tsetse flies (*Glossina*)	Cyclical; insect bite
T. rhodesiense	Probably humans and wild animals	Sleeping sickness	East Tropical Africa	Tsetse flies (*Glossina*)	Cyclical; insect bite
T. cruzi	Human, dog, armadillo, opossum, and other animals	Chagas' disease	South and Central America	Kissing bugs (*Triatoma*)	Cyclical; feces of bug
T. brucei	Domestic and wild mammals	Nagana	Tropical Africa	Tsetse flies (*Glossina*)	Cyclical insect bite
T. vivax	Domestic and wild mammals	Souma	Tropical Africa and South America	Tsetse flies (*Glossina*); stable flies (*Stomoxys*)	Cyclical (*Glossina*) and mechanical (*Stomoxys*); insect bite
T. equinum	Domestic and wild mammals	Mal de caderas	Tropical and South America	Biting flies (*Tabanus, Stomoxys*)	Mechanical; insect bite
T. evansi	Domestic and wild mammals	Surra	Asia, Australia, Madagascar	Biting flies (*Tabanus, Stomoxys*)	Mechanical; insect bite
T. hippicum	Domestic mammals, especially horses and mules	Murrina de caderas	Central America	Nonbiting flies (*Musca*)	Mechanical; by flies
T. equiperdum	Horses and donkeys	Dourine	Mediterranean countries	Usually none	Contamination; sexual act

presents a somewhat different cycle in humans and triatomid bugs. These rather large arthropods, about 1 in. (2.5 cm) in length, frequently feed around the face of the sleeping person and therefore are commonly referred to as kissing or barber bugs. In the human body the trypanosomal stages circulating in the bloodstream do not divide. All multiplication takes place in tissue cells. The parasites enter various tissues including the heart muscles, transform into leishmanial stages, and divide by binary fission until a cluster of parasites fills each cell. Before breaking out of the cells into the bloodstream, they return to the trypanosomal stage. The parasite may reenter other cells and multiply again. The trypanosome, after its ingestion by the kissing bug, multiplies first in the midgut and later in the rectum. Multiplication in the bug is primarily during the crithidial stage. Metacyclic trypanosomes develop in the rectum and pass out of the bug when it defecates. The infective stages therefore are not inoculated into the vertebrate but gain entrance into the body through contamination or breaks in the skin or through the conjunctiva of the eye. *See* HEMIPTERA.

Although *T. cruzi* is apparently present in triatomid bugs and lower animals such as the raccoon, skunk, opossum, and armadillo in various parts of the southern half of the United States, the first proved human infection in this country was not reported until 1955.

Another species, *T. rangeli*, has been recognized as a parasite of humans. In South and Central America this apparently nonpathogenic species must be differentiated from *T. cruzi* in examining the intestinal contents of triatomid bugs and the blood of humans and other vertebrates.

Leishmania. This is the second most important genus, at least from humanity's standpoint. Species of this genus occur as typical leishmanial forms in vertebrate hosts and as leptomonad stages in invertebrate hosts. Three species parasitize humans but have also been found naturally infecting dogs, cats, and perhaps other lower animals. The sand fly, *Phlebotomus*, transmits the parasite from vertebrate to vertebrate. After ingestion the leishmanial stage transforms into the leptomonad form in the gut of the fly and multiplies. The leptomonads are the infective stage for humans and are introduced by the bite of the infected fly. They enter various cells of the body such as the skin, capillaries, spleen, and liver, transform to the leishmanial stage, and multiply.

The three species of *Leishmania* are morphologically identical. Their small, oval bodies, about 5 μm in length, have a relatively large nucleus and kinetoplast, but no flagellum. The species are distinguished by their geographical distribution, the tissues they infect, and their immune reactions. They all produce serious diseases in humans which are difficult to control and treat. *Leishmania donovani* infects primarily the internal organs, causing kala azar or visceral leishmaniasis. *Leishmania tropica* is limited to the surface of the body, producing skin lesions (oriental sore) or cutaneous leishmaniasis. *Leishmania brasiliensis* is also limited to the surface of the body, but produces skin lesions (espundia, forest yaws) which frequently involve the mucous membranes of the nose, mouth, and pharynx; this is mucocutaneous leishmaniasis.

Leptomonas. In this genus, the kinetoplast is situated near the anterior end of the elongated body. The axoneme arising from the blepharoplast extends

directly out of the body as a free flagellum. Leptomonads are exclusively parasitic in invertebrates; for example, they occur in the hindgut of the common dog tick, *Ctenocephalus canis*. The nonflagellated leishmanial form is the infective stage.

Phytomonas. Morphologically similar to *Leptomonas*, phytomonads infect the latex of certain plants, for example, milkweed. Multiplication of the flagellates may cause degeneration of the infected part of the plant. Hemipterous insects which feed on latex transmit the parasite from plant to plant.

Crithidia. Species of this genus are parasitic in various arthropods. The kinetoplast is anterior to the centrally placed nucleus. There are generally a short undulating membrane and a free anterior flagellum. Depending upon the species, the organism may be found in the intestinal tract or body cavity of invertebrates including water bugs and ticks. Leptomonad and leishmanial stages develop, and encysted forms may serve as the infective stage.

Herpetomonas. All four stages, trypanosome, crithidial, leptomonas, and leishmanial, occur in the life cycle of species in this genus. The several species are exclusively parasitic in invertebrates. They infect the intestinal tracts of various species of flies. Transmission is by ingestion of encysted forms.

M. M. Brooke; Horace W. Stunkard

Trypanosomiasis

A potentially fatal infection caused by parasites of the genus *Trypanosoma*.

African trypanosomiasis. The African trypanosomes, the cause of African trypanosomiasis or African sleeping sickness, are flagellated protozoan parasites. They are members of the *T. brucei* group. *Trypanosoma brucei rhodesiense* and *T. b. gambiense* cause disease in humans. *Trypanosoma brucei* is restricted to domestic and wild animals. The trypanosomes are transmitted by the tsetse fly (*Glossina*), which is restricted to the African continent. The trypanosomes are taken up in a blood meal and grow and multiply within the tsetse gut. After 2–3 weeks, depending upon environmental conditions they migrate into the salivary glands, where they become mature infective forms, and are then transmitted by the injection of infected saliva into a new host during a blood meal. The survival of the tsetse fly is dependent upon both temperature and humidity, and the fly is confined by the Sahara to the north and by the colder drier areas to the south, an area approximately the size of the United States. Approximately 50 million people live within this endemic area, and 15,000–20,000 new human cases of African trypanosomiasis are reported annually.

In humans and other mammals the trypanosomes are extracellular. During the early stages of infection, the trypanosomes are found in the blood and lymph but not in cerebrospinal fluid. There is fever, malaise, and enlarged lymph nodes. In the absence of treatment the disease becomes chronic and the trypanosomes penetrate into the cerebrospinal fluid and the brain. The symptoms are headaches, behavioral changes, and finally the characteristic sleeping stage. Without treatment the individual sleeps more and more and finally enters a comatose stage which leads to death. Treatment is more difficult if the infection is not diagnosed until the late neurological stage. The two species *T. b. rhodesiense* and *T. b. gambiense* are found in different geographical areas, have different vectors and different reservoir hosts, and produce different types of disease (see **table**).

The epidemiology of African trypanosomiasis is influenced by the environment, the abundance and type of vector, the type and availability of reservoir hosts, the type of human activities, and numbers and activities of domestic livestock. Control depends on a detailed knowledge of these factors and includes avoiding areas where there is a high density of tsetse flies; altering the environment by brush clearing; removal of reservoir hosts; use of insect traps and insecticides; and drug treatment of humans and domestic animals. By use of these protocols in an environmentally sound manner, African trypanosomiasis

Comparison of the biology of the human infective *Trypanosoma brucei gambiense* and *T. b. rhodesiense*

Characteristics	*T. b. gambiense*	*T. b. rhodesiense*
Vector	*Glossina palpalis* group, riverine tsetse	*G. morsitans* group, savana, woodland tsetse
Method of transmission	Human → Tsetse → Human ↓ Animal reservoirs	Animal reservoirs → Tsetse → Animal reservoirs ↓ Human
Resevoir hosts	Possibly kob, hartebeest, domestic pigs, dogs	Bushbuck, waterbuck, hartebeest, hyena, lion, domestic cattle, possibly warthog and giraffe
Geographical area	West and North-Central Africa	Central and East Africa
Disease	Chronic, low parasitemia, incubation period of months to years	Acute, high parasitemia; incubation period of days to weeks

can be controlled but not eradicated. *See* MEDICAL PARASITOLOGY.

Chagas' disease. *Trypanosoma cruzi*, the cause of American trypanosomiasis or Chagas' disease, is transmitted by biting insects of the Reduviidae family (subfamily Triatominae). The parasite is distributed throughout most of South and Central America, where it infects an estimated 12–20 million people and over 35 million individuals are exposed to the parasite. Over 100 different species of wild and domestic animals have also been found to be infected. Animals that live in proximity to humans such as dogs, opossums, and wood rats are important reservoir hosts. Infected animals have been detected in the southwestern United States. However, only a few infections have ever been diagnosed in individuals from the United States who have not traveled to Central or South America.

Trypanosoma cruzi is predominantly an intracellular parasite in the mammalian host. During the intracellular stage, *T. cruzi* loses its flagellum and grows predominantly in cells of the spleen, liver, lymphatic system, and cardiac, smooth, and skeletal muscle. The cells of the autonomic nervous system are also frequently invaded. The parasites grow rapidly, forming closely packed pockets of parasites called pseudocysts. Rupture of the infected cells releases these intracellular stages into the surrounding tissue spaces, where they become flagellated and invade the blood of the host. These flagellated forms are taken up in the reduviid blood meal. The trypanosomes migrate into the insect gut, where they multiply and finally are excreted.

The disease has an acute and a chronic stage. The acute stage is initiated by scratching the trypanosomes that are in insect feces into the wound produced by the insect bite. Thus, individuals are infected by contamination following the bug bite rather than by direct inoculation. The acute phase is characterized by local inflammation at the site of the bite and by fever, muscle and bone pain, headaches, and anemia. However, the symptoms are extremely varied, and a clear diagnosis is difficult to obtain. The acute symptoms disappear spontaneously within several months in 90–95% of the cases. In many individuals the infection remains silent, and death results from causes unrelated to *T. cruzi*. However, in some individuals, the disease becomes chronic, and over a period of years serious symptoms appear. One sympton is chronic myocarditis, an inflammation of the heart muscles. In some areas, Chagas' disease accounts for a majority of cardiac deaths in young adults. In other geographical regions a second condition is observed and is referred to as megaesophagus or megacolon. It is caused when *T. cruzi* invades and destroys the autonomic ganglia of the esophagus or colon. This leads to enlargement of the colon or esophagus, and in some cases death. Although there is only one known species of *T. cruzi*, different isolates appear to have differences in virulence. Possibly this explains the different symptoms observed in different geographical regions.

In villages located in endemic areas, the insect vector lives within thatched roofs, cracked walls, or trash-filled rooms, and infection usually occurs at night while the individual is asleep. In addition, individuals become infected when villagers go into forest areas in which reservoir hosts and vectors are present. Transmission also occurs by the tranfusion of blood donated by infected individuals.

There is no satisfactory chemotherapeutic agent for treatment, and a vaccine is unavailable. Therefore the methods for control of Chagas' disease are directed toward the vector. For example, houses are treated with insecticides, cracks within walls are filled, and metal roofs (replacing thatched roofs) are used to eliminate vector breeding sites. These techniques are effective in reducing the incidence of Chagas' disease. Unfortunately, they are all costly and depend upon improved socioeconomic conditions in endemic areas. *See* PARASITOLOGY.

John Richard Seed

Bibliography. J. R. Baker (ed.), *Parasitic Protozoa*, vol. 2, 2d ed., 1992; L. S. Roberts, *Foundations of Parasitology*, 5th ed., 1995.

Tsunami

A set of ocean waves caused by any large, abrupt disturbance of the sea surface. If the disturbance is close to the coastline, a local tsunami can demolish coastal communities within minutes. A very large disturbance can cause local devastation and destruction thousands of miles away. Tsunami comes from the Japanese language, meaning harbor wave.

Tsunamis rank high on the scale of natural disasters. Since 1850, they have been responsible for the loss of over 420,000 lives and billions of dollars of damage to coastal structures and habitats. Most casualties were caused by local tsunamis that occur about once per year somewhere in the world. For example, the Indian Ocean tsunami on December 26, 2004, killed about 164,000 people close to the earthquake and about 63,000 people on distant shores. Predicting when and where the next tsunami will strike is currently impossible. Once a tsunami is generated, forecasting its arrival and impact is possible through modeling and measurement technologies.

Generation. Tsunamis are most commonly generated by earthquakes in marine and coastal regions. Major tsunamis are produced by large (greater than magnitude 7), shallow-focus (less than 30 km or 19 mi deep in the Earth) earthquakes associated with the movement of oceanic and continental plates. They frequently occur in the Pacific, where dense oceanic plates slide under the lighter continental plates. When these plates fracture, they provide a vertical movement of the sea floor that allows a quick and efficient transfer of energy from the solid earth to the ocean. When a powerful (magnitude 9.3) earthquake struck the coastal region of Indonesia in 2004, the movement of the sea floor produced a tsunami in excess of 30 m (100 ft) in height along the adjacent coastline. From this source, the tsunami radiated

outward and within 2 hours had claimed 63,000 lives in Thailand, Sri Lanka, and India. *See* EARTHQUAKE; PLATE TECTONICS.

Underwater landslides associated with smaller earthquakes are also capable of generating destructive tsunamis. The tsunami that devastated the northwestern coast of Papua New Guinea on July 17, 1998, was generated by a magnitude 7 earthquake that apparently triggered a large underwater landslide. Three waves measuring more than 7 m (23 ft) high struck a 10-km (6-mi) stretch of coastline within 10 minutes of the earthquake/slump. Three coastal villages were swept completely clean by the attack, leaving nothing but sand and 2200 people dead. Other large-scale disturbances of the sea surface that can generate tsunamis are explosive volcanoes and asteroid impacts. The eruption of the volcano Krakatoa in the East Indies on August 27, 1883, produced a 30-m (100-ft) tsunami that killed over 36,000 people. In 1997, scientists discovered evidence that a 4-km-diameter (2.5-mi) asteroid landed offshore of Chile approximately 2 million years ago and produced a huge tsunami that swept over portions of South America and Antarctica. *See* ASTEROID; SEISMIC RISK; VOLCANO.

Wave propagation. Because earth movements associated with large earthquakes are thousand of square kilometers in area, any vertical movement of the sea floor immediately changes the sea surface. The resulting tsunami propagates as a set of waves whose energy is concentrated at wavelengths corresponding to the earth movements (~100 km or 60 mi), at wave heights determined by vertical displacement (~1 m or 3 ft), and at wave directions determined by the adjacent coastline geometry. Because each earthquake is unique, every tsunami has unique wavelengths, wave heights, and directionality. From a tsunami warning perspective, this makes forecasting tsunamis in real time daunting. *See* OCEAN WAVES.

Warning systems. Since 1946, the tsunami warning system in the Pacific basin has monitored earthquake activity and the passage of tsunami waves at tide gauges. However, neither seismometers nor coastal tide gauges provide data for accurately predicting the impact of a tsunami at a particular coastal location. Monitoring earthquakes gives a good estimate of the potential for tsunami generation, based on earthquake size and location, but gives no direct information about the tsunami itself. Tide gauges in harbors provide direct measurements of the tsunami, but the tsunami is significantly altered by local bathymetry and harbor shapes, which severely limits their use in forecasting tsunami impact at other locations. Partly because of these data limitations, 15 of 20 tsunami warnings issued since 1946 were considered false alarms because the tsunami that arrived was too weak to cause damage. *See* SEISMOGRAPHIC INSTRUMENTATION.

Forecasting impacts. Real-time, deep-ocean tsunami detectors (tsunameters) will provide the data necessary to make tsunami forecasts (**Fig. 1**). On November 17, 2003, the Rat Islands tsunami in Alaska provided the most comprehensive test for this forecast methodology. The magnitude-7.8 earthquake on the shelf near the Rat Islands generated a tsunami that was detected by three tsunameters located along the Aleutian Trench. It was the first tsunami detected by the newly developed tsunameter system. These real-time data combined with the model database were then used to produce the real-time model tsunami forecast (**Fig. 2**). For the first time, tsunami model predictions were obtained, during the tsunami propagation, before the waves had reached many coastlines. The initial offshore forecast was obtained immediately after preliminary earthquake parameters (location and magnitude 7.5) became available from the West Coast/Alaska Tsunami Warning Center (about 15–20 min after the earthquake). The model estimates provided expected tsunami time series at tsunameter locations. When the closest tsunameter recorded the first tsunami wave, about 80 min after the tsunami, the model predictions were compared with the deep-ocean data and the updated forecast was adjusted immediately.

These offshore model scenarios were then used

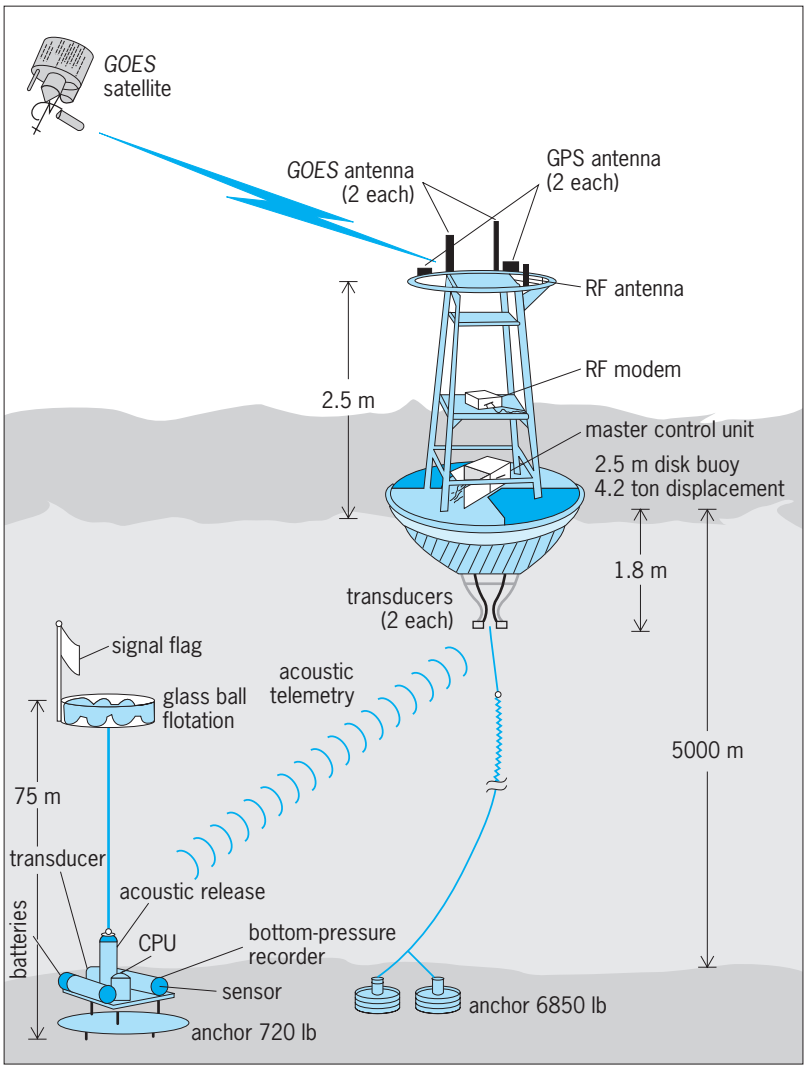


Fig. 1. Deep-ocean tsunami detection system.

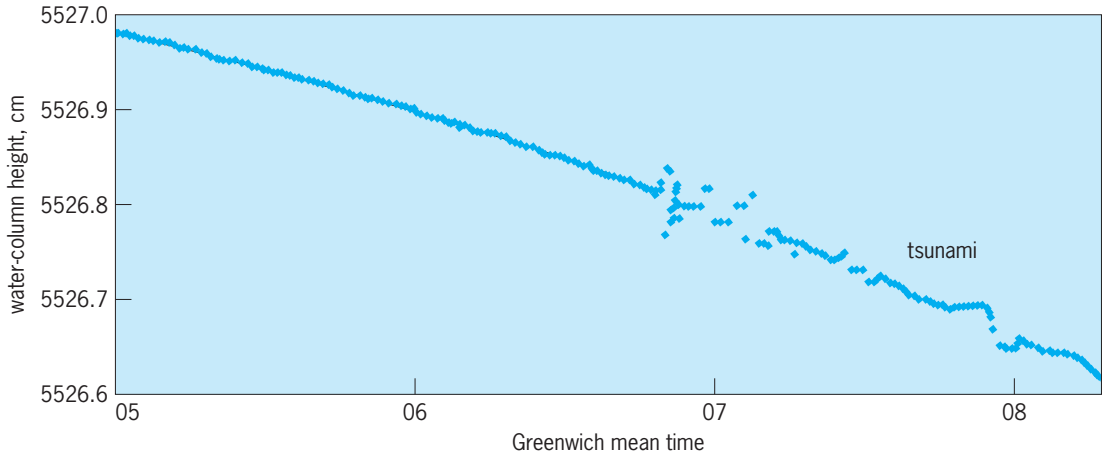


Fig. 2. Tsunami at Rat Islands, Alaska, on November 17, 2003, as measured at the tsunameter located at 50°N 171°W.

as input for the high-resolution inundation model for Hilo Bay, Hawaii. The model computed tsunami dynamics on several nested grids, with the highest spatial resolution of 30 m (100 ft) inside Hilo Bay. None of the tsunamis produced inundation at Hilo, but all of them recorded a nearly 0.5-m (1.6-ft) peak-to-trough signal at the Hilo tide gauge. Model forecast predictions for this tide gauge are compared with observed data in **Fig. 3**. The comparison demonstrates that amplitudes, arrival time, and periods of several first waves of the tsunami wave train were correctly forecasted. More tests are required to ensure that the inundation forecast will work for every tsunami likely to occur. When implemented, such a forecast will be obtained even faster, and would provide enough lead time for potential evacuation or warning cancellation for Hawaii and the west coast of the United States.

Reduction of impact. The recent development of real-time deep-ocean tsunami detectors and tsunami inundation models has given coastal communities the tools they need to reduce the impact of future tsunamis. If these tools are used in conjunction with a continuing educational program at the community level, at least 25% of the tsunami-related deaths might be averted. By contrasting the casualties from the 1993 Sea of Japan tsunami with that of the 1998 Papua New Guinea tsunami, we can conclude that these tools work. In the case of Aonae, Japan, about 15% of the population at risk died from a tsunami that struck within 10 minutes of the earthquake because the population was educated about tsunamis, evacuation plans had been developed, and a warning was issued. In the case of Warapa, Papua New Guinea, about 40% of the at risk population died from a tsunami that arrived within 15 min of the earthquake because the population was not educated, no evacuation plan was available, and no warning system existed. Eddie N. Bernard

Bibliography. E. N. Bernard (ed.), *Developing Tsunami-Resilient Communities: The National Tsunami Hazard Mitigation Program* (NTHMP) [reprinted from *Nat. Hazards*, 35(1), 2005], Springer, The Netherlands, 2005; E. N. Bernard, The U.S. National Tsunami Hazard Mitigation Program: A successful state-federal partnership, *Nat. Hazards*, 35(1):5–24, NTHMP, 2005; F. I. González et al., The NTHMP tsunameter network, *Nat. Hazards*, 35(1):25–39, NTHMP, 2005; F. I. González et al., Progress in NTHMP hazard assessment, *Nat. Hazards*, 35(1):89–110, NTHMP, 2005; V. V. Titov et al., The global reach of the 26 December 2004 Sumatra tsunami, *Science*, 309(5743):2045–2048, 2005; V. V. Titov et al., Real-time tsunami forecasting: Challenges

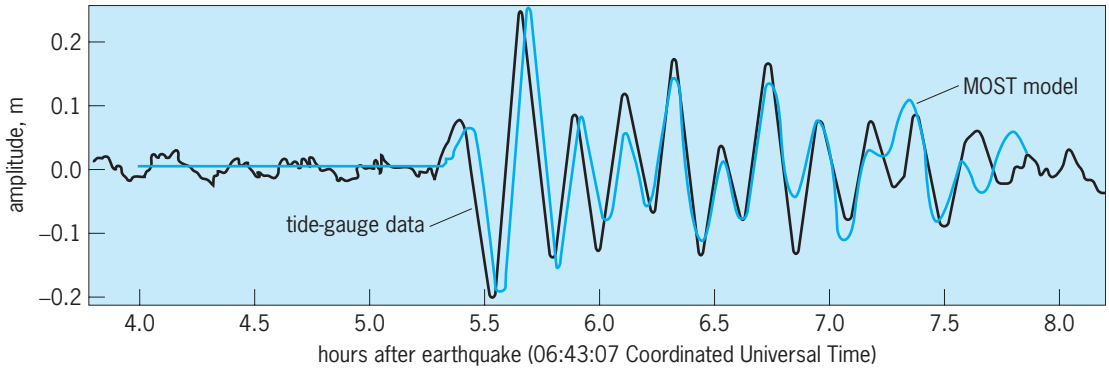


Fig. 3. Coastal forecast at Hilo, Hawaii, for 2003 Rat Islands tsunami, comparing the forecasted and measured gauge data.

and solutions, *Nat. Hazards*, 35(1):41–58, NTHMP, 2005.

Tuberculosis

An infectious disease caused by the bacillus *Mycobacterium tuberculosis*. It is primarily an infection of the lungs, but any organ system is susceptible, so its manifestations may be varied.

The tubercle bacillus was discovered by Robert Koch in 1882. Effective therapy and methods of control and prevention of tuberculosis have been developed, but the disease remains a major cause of mortality and morbidity throughout the world.

Tuberculosis is an acute health problem, particularly in the developing countries of Latin America, Africa, and Asia. The resurgence in tuberculosis between 1985 and 1993 was mainly the result of its occurrence in persons infected with human immunodeficiency virus (HIV) and with acquired immune deficiency syndrome (AIDS). In addition to individuals infected with HIV, there are other groups with a higher incidence of tuberculosis, including those having close contacts with infectious tubercular individuals; those with medical conditions such as silicosis, diabetes, end-stage renal disease, hematologic disease, and immunosuppressive therapy; intravenous drug users and alcoholics; and those in long-term care facilities, such as prisons and nursing homes. The treatment of tuberculosis has been complicated by the emergence of drug-resistant organisms, including multiple-drug-resistant tuberculosis, especially in those with HIV infection. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS).

Most new cases of clinical tuberculosis arise in individuals who have been previously infected. The eradication of tuberculosis and its prevention depend upon the detection and prophylactic treatment of infected individuals so that clinical disease does not occur, and effective treatment of those with symptomatic clinical disease so that transmission of tuberculosis to others is prevented.

Pathogenesis. *Mycobacterium tuberculosis* is transmitted by airborne droplet nuclei (1–10 micrometers in diameter) produced when an individual with active disease coughs, speaks, or sneezes. When inhaled, the droplet nuclei reach the alveoli of the lung. In susceptible individuals the organisms may then multiply and spread through lymphatics to the lymph nodes, and through the bloodstream to other sites such as the lung apices, bone marrow, kidneys, and meninges.

The development of acquired immunity in 2 to 10 weeks results in a halt to bacterial multiplication. Lesions heal and the individual remains asymptomatic. Such an individual is said to have tuberculous infection without disease, and will show a positive tuberculin test. The risk of developing active disease with clinical symptoms and positive cultures for the tubercle bacillus diminishes with time and may never occur, but is a lifelong risk.

Only close contacts such as family members or other sharing a closed environment are liable to become infected from a clinical case of tuberculous disease.

Only 5% of individuals with tuberculous infection progress to active disease. Progression occurs mainly in the first 2 years after infection; household contacts and the newly infected are thus at risk.

The classic pathologic lesion in tuberculosis is the granuloma or tubercle, the result of a cell-mediated hypersensitivity response. The tubercle consists of an aggregation of epithelioid cells including Langhans multinucleated giant cells, surrounded by a rim of monocytes, lymphocytes, and fibroblasts. Although the granuloma is not unique to tuberculosis, a distinctive feature in tuberculosis is a friable, cheeselike appearance in diseased tissue. This results from liquefaction of the epithelioid center of the tubercle.

Clinical manifestations. Many of the symptoms of tuberculosis, whether pulmonary disease or extrapulmonary disease, are nonspecific. Fatigue or tiredness, weight loss, fever, and loss of appetite may be present for months. A fever of unknown origin may be the sole indication of tuberculosis, or an individual may have an acute influenzalike illness. Erythema nodosum, a skin lesion, is occasionally associated with the disease.

The lung is the most common location for a focus of infection to flare into active disease with the acceleration of the growth of organisms. There may be complaints of cough, which can produce sputum containing mucus and pus. Though pulmonary tuberculosis is one of the major cases of blood in sputum, it is uncommon. Other chest complaints, such as shortness of breath or pain, are unusual. On examination, listening to the lungs may disclose rales or crackles and signs of pleural effusion (the escape of fluid into the lungs) or consolidation if present. In many, especially those with small infiltration, the physical examination of the chest reveals no abnormalities.

Miliary tuberculosis is a variant that results from the blood-borne dissemination of a great number of organisms resulting in the simultaneous seeding of many organ systems. The meninges, liver, bone marrow, spleen, and genitourinary system are usually involved. The term miliary refers to the lung lesions being the size of millet seeds (about 0.08 in. or 2 mm). These lung lesions are present bilaterally. Symptoms are variable and the diagnosis can be difficult.

Extrapulmonary tuberculosis is much less common than pulmonary disease. However, in individuals with AIDS, extrapulmonary tuberculosis predominates, particularly with lymph node involvement. Fluid in the lungs and lung lesions are other common manifestations of tuberculosis in AIDS. The lung is the portal of entry, and an extrapulmonary focus, seeded at the time of infection, breaks down with disease occurring.

Development of renal tuberculosis can result in symptoms of burning on urination, and blood and

white cells in the urine; or the individual may be asymptomatic. The symptoms of tuberculous meningitis are nonspecific, with acute or chronic fever, headache, irritability, and malaise.

A tuberculous pleural effusion can occur without obvious lung involvement. Fever and chest pain upon breathing are common symptoms.

Bone and joint involvement results in pain and fever at the joint site. The most common complaint is a chronic arthritis usually localized to one joint. Osteomyelitis is also usually present.

Pericardial inflammation with fluid accumulation (pericardial effusion) or constriction of the heart chambers secondary to pericardial scarring (constructive pericarditis) are two other forms of extrapulmonary disease.

Diagnosis. The principal methods of diagnosis for pulmonary tuberculosis are the tuberculin skin test, sputum smear and culture, and the chest x-ray. Culture and biopsy are important in making the diagnosis in extrapulmonary disease.

Tuberculin skin test. A positive tuberculin skin test defines tuberculous infection; it is the major and essential screening test. Some individuals with a positive test will have further studies done to rule out tuberculous disease; others will be selected for preventive drug therapy to thwart progression to disease.

Tuberculin is composed mainly of tuberculoprotein obtained from cultures of the tubercle bacillus. The Mantoux test is the standard tuberculin test. An intracutaneous injection of purified protein derivative tuberculin is performed, and the injection site is examined for reactivity (as manifested by induration) 48–72 h later. An individual with tuberculous infection develops a delayed hypersensitivity response to the tubercle bacillus, one manifestation of which is a positive reaction to the tuberculin skin test. A multiple-puncture technique is occasionally used for screening, but such a test is not as accurate as the Mantoux test, and any positive reaction has to be confirmed with a standard Mantoux test. Skin sensitivity is not affected by therapy and usually lasts for one's lifetime, though it may wane with age. *See* HYPERSENSITIVITY.

While a positive Mantoux test confirms tuberculous infection, a negative reaction does not necessarily mean tuberculosis is not present. A false negative reaction, or anergy, can occur with viral illness, AIDS, neoplastic disease, the use of immunosuppressive drug therapy, and fulminant disseminated tuberculosis, and in any condition affecting T-lymphocyte function.

Pulmonary disease. In those with symptoms suggesting clinical pulmonary disease, the chest x-ray film is essential since it reveals the extent and severity of disease. The four major radiographic abnormalities are segmental or lobar infiltration, pleural effusion, lymphadenopathy, and miliary densities. The most common pattern would be a nodular infiltrate with cavitation (see **illus.**). When an affected area heals, calcification, fibrosis, and volume loss may occur.

Activity or bacteriologic status cannot be determined from the chest film: while the film may be strongly suggestive of tuberculosis, other bacterial and fungal disease can give similar findings. Further diagnostic studies to demonstrate the organism via sputum smear and sputum culture are mandatory.

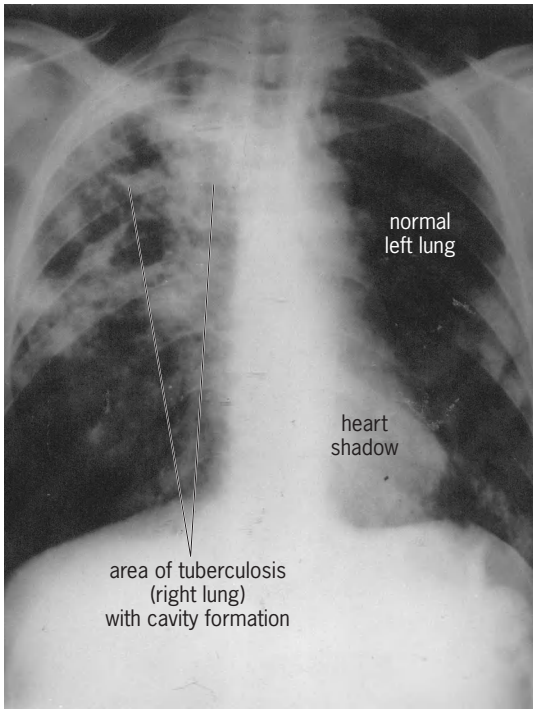
An individual with tuberculous infection without disease will have a negative chest x-ray, since the initial focus has healed and is not seen. However, there may be residual enlargement of some of the lymph nodes.

The best sputum sample for *M. tuberculosis* is a single early-morning specimen. At least three single morning specimens are obtained and sent for culture.

Extrapulmonary disease. The diagnostic approach to extrapulmonary disease depends upon the site involved, but usually specific tissue biopsy is required to identify the granulomas of tuberculosis.

The diagnosis of tuberculous pericarditis and peritonitis invariably requires biopsy. In addition to pathologic examination, all tissue specimens are cultured. The diagnosis of tuberculous meningitis is made by culture of cerebrospinal fluid. Synovial tissue (surrounding joints and connective tissue) biopsy and culture are done for tuberculosis arthritis. In genitourinary tuberculosis, urinalysis shows pus in the urine without bacteria the diagnosis depends on urine culture.

Treatment and control. The modern era of chemotherapy for tuberculosis began with the use of streptomycin in experimental tuberculosis in 1945. For the first time a drug was available that was remarkably effective against tuberculosis. Other agents followed: paraaminosalicylic acid in 1946, isoniazid and pyrazinamide in 1952, and ethambutol in 1961. By using a combination of agents for



Chest radiograph: posteroanterior view showing extensive tuberculosis with cavitation in right lung.

prolonged periods, drug resistance is overcome and potentially all cases can be cured with very low relapse rates. In 1966 the most potent drug to date, rifampin, was introduced for clinical use, permitting shorter drug regimens. *See* ANTIBIOTIC; DRUG RESISTANCE.

Preventive therapy. Screening for tuberculosis relies upon the tuberculin skin test. Since mass screening of the general population is not cost-effective, screening is directed at those subpopulations who have a high risk of becoming infected or have an increased incidence of tuberculous disease. Such subpopulations might include those in correctional institutions or health care facilities. After the individual with tuberculous infection is identified and a negative chest x-ray rules out clinical tuberculous disease, the use of isoniazid therapy daily for 6 months to 1 year will prevent development of tuberculous disease.

Since isoniazid is hepatotoxic (toxic to the liver), it cannot be used in everyone with a significant tuberculin reaction, but only in those in whom the risk of progression to tuberculous disease is greater than the risk of hepatotoxicity, which can cause hepatitis. Hepatitis due to isoniazid occurs mainly in those over 35 years of age. For the individual over 35 with tuberculous infection, another risk factor must be present to justify isoniazid use.

Tuberculous disease therapy. A combination of two or more drugs is used in the initial therapy of tuberculous disease. Drug combinations are used to lessen the chance of drug-resistant organisms surviving. Valuable first-line drugs include isoniazid, rifampin, pyrazinamide, ethambutol, and streptomycin. Isoniazid and rifampin are bactericidal and especially useful. Since it is injectable, streptomycin is useful in supervised programs when patient compliance is in doubt.

Other agents with activity against *M. tuberculosis* include capreomycin, kanamycin, ethionamide, paraaminosalicylic acid, and cycloserine. These agents are considered second-line drugs, and they are used when drug resistance causes treatment failure.

The preferred treatment regimen for both pulmonary and extrapulmonary tuberculosis is a 6-month regimen of the antibiotics isoniazid, rifampin, and pyrazinamide given for 2 months, followed by isoniazid and rifampin for 4 months. Because of the problem of drug-resistant cases, ethambutol can be included in the initial regimen until the results of drug susceptibility studies are known.

Once treatment is started, improvement occurs in almost all individuals. In pulmonary tuberculosis, sputum smears are the best way to ascertain effectiveness of therapy, and smears usually become negative for acid-fast organisms in a few weeks to 3 months. In 3–5 months, cultures usually become negative. After 10–14 days of adequate therapy, the patient is considered noninfectious due to rapid reduction in the total number of organisms in the sputum. After a course of therapy has been successfully completed, the patient is considered cured.

Any treatment failure or individual relapse is usually due to drug-resistant organisms. Retreatment is guided by drug susceptibility testing. Primary drug resistance where drug-resistant organisms are present at the start of therapy once occurred in less than 3% of cases; but with disease spread from persons with drug-resistant organisms, primary resistance is becoming more common. Acquired or secondary drug resistance is resistance that develops during treatment. In any case, drug resistance is associated with individuals not taking medication correctly. *See* DRUG RESISTANCE.

The community control of tuberculosis depends on the reporting of all new suspected cases so case contacts can be evaluated and treated appropriately as indicated. Individual compliance with medication is essential. Furthermore, measures to enhance compliance, such as directly observed therapy, may be necessary. *See* MYCOBACTERIAL DISEASES.

George Lordi

Bibliography. American Thoracic Society, Control of tuberculosis in the United States, *Amer. Rev. Respir. Dis.*, 146:1623–1633, 1992; American Thoracic Society, Treatment of tuberculosis and tuberculosis infection in adults and children, *Amer. J. Respir. Crit. Care Med.*, 149:1359–1374, 1994; B. R. Bloom (ed.), *Tuberculosis: Pathogenesis, Protection and Control*, 1994; P. D. Davies (ed.), *Clinical Tuberculosis*, 1993.

Tubeworms

The name given to marine polychaete worms (particularly to many species in the family Serpulidae) which construct permanent calcareous tubes on rocks, seaweeds, dock pilings, and ship bottoms. The individual tubes with hard walls of calcite-aragonite, ranging from 0.04 to 0.4 in. (1 mm to 1 cm) in diameter and from 0.16 to 4 in. (4 mm to 10 cm) in length, are firmly cemented to any hard substrate and to each other.

Economically they are among the most important fouling organisms both on ship hulls (where they are second only to barnacles) and inside seawater cooling pipes of power stations. A moderate growth of tubeworms can add more than 12% to the fuel costs of a medium-sized freighter. Mass settlements of tubeworm species in the genera *Hydroides*, *Ficopomatus*, *Spirobranchus*, *Serpula*, and *Pomatoceros* can deposit massive layers of calcareous tubes (3.2–8 in. or 8–20 cm in thickness per year) on dock walls and other submerged marine installations. On such structures as the legs of drilling platforms, the accumulated weight can lead to major structural damage.

Tubeworms feed by filtering suspended material from the water, using a crown of ciliated pinnate tentacles. For this reason, life on a ship bottom or inside an intake pipe can provide an enhanced water flow and food supply, and thus more rapid growth. Sexual reproduction in tubeworms results in planktonic trochosphere larvae, which are the natural dispersal

stage. After a varying period of larval life, settling (which may be gregarious on suitable surfaces) is induced by a combination of chemical and physical stimuli. Initial tube secretion after settlement can be rapid (up to 0.4 in. or 1 cm per week), and a new generation of tubeworms can quickly overgrow and smother a slower-growing parental one.

About 340 valid species of serpulid tubeworms have been described. The majority are truly marine, but several species of *Ficopomatus* thrive in brackish waters of low salinity, and one species occurs in fresh waters in Karst limestone caves. The wide geographical distribution of certain abundant species owes much to human transport on the bottoms of relatively fast ships and occurred within the last 120 years. *See* ANNELIDA; POLYCHAETA.

W. D. Russell-Hunter

Tubulidentata

An order of mammals containing a single living genus *Orycteropus*, the aardvark. Aardvarks occur in suitable habitats throughout sub-Saharan Africa. This order exhibits the results of an extreme adaptation for burrowing and feeding on small food items (particularly termites and ants). *See* MAMMALIA.

Aardvarks, also known as antbears, resemble a medium-sized to large pig. The body is massive with an elongate head and a piglike snout. The tough thick skin is sparsely covered with bristly hair. The ears are large and donkeylike and can be moved independently of each other. The strong muscular tail is kangaroolike. The short thick legs possess powerful sharp claws that are used to excavate burrows and to open the nests of termites and ants which are gathered by the aardvark's long sticky tongue. Adult aardvarks lack incisor and canine teeth. The simple peglike teeth on the sides of the jaws consist of tubular dentin covered by cement ("tubule teeth"). They lack enamel and grow continuously during the animal's life. The dental formula is I 0/0 C 0/0 Pm 2/2 M 3/3 × 2 for a total of 20 teeth. Adult aardvarks have a head and body length of 1000–1580 mm (39–62 in.), a tail length of 443–710 mm (17–28 in.), and a shoulder height of 600–650 mm (23–25 in.). Most weigh 50–70 kg (110–153 lb). *See* AARDVARK.

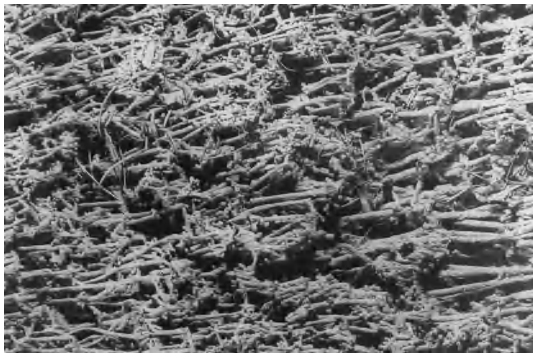
In the past, tubulidentates were often considered closely related to ungulates. However, recent mitochondrial and nuclear gene data show a close relationship to elephant-shrews, paenungulates (hyraxes, sirenians, and proboscideans), and golden moles (Chrysochloridae). All of these ecologically divergent forms probably originated in Africa. Molecular evidence implies that they all may have arisen from a common ancestor that existed in the Cretaceous Period when Africa was isolated from other continents. Three genera of the family Orycteropodidae are known: *Leptorycteropus*, *Myorycteropus*, and *Orycteropus*. The earliest known tubulidentate (*Myorycteropus*) is from early Miocene deposits found in Kenya in East Africa. *Orycteropus gaudryi*, a species from the late Miocene epoch, is similar to *O. afer* except that the former had a greater number of cheekteeth. A relatively unspecialized form, *Leptorycteropus*, dates from the mid-Pliocene epoch. Pleistocene remains are known from France, Greece, India, and Turkey. A genus from Madagascar (*Plesiorycteropus*) may be related to them.

Donald W. Linzey

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, Freeman, 1998; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Tufa

A spongy, porous limestone formed by precipitation from evaporating spring and river waters; also known as calcareous sinter. Calcium carbonate commonly precipitates from supersaturated waters on the leaves and stems of plants growing around the springs and pools and preserves some of their plant structures (see **illus.**). Tufa tends to be fragile and friable. Tufa deposits are limited in extent and are found mainly in the youngest rocks, Pleistocene or Holocene. *See* LIMESTONE; TRAVERTINE.



Calcareous tufa deposited on plant stems. (*Copyright © 1975 by Francis J. Pettijohn*)

Raymond Siever

Tuff

Fragmental volcanic products from explosive eruptions that are consolidated, cemented, or otherwise hardened to form solid rock. In strict scientific usage, the term "tuff" refers to consolidated volcanic ash, which by definition consists of fragments smaller than 2 mm. However, the term is also used for many pyroclastic rocks composed of fragments coarser than ash and even for pyroclastic material that has undergone limited posteruption reworking. If the thickness, temperature, and gas content of a tuff-forming pyroclastic flow are sufficiently high, the constituent fragments can become compacted and fused to form welded tuff. The term "tuff" is also used

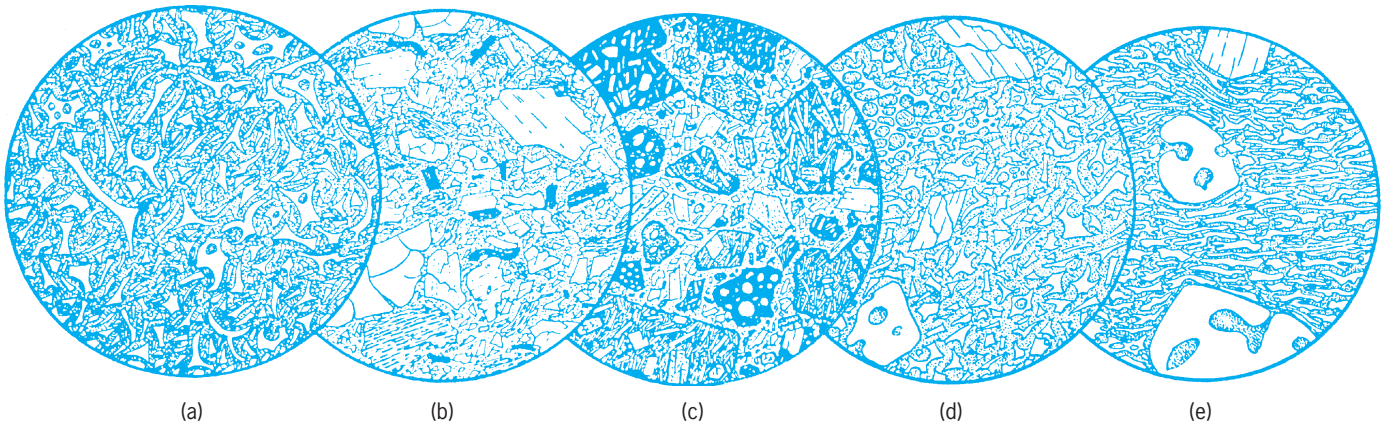


Fig. 1. Some common types of tuffs as viewed through the petrographic microscope. The field of view of the images is 2 mm for *a*–*c*, 3 mm for *d* and *e*. (*a*) Rhyolitic vitric tuff erupted from Mount Shasta Volcano, California, showing aggregates of glass fragments (shards) set in a matrix of very fine ash (volcanic dust). (*b*) Rhyolitic crystal tuff, Etsch Valley, Italy; broken crystals of quartz, plagioclase feldspar, and biotite set in a fine-grained matrix of glass and pumice fragments. (*c*) Andesitic lithic tuff, near Managua, Nicaragua, composed of andesitic rock fragments set in a finer-grained matrix of plagioclase and pyroxene and glass; there are no discrete large crystals. Each lithic fragment itself contains its own crystals set in a finer-grained matrix. (*d*) Tuff from the mostly unwelded top of the Bishop Tuff, an ignimbrite erupted from Long Valley Caldera, California; crystals of quartz and alkali feldspar are set in a matrix of undeformed glass shards and volcanic dust. (*e*) Welded tuff from the same locality as *d*, but the specimen was collected from the mostly welded interior part of the same ignimbrite; same constituents as in *d*, but here the glass shards are stretched, bent, and flattened by the welding (compaction) process. (*After H. Williams et al., Petrography, 2d ed., W. H. Freeman, 1982*)

in the naming of several related types of small volcanic edifices formed by hydrovolcanic eruptions, triggered by the explosive interaction of hot magma or lava with water. *See* IGNIMBRITE; PYROCLASTIC ROCKS.

As with their unconsolidated counterparts, the fragmental volcanic materials making up tuffs can vary widely in chemical and mineralogical composition, texture (crystalline or glassy), shape, and other properties. A common classification of the various types of tuff is based on the proportions of the constituent fragments: volcanic glass, crystals, and rock fragments (lithics). A tuff composed mostly of shards of glass is called vitric tuff, one containing mainly crystal is termed crystal tuff, and one in which rock fragments dominate is a lithic tuff (**Fig. 1**).

Tuffs are further distinguished by their approximate bulk chemical composition, as determined by the composition of the constituent fragments taken as a whole. Thus, depending on its bulk composition, a tuff can be more fully described as basaltic vitric tuff, andesitic crystal tuff, trachytic lithic tuff, rhyolitic vitric tuff, and the like. Vitric tuffs are generally characteristic of highly explosive and large-volume eruptions, and they can be deposited and preserved at hundreds of kilometers from their sources. Because of higher viscosity and gas content, these voluminous eruptions generally involve rhyolitic magmas; less commonly, intermediate composition magmas (andesitic, trachytic, dacitic); and, rarely, basaltic magmas. *See* IGNEOUS ROCKS; LAVA; MAGMA; VOLCANO.

Glass fragments (or shards) in tuffs are highly irregular in shape because of the violent shredding and sudden cooling (quenching) of gas-charged liquid lava when it is explosively ejected into the atmosphere. Typically, the shards are bounded by concave surfaces that represent the walls of ruptured gas bubbles (vesicles) [**Fig. 2**]. Larger glassy fragments, such as chunks of pumice not totally explosively disrupted, may contain numerous vesicles that have been preserved intact; a diagnostic feature of welded tuffs is the flattening and elongation of the vitric components. "Fresh" glass shards (that is, those unaltered by posteruption processes) are generally clear and colorless, but they also can have brownish, yellowish, or pinkish tones depending on chemical composition and trace impurities. In some samples, the glass is clouded by the suspension of tiny crystals (microlites or crystallites) of iron-titanium oxide or other dark-colored minerals. The larger crystals contained in tuffs are rarely preserved in their entirety; instead, they are broken into pieces, shattered by the force of explosive eruption. Rock fragments in tuffs are most commonly derived from solidified volcanic

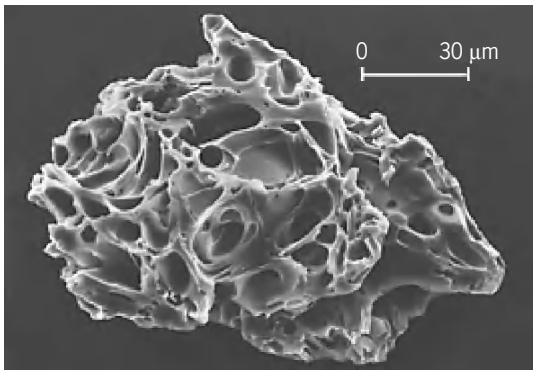


Fig. 2. Scanning electron microscope (SEM) image of a single ash particle (vitric fragment) from the May 18, 1980, eruption of Mount St. Helens, Washington. The tiny voids (vesicles) were created by expanding bubbles of volcanic gas during the rise and eruption of magma. If this fragment were further explosively shattered, the resulting tinier fragments (glass shards) would represent pieces of vesicle walls. (*Image by A. M. Sarna-Wojcicki, USGS*)

rocks erupted earlier from the same volcanic system (cognate pyroclasts), but they can also include small amounts of solid fragments unrelated to the volcano (accidental pyroclasts).

The vitric fragments of a tuff are readily altered by postdepositional processes, initially by hydration (chemical reaction with water) because of the percolation of hydrothermal fluids and ground water. For example, brown-colored glass of basaltic composition, upon hydration and weathering, is converted to yellowish or orange-colored palagonite; a pyroclastic rock composed mostly of hydrothermally altered or weathered basaltic glass is called palagonite tuff. Hydrated glassy material then begins to devitrify, that is, to crystallize to form extremely fine-grained aggregates of silica (colloidal and crystalline), feldspar, and clay minerals. Glassy fragments of rhyolitic composition are commonly altered to clay minerals of the montmorillonite (smectite) group to form a porous, light-colored rock called bentonite, widely found in volcanic regions. The postdepositional breakdown of crystals in tuffs is similar to that of the mineral constituents in most igneous rocks. The erosion and reworking of tuff by wind and water create sedimentary materials that are ultimately redeposited on land or in water, along with sediments of nonvolcanic origin, to form widespread submarine or subaerial volcaniclastic deposits. The term "tuffaceous" is used to describe sedimentary deposits which, while composed mostly of nonvolcanic fragments, contain an appreciable and recognizable component of ash-size pyroclasts. *See* CLAY MINERALS; SEDIMENTARY ROCKS.

In many volcanic regions of Italy, Mexico, and other countries, tuff is the common and preferred building stone. Many tuffs are colorful and visually attractive, soft enough to be quarried and shaped by hand and yet with sufficient structural strength to be set into walls with mortar. Dwellings and structures, including churches, were carved into thick, massive ignimbrite deposits by the inhabitants of the Cappadocia volcanic region (Anatolia, central Turkey) as early as mid-third century B.C. Robert Tilling

Bibliography. R. A. F. Cas and J. V. Wright, *Volcanic Successions: Modern and Ancient*, Chapman & Hall, London, 1988; R. V. Fisher and H.-U. Schmincke, *Pyroclastic Rocks*, Springer-Verlag, Berlin, 1984; G. Heiken, *An Atlas of Volcanic Ash: Smithsonian Contributions to the Earth Sciences*, Smithsonian Institution, Washington, DC, 1974; H. Williams, F. J. Turner, and C. M. Gilbert, *Petrography: An Introduction to the Study of Rocks in Thin Sections*, 2d ed., W. H. Freeman, San Francisco, 1982.

Tularemia

A worldwide disease caused by infection with the bacterium *Francisella tularensis*, which affects multiple animal species, including humans. Tularemia cases have been reported from North America, Europe, Russia, China, and Japan. In the winter, infection in adult humans occurs frequently from skinning infected rabbits, hares, muskrats, or beavers bare-handed; in the summer, infections occur in both adults and children from transmission of the bacteria through the bites of ticks or deer flies.

Clinical signs. Clinical signs typically occur 1–14 days after exposure. Tularemia can be difficult to differentiate from other diseases because it can have multiple clinical manifestations. Nonspecific signs frequently include fever, lethargy, anorexia, and increased pulse and respiration rates. The disease can overlap geographically with plague, and both may lead to enlarged lymph nodes (buboes). However, with tularemia, the buboes are more likely to ulcerate. If tularemic infection results from inhalation of dust from contaminated soil, hay, or grain, either pneumonia or a typhoidal syndrome can occur. Rarely, the route of entry for the bacteria is the eyes, leading to the oculoglandular type of tularemia. If organisms are ingested from soil, water, or contaminated wildlife, the oropharyngeal form can develop, characterized by abdominal pain, diarrhea, vomiting, and ulcers. *See* PLAGUE.

The mortality rate varies by species, although with treatment it is low. Ungulates are frequently infected but suffer low mortality from uncomplicated infections, except for documentation of high mortality in sheep on the range (up to 15% in untreated lambs), related to heavy tick infestations.

Tularemia is not transmitted directly from person to person. If the infected person or animal is untreated, blood remains infectious for 2 weeks and ulcerated lesions are infectious for a month. Deer flies (*Chrysops discalis*) are infective for 2 weeks, and ticks are infective throughout their lifetime (usually 2 years). Rabbit meat is infective even after being frozen for 3 years.

Ticks most frequently implicated in disease transmission include wood ticks (*Dermacentor andersoni*), dog ticks (*D. variabilis*), and Lone Star ticks (*Amblyomma americanum*). Deer flies (*C. discalis*) may also spread the organism. In Sweden, the implicated mosquito species is *Aedes cinereus*. Humans can be exposed through the bites of infected coyotes, squirrels, skunks, hogs, cats, or dogs.

Diagnosis and treatment. Confirmation of infection is usually made with a fourfold rise in specific antibody titers. Other tests, such as fluorescent antibody and culture, utilizing special media, may also be used. The liver, spleen, and lymph nodes can be enlarged, with whitish foci of necrosis.

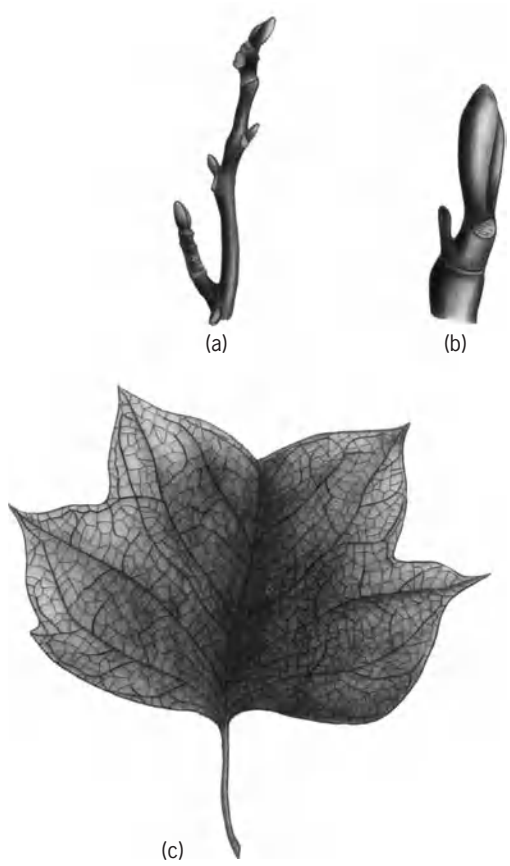
A number of antibacterial agents are effective against *F. tularensis*, the most effective being streptomycin. Gentamicin and tobramycin may be effective; and the tetracyclines and chloramphenicol can be utilized, but relapses occur, and thus treatment must continue until the temperature has been normal for 4–5 days. Penicillin and the sulfonamides have no therapeutic effect. *See* ANTIBIOTIC. Millicent Eidson

Bibliography. A. S. Benenson (ed.), *Control of Communicable Diseases Manual*, 1995; T. Morner, The ecology of tularaemia, *Rev. Sci. Technol.*, 11:1123–1130, 1992.

Tulip tree

A tree, *Liriodendron tulipifera*, also known in forestry as yellow poplar, belonging to the magnolia family, Magnoliaceae. One of the largest and most valuable hardwoods of eastern North America, it is native from southern New England and New York westward to southern Michigan, and south to Louisiana and northern Florida. In rich, moist soil it may grow 150 ft (45 m) tall and have a diameter of 8–10 ft (2.4–3 m). *See* MAGNOLIALES.

This tree is distinguished by leaves which are squarish at the tip as if cut off, true terminal buds flattened and covered by two valvate scales, stipular scars encircling the twig, an aromatic odor resembling that of magnolia, chambered white pith, and cone-shaped fruit which is persistent in winter



Tulip tree (*Liriodendron tulipifera*). (*a*) Twig, (*b*) terminal bud, and (*c*) leaf.

(see **illus.**). The name tulip refers to the large greenish-yellow and orange-colored flowers.

The wood of the tulip tree is light yellow to brown, hence the common name yellow poplar, which is a misnomer. It is a soft and easily worked wood, used for construction, interior finish, containers (boxes, crates, baskets), woodenware, excelsior, veneer, and sometimes for paper pulp. Because of its wide natural dimensions, the tulip tree often yields lumber as wide as 60 in. (150 cm) which is valuable for certain articles of furniture. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Tumbling mill

A grinding and pulverizing machine consisting of a shell or drum rotating on a horizontal axis. The material to be reduced in size is fed into one end of the mill. The mill is also charged with grinding material such as iron balls. As the mill rotates, the material and grinding balls tumble against each other, the material being broken chiefly by attrition.

Tumbling mills are variously classified as pebble, ball, or rod depending on the grinding material, and as cylindrical, conical, or tube depending on the shell shape. *See* CRUSHING AND PULVERIZING; GRINDING MILL; PEBBLE MILL.

Ralph M. Hardgrove

Tumor

Literally, a swelling; in the past the term has been used in reference to any swelling of the body, no matter what the cause. Thus a swollen region produced by edema, congestion, or hemorrhage into a tissue has been called a tumor. However, the word is now being used almost exclusively to refer to a neoplastic mass, and the more general usage is being discarded.

Neoplasm. A neoplastic mass or neoplasm is a pathological lesion characterized by the progressive or uncontrolled proliferation of cells. The cells involved in the neoplastic growth have an intrinsic heritable abnormality such that they are not regulated properly by normal methods. The stimulus which elicits this growth is not usually known. The cellular proliferation serves no useful function and often is very detrimental. Most cells in the body can undergo neoplastic changes, and hence there are many kinds of neoplasms. All classes of vertebrates have members which have developed neoplasms. These growths also have been seen in some invertebrates and plants.

Tumors are composed of two basic components: the parenchyma, which consists of the neoplastic proliferating cells, and the stroma, which is the supporting framework of connective tissue that includes the vascular supply. The stroma is derived from normal tissue, and the amount within a tumor varies greatly. It is the parenchyma which determines the biological behavior of the neoplasms.

Benign and malignant tumors. It is common to divide tumors into benign or malignant. The decision as to which category a tumor should be assigned is usually based on information gained from gross or microscopic examination, or both. Benign neoplasms usually grow slowly, remain so localized that they cause little harm, and generally can be successfully and permanently removed. Malignant or cancerous neoplasms tend to grow rapidly, spread throughout the body, and recur if removed.

Degree of harm. Not all tumors which have been classified as benign are harmless to the host, and some can cause serious problems. Difficulties may occur as a result of mechanical pressure. As the mass of cells increases in size, it may press against another

structure. In this way a blood vessel or duct may be occluded or a vital organ compressed. Benign tumors of glandular tissue can induce illness as a consequence of overproduction of certain hormones. Clinically there may be very abnormal signs. In contrast, some malignant neoplasms are relatively harmless in that they are slow-growing and often can be successfully removed. Examples of tumors with these characteristics are certain skin cancers.

Growth patterns. Neoplasms exhibit a wide range of abnormal growth patterns. It is not always easy to decide if a given tumor should be classified as benign or malignant. Nor is it known if certain benign tumors can, on occasion, progressively change and become malignant. The early phases of many malignancies greatly resemble benign growths.

Characteristics. There are a number of characteristics which are used to differentiate benign from malignant tumors. It is important to remember that these are generalities and that it is possible to find exceptions for some of the characteristics given.

The cells of benign tumors are well differentiated. This means that the cells are very like the normal tissue in size, structure, and spatial relationship. The cells forming the tumor usually function normally. Cell proliferation usually is slow enough so that there is not a large number of immature cells. Also, because growth is relatively slow, the stroma proliferation keeps pace with that of the parenchyma, and hemorrhage and ischemic necrosis are not common. As the cellular mass increases in size, most benign tumors develop a fibrous capsule around them which separates them from the normal tissue. The cells of a benign tumor remain at the site of origin and do not spread throughout the body. Anaplasia (loss of differentiation) is not seen in benign tumors.

The cells of malignant tumors may be well differentiated, but most have some degree of anaplasia. Anaplastic cells tend to be larger than normal and are abnormal, even bizarre, in shape. The nuclei tend to be very large, and irregular, and they often stain darkly. Mitoses are seen frequently, and necrosis and hemorrhage are common. Malignant tumors may be partially but never completely encapsulated. The cells of the cancer infiltrate and destroy surrounding tissue. They have the ability to metastasize; that is, cells from the primary tumor are disseminated to other regions of the body where they are able to produce secondary tumors called metastases. In summary, the characteristics which are most important in separating malignant tumors from benign growths are their anaplasia, invasiveness, and ability to metastasize.

In most cases the formation of a neoplasm is irreversible. It results from a permanent cellular defect which is passed on to daughter cells. Tumors should undergo medical appraisal to determine what treatment, if any, is needed. N. Karle Mottet; Carol Quaife

Tumor suppressor genes. Tumor suppressor genes are a class of genes which, when mutated, predispose an individual to cancer. The mutations result in the loss of function of the particular tumor suppressor protein encoded by the gene. Although this class of genes was named for its link to human cancer, it is now clear that these genes play a critical role in the normal development, growth, and proliferation of cells and organs within the human body. The protein product of many tumor suppressor genes constrains cell growth and proliferation so that these events occur in a controlled manner. Thus, these genes appear to act in a manner antagonistic to that of oncogenes, which promote cell growth and proliferation.

Currently, the retinoblastoma (RB) p53 and p16 genes are the best-understood tumor suppressors (see **table**). Inactivating mutations in the RB gene have been observed in retinoblastomas, osteosarcomas (cancer of the bone), as well as cancers of the lung, breast, and bladder. The RB gene regulates proliferation, growth, and replication of cells. It also regulates specific transcription factors which activate genes whose protein products are involved in cell cycle progression (the cycle of growth and division of cells). Cell proliferation and division occurs in response to external cellular signals (such as growth factors). The RB gene functions by coordinating the expression of a number of genes which play a significant role in cell proliferation.

Tumor suppressor genes

Syndrome (or tumor)	Name of gene	Tumor type
Retinoblastoma	RB	Retinoblastoma, small cell lung carcinomas, osteosarcomas
Li-Fraumeni	p53	Sarcomas, breast carcinomas, brain tumors
Familial adenomatous polyposis	APC	Colon cancer
Neurofibromatosis I	NF1	Neurofibromas
Neurofibromatosis II	NF2	Schwannomas, meningiomas
Wilms' tumor	WT-1	Nephroblastoma
Von Hippel-Lindau	VHL	Renal cell carcinoma
Familial breast cancer	BRCA1	Breast, ovary
Colorectal cancer	DCC	Colon cancer
Familial melanoma	p16	Melanoma, many others
Hereditary nonpolyposis colon cancer	hMSH2	Colon cancer
	hMLH1	Colon cancer
	hPMS1,2	Colon cancer

The p16 mutations have been observed in cancers of the skin, lung, breast, brain, bone, bladder, kidney, esophagus, and pancreas. By regulating key factors involved in cell cycle progression, the p16 gene functions to modulate the activity of the RB gene. RB gene function is modulated during the cell cycle by phosphorylation (covalent addition of phosphate). The enzymes responsible for the phosphorylation of the RB gene are called cyclin-dependent kinases (kinases are enzymes which enzymatically add phosphate to their substrates; cyclins play a key role in activating cyclin-dependent kinases). Phosphorylation of the RB gene inactivates its growth suppressive function. The p16 gene specifically inhibits the cyclin-dependent kinases responsible for the phosphorylation of the RB gene. Thus, p16 functions by keeping RB in its growth-suppressive form. A delicate balance between the activity of cyclin-dependent kinases and the tumor suppressor p16 allows for controlled cell cycle progression by the RB gene. Loss of p16 prevents the RB gene from controlling cell proliferation and growth.

The tumor suppressor p53 is the most frequently mutated gene associated with the development of many different types of human cancer, including those of the breast, lung, and colon. It is also associated with the rare inherited disease, Li-Fraumeni syndrome. Affected individuals manifest an increased likelihood of breast carcinomas (invasive cancers of epithelial origin), soft tissue sarcoma (cancers of the connective tissue), brain tumors, osteosarcoma, leukemia, and adrenocortical carcinoma. Like RB and p16, p53 has a role in cell cycle regulation. In response to DNA damage, p53 can cause cells to stop replicating, thus allowing time for DNA repair. In addition, p53 functions in the cell's decision on whether to undergo programmed cell death (apoptosis), a process that plays a critical role in the normal development and functioning of many organs in the human body. Deregulated cell proliferation and escape from apoptosis appear to be two common pathways leading to tumor formation. The p53 gene functions to provide a regulated balance between these two processes. This gene is able to bind to DNA in a specific manner and activate the transcription (production of mRNA) of a number of genes in order to regulate cell proliferation. The p53 gene can induce the expression of p21, which like p16 is an inhibitor of cyclin-dependent kinases involved in cell proliferation. By inducing the expression of p21, p53 is able to cause cells to stop proliferating. There is indirect evidence that p53 may have biochemical functions apart from its ability to bind DNA. These other functions are likely to play a significant role in the ability of p53 to control apoptosis and cancer.

The normal counterparts of tumor suppressors and oncogenes collaborate with each other to ensure appropriate growth and cell proliferation within the context of the body. It has been suggested that when these genes are mutated they work together to allow tumor growth, consistent with the observation that the accumulation of multiple loss-of-function mutations in tumor suppressor genes and gain of function mutations in oncogenes within a single cell gives rise to cancer. *See* CANCER (MEDICINE); GENE; MUTATION; ONCOLOGY; TUMOR VIRUSES.

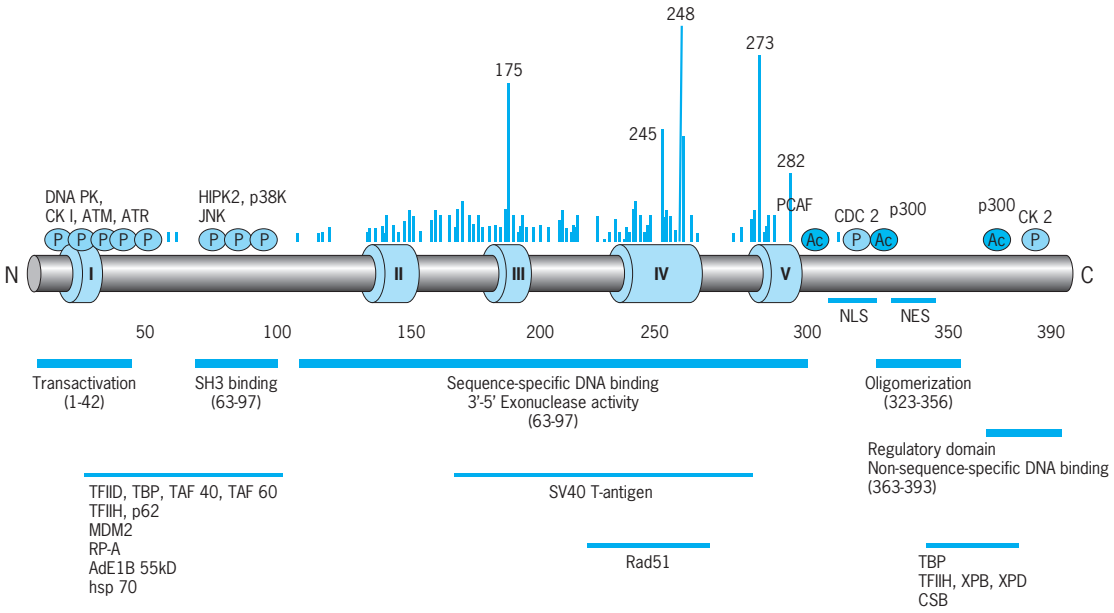
Mark Ewen

Bibliography. M. E. Ewen, The cell cycle and the retinoblastoma protein family, *Cancer Metastasis Rev.*, 13:45–66, 1994; E. R. Fearon and B. Vogelstein, A genetic model for colorectal tumorigenesis, *Cell*, 61:759–767, 1990; S. H. Friend et al., A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma, *Nature*, 323:643–646, 1986; A. J. Levine, The tumor suppressor genes, *Annu. Rev. Biochem.*, 62:623–651, 1993; R. A. Weinberg, Finding the anti-oncogene, *Sci. Amer.*, 259:44–51, 1988.

Tumor suppressor genes

A functionally heterogenous class of genes, ranging from nuclear transcription factors to cell adhesion molecules having as their common denominator that their function must be compromised during development and progression of a given tumor. The most important and most widely studied tumor suppressor is called p53. Its importance becomes evident from the fact that mutations in its gene are found in approximately 50% of all human cancers, thus constituting the most frequent alteration in a single human cancer-associated gene. The prevalence of p53 mutations in human cancer has spurred research worldwide to decipher p53 functions. It is now clear that the major function of p53 is to preserve the integrity of the genome of a cell under various conditions of cellular stress, endowing p53 with the title "guardian of the genome," coined by David Lane, one of the discoverers of p53 in 1979.

Tumorigenesis is a very complex process that is initiated by the accumulation of genetic alterations in at least two classes of genes: Activation of proto-oncogenes to oncogenes drives tumor initiation and progression by furthering uncontrolled cell growth and division, while inactivation of tumor suppressor genes leads to the elimination of important checkpoints in cellular proliferation that normally prevent uncontrolled proliferation. This implies that the products of tumor suppressor genes, the tumor suppressor proteins, exert functions that are disadvantageous to tumor development and progression. Inactivation of tumor suppressor genes usually occurs through mutations but can also occur by other means, like gene silencing by epigenetic means, for example, promoter inactivation through DNA methylation. Although both activation of proto-oncogenes and inactivation of tumor suppressor genes normally are the result of mutations as a consequence of manifold endogenous and exogenous DNA-damaging events, there is an important difference: as activation of a proto-oncogene to an oncogene leads to a "gain of function," already the mutational alteration of one of the two proto-oncogene alleles inherited from father or mother suffices to achieve an oncogenic



Landmarks for p53. Roman numerals represent the five regions that are conserved within p53 from all vertebrates. Known phosphorylation (P) and acetylation (Ac) sites are indicated. The vertical bars, clustered in the center of the p53 molecule, indicate amino acid residues mutated in human tumors (hot spots are identified by amino acid number). Below the molecule, horizontal bars indicate the current information concerning various domains of p53 for biological activities, p53 DNA interactions, and p53–protein complex formation; respective amino acids are given in brackets. Horizontal bars farther below mark interaction areas for various cellular and viral proteins interacting with p53. ATM indicates ATM kinase; ATR, ATR kinase; CK, casein kinase; CSB, Cockayne's syndrome B protein; DNA PK, DNA-dependent protein kinase; HIPK2, HIP kinase 2; JNK, Jun kinase; MDM2, mouse double minute protein 2; NES, nuclear export signal; NLS, main nuclear localization signal; RP-A, replication protein A; SV 40, Simian Virus 40; TAF, transcription-activating factor; TBP, TATA-Box binding protein; TF, transcription factor; XPB, xeroderma pigmentosum B protein; XPD, xeroderma pigmentosum D protein.

effect. As mutations in a tumor suppressor gene usually are characterized by a "loss of function," mutational inactivation of one allele of a tumor suppressor gene is not sufficient to eliminate the protective effect of the tumor suppressor, as the remaining nonmutated allele is still active. Inactivation of a tumor suppressor gene thus always requires that both of its two alleles are mutated.

Multifunctionality of p53. The protein p53 is truly multifunctional. It consists of several domains exhibiting a number of biochemical activities and interacting with a large variety of cellular and viral proteins (see **illustration**). The p53 core domain (amino acids 102–292) harbors at least three biochemical activities: (1) It mediates sequence-specific DNA binding to p53 consensus DNA sequences. Together with the bipartite *N*-terminal transactivation domain, this function enables p53 to act as a sequence-specific transcriptional activator. (2) The p53 core domain also recognizes complex non-B DNA structures, in both a sequence-specific and nonsequence-specific fashion. Recognition of structural features in nucleic acids also extends to the recognition of structural motifs in RNA, forming the molecular basis for p53 being a regulator of mRNA translation. (3) In addition, the p53 core domain exhibits an intrinsic 3′–5′ exonuclease activity (removes nucleotides from the ends of DNA).

As the p53 core domain is the major target for mutational inactivation of the p53 gene (see illustration), its multiple activities obviously are relevant to the tumor suppressor functions of p53. The C-terminal domain of p53 (amino acids 323–393), separated from the core domain by a flexible hinge region, harbors the p53 nuclear localization signals, an oligomerization domain, and a basic domain, which is able to bind non-sequence-specifically to DNA and RNA, and to exert an RNA and DNA reannealing activity. The basic C-terminal end also is the major regulatory domain of p53, affecting its DNA binding and p53 exonuclease activities.

All major domains of p53, but especially the p53 N- and C-termini, can interact with a large number of cellular and viral proteins. These proteins are partners of p53 in various p53 functions and regulate either positively (support) or negatively (inactivate), p53 functions. The physiologically most relevant cellular protein in this respect is MDM2, which binds to the p53 N-terminus and inhibits the transactivation function of p53, and it also targets p53 to proteasomal degradation. As the *mdm2* gene is a transcriptional target for the p53 protein, MDM2 and p53 are connected via a negative autoregulatory feedback loop.

Genomic integrity. The best-analyzed tumor suppressor function of p53 is that of a sequence-specific transcriptional activator (transactivator) of cellular genes which are involved in cell-cycle checkpoint control or apoptosis. In the absence of cellular stress, p53 is rapidly turned over and is present in normal cells only in small copy numbers of 1000–10,000 molecules. The high turnover is due to the interaction of p53 with MDM2. Various cellular stress signals, most importantly DNA damage, activate different signaling cascades, which lead to phosphorylation of p53 at N-terminal and C-terminal

phosphorylation sites. Phosphorylation of p53 first disrupts the p53–MDM2 interaction, resulting in metabolic stabilization of p53. Then p53 can act as a sequence-specific transactivator, leading to the upregulation of genes encoding cell-cycle inhibitors, like $p21^{waf1}$ or 14.3.3σ, which induce growth arrest in the G1 or G2 phase, respectively, or of repair-associated genes, like *gadd45*. The p53-induced growth arrest allows time to repair the inflicted DNA damage before it can be fixed as a mutation during the next round of DNA replication. Concerning tumor suppression, the most important p53 response is the induction of apoptosis in damaged cells, leading to their elimination. Consequently, p53 also activates a number of genes involved in apoptosis, like the *bax*, *fas-R*, *PUMA*, and several redox-related *PIG* genes. However, p53-mediated apoptosis is further complicated by the finding that induction of apoptosis by p53 involves nontranscriptional pathways in addition to transcriptional ones.

The p53 interacts with a variety of repair-associated proteins, and accumulating evidence strongly suggests that p53 is directly involved in repair processes and in the control of their fidelity. The idea of a direct involvement of p53 in repair processes is also strongly supported by the discovery that p53 exhibits an intrinsic 3′-5′ exonuclease activity. Exonucleases are required for nearly all processes of DNA metabolism, such as DNA replication, long-patch DNA repair, postreplicative mismatch repair, and DNA recombination. The p53 exonuclease activity thus strongly expands p53's possibilities as a guardian of the genome. So far, the exonuclease activity of p53 has been demonstrated to play a role in the fidelity control of homologous DNA recombination, where it removes mismatches in heteroduplex joints. Whether it also plays a role in DNA proofreading, for example, in DNA synthesized by the proofreader defective DNA polymerases α and β, is still a matter of debate. *See* CELL CYCLE (CANCER).

The p53 not only acts as a transactivator but also represses the expression of a large number of genes. In fact, in genewide microarray screens, more genes seem to be repressed than activated by p53. Although much less is known about the transrepression activity of p53 compared to its transactivation function, it is clear that transrepression of genes by p53 is even more complex than transactivation. Most likely, transrepression is not directly mediated by sequence-specific binding, but requires the interaction of p53 with other transcription factors, for example, with the Ets-1 protein, a transcription factor that activates a large number of genes that are involved in cell growth and invasion. By binding to Ets-1, p53 blocks the transcriptional activity of Ets-1, thereby exerting a tumor-suppressive function. Also in transrepression, the interaction of p53 with corepressors, like the Sin3a protein, might promote a local "closing" of the chromatin structure, thereby preventing transactivation.

Mutant p53. More than 80% of all mutations in the p53 gene are missense point mutations in the coding region of the p53 core domain, leading to the expression of a full-length p53 protein with a single amino acid exchange. Such a mutational spectrum is quite unusual for a tumor suppressor, as most tumor suppressors are inactivated by gene truncation or deletion or by promoter inactivation. The unusual mutational spectrum implies a selection advantage for tumor cells expressing such a mutant p53, and thus that mutant p53 exerts oncogenic properties. The oncogenic properties of mutant p53 are not yet understood at the molecular level. However, evidence obtained from the analyses of tumor banks strongly supports the concept that at least certain point mutations within the p53 gene not only serve to eliminate the tumor suppressor functions of p53 but enhance the aggressiveness of the respective tumors, leading to a poorer prognosis for tumor patients. In this regard, an important characteristic of at least most epithelial cancers is that mutations in the p53 gene are late events during their development, often associated with the onset of invasiveness. The p53 mutations thus are associated with tumor progression rather than with tumor initiation. The genetic instability of tumor cells expressing mutant p53 will allow the selection of more aggressive tumor cell variants and of tumor cells resisting tumor therapy. In addition, mutant p53 is suspected to actively suppress the responsiveness of tumor cells to various anticancer treatments.

Prognosis and therapy. Like no other tumor suppressor, the central role of p53 at the crossroad of cell cycle control, DNA repair, and apoptosis, as well as the fact that p53 is mutated in a large number of tumors, renders p53 a prominent target for applied tumor research. In contrast to wild-type p53, which is usually an unstable protein, mutant p53 accumulates in tumor cells. Detection of enhanced levels of p53 in tumor tissue is indicative for the expression of a mutant p53 and is already commonly used in tumor diagnosis. The value of mutant p53 as a diagnostic and prognostic marker will be enhanced by sequence analyses, allowing the classification of the aggressiveness of the respective mutant p53 as derived from the analyses of tumor banks.

The most challenging aspect of applied p53 research is the development of therapy regimens based on the p53 status of the tumor cells.

Gene therapy approaches. Wild-type p53 is able to induce apoptosis under appropriate conditions, whereas mutant p53 has lost this activity and even seems to actively block at least certain apoptotic pathways. Therefore, gene replacement, that is, introduction of wild-type p53 into tumors by various vector systems, is one of the aims of a p53-based gene therapy. In addition to the problems common to all gene therapy approaches, such as tumor-specific delivery of the therapeutic gene and its high-level expression in tumor cells only, several p53-specific problems must be resolved: the transduced wild-type p53 should not hetero-oligomerize with the endogenous mutant p53 because this would result in its inactivation; and it should specifically induce

apoptosis rather than growth arrest. Both problems can be overcome by genetically modifying the wild-type p53 to create a "super p53," which is more potent in induction of apoptosis than wild-type p53, and which contains a "foreign" oligomerization domain preventing hetero-oligomerization.

Conventional approaches. Development of low-molecular-weight compounds to be used as drugs in p53-based therapies is a promising alternative to p53-based gene therapy. One possibility is the restoration of the wild-type function of mutant p53 by drugs directly binding to the mutant p53. This has been demonstrated to work in vitro with tumor cells. However, the challenge will be to develop a drug that specifically reactivates mutant p53 without activating wild-type p53 in healthy tissues. Conversely, inactivation of the oncogenic properties of mutant p53 will possibly render otherwise resistant tumors amenable to conventional tumor therapy. A prerequisite, however, will be the identification of the oncogenic functions of mutant p53 at the molecular level. Finally, inhibition of the transactivation function of wild-type p53 in healthy tissue by a systemically applicable compound could prevent radiation-induced apoptosis in tissues surrounding the tumor, thereby allowing the use of very high doses of radiation in tumor therapy. Obviously, treatment with such compounds should not abolish p53-mediated protection of the healthy tissue from radiation-induced DNA damage, as predicted by the concept of an active involvement of wild-type p53 in repair processes in the absence of p53 transcriptional activation.

Like no other tumor-associated protein so far, p53 has stimulated both basic and clinically oriented molecular tumor research. Although current understanding of the molecular functions of p53 is still far ahead of its translation into therapies, this avenue of p53 research is picking up momentum. Given the importance of p53 in tumor treatment, any successful p53-based tumor therapy should provide a major breakthrough in the fight against cancer. *See* CANCER (MEDICINE); CELL CYCLE; GENETICS; MUTATION.

Wolfgang Deppert

Bibliography. E. Kim and W. Deppert, *Biochem. Cell Biol.*, 81(3):141–150, 2003; Reviews: p53, *Human Mutation*, vol. 21, 2003; Reviews: p53, *Oncogene*, vol. 18, no. 53, 1999.

Tumor viruses

Viruses associated with tumors can be classified in two broad categories depending on the nucleic acid in the viral genome and the type of strategy to induce malignant transformation.

RNA viruses. The ribonucleic acid (RNA) tumor viruses are retroviruses. When they infect cells, the viral RNA is copied into deoxyribonucleic acid (DNA) by reverse transcription and the DNA is inserted into the host genome, where it persists and can be inherited by subsequent generation of cells. Transformation of the infected cells can be traced to oncogenes that are carried by the viruses but are not necessary for viral replication. The viral oncogenes are closely similar to cellular genes, the proto-oncogenes, which code for components of the cellular machinery that regulates cell proliferation, differentiation, and death. Incorporation into a retrovirus may convert proto-oncogenes into oncogenes in two ways: the gene sequence may be altered or truncated so that it codes for proteins with abnormal activity; or the gene may be brought under the control of powerful viral regulators that cause its product to be made in excess or in inappropriate circumstances. Retroviruses may also exert similar oncogenic effects by insertional mutation when DNA copies of the viral RNA are integrated into the host-cell genome at a site close to or even within proto-oncogenes.

RNA tumor viruses cause leukemias, lymphomas, sarcomas, and carcinomas in fowl (avian leukosis virus, Rous sarcoma virus); rodents (Gross leukemia virus, Friend erythroleukemia virus, Harvery sarcoma virus, Moloney sarcoma virus, Bittner mammary adenocarcinoma virus); primates (simian sarcoma virus, gibbon ape leukemia virus); and other species. The human T-cell leukemia virus (HTLV) types I and II are endemic in Southeast Asian populations and cause adult T-cell leukemia and hairy-cell leukemia. *See* AVIAN LEUKOSIS; LEUKEMIA; LYMPHOMA; ROUS SARCOMA.

DNA viruses. DNA viruses replicate lytically and kill the infected cells. Transformation occurs in nonpermissive cells where the infection cannot proceed to viral replication. The transforming ability of DNA tumor viruses has been traced to several viral proteins that cooperate to stimulate cell proliferation, overriding some of the normal growth control mechanisms in the infected cell and its progeny.

Unlike retroviral oncogenes, DNA virus oncogenes are essential components of the viral genome and have no counterpart in the normal host cells. Some of these viral proteins bind to the protein products of two key tumor suppressor genes of the host cells, the retinoblastoma gene and the p53 gene, deactivating them and thereby permitting the cell to replicate its DNA and divide. Other DNA virus oncogenes interfere with the expression of cellular genes either directly or via interaction with regulatory factors. There is often a delay of several years between initial viral infection in the natural host species and the development of cancer, indicating that, in addition to virus-induced transformation, other environmental factors and genetic accidents are involved. A specific or general impairment of the host immune responses often plays an important role.

DNA tumor viruses belong to the families of papilloma, polyoma, adeno, hepadna, and herpes viruses and produce tumors of different types in frog (Luke's renal adenocarcinoma of leopard frog), chicken (Marek's disease, a herpesvirus-induced malignant lymphoma), woodchuck and ducks (hepatitis virus), rodents (Shope papilloma virus in rabbit, SV40 and adenovirus 12 in hamsters, polyoma virus

in mice, hamster, rat, guinea pigs, and rabbits), and primates (*Herpesvirus saimiri* and *H. ateles* in New World monkeys). *See* INFECTIOUS PAPILLOMATOSIS; MUTATION.

DNA tumor viruses are thought to play a role in the pathogenesis of about 15–20% of human cancers. These include African Burkitt's lymphoma, nasopharyngeal carcinoma, immunoblastic lymphomas in immunosuppressed individuals and a proportion of Hodgkin's lymphomas that are all associated with the Epstein-Barr virus of the herpes family; and liver carcinoma in Southeast Asia and tropical Africa associated with chronic hepatitis B virus infection. *See* ANIMAL VIRUS; CANCER (MEDICINE); EPSTEIN-BARR VIRUS; HODGKIN'S DISEASE; ONCOLOGY. Maria G. Masucci

Bibliography. J. M. Bishop, Viral oncogenes, *Cell*, 42:23–38, 1985; J. F. Nevins, E2F: A link between the Rb tumor suppressor protein and viral oncoproteins, *Science*, 258:424–429, 1992; R. A. Wienberg, Tumor suppressor genes, *Science*, 254:1138–1146, 1991; H. zur Hausen, Viruses in human cancers, *Science*, 254:1167–1173, 1991.

Tuna

Certain perciform (spiny-rayed) fishes in the family Scombridae. Like all other scombrids (such as mackerels, bonitos, wahoo, and sierras), tunas have a fusiform (tapering toward each end) and moderately compressed body and certain other characteristics that adapt them for sustained swimming at high speeds. The long spinous dorsal fin is depressible in a groove in the back; the pelvic fins and usually the pectoral fins are small and retractable in shallow depressions; the scales are typically small, but enlarged scales usually cover the anterior part of the body and lateral line, and form an envelopment called the corselet. The eyes protrude very little, if at all, beyond the surface of the head; the mouthparts fit snugly against the pointed head; and the gill covers fit snugly against the body. These features eliminate almost all irregularities that would cause resistance to the water. Tunas are also recognized by the finlets (independent multibranched rays, each appearing as a small fin) behind the dorsal and anal fins. The slender caudal peduncle, supported on each side by two small keels and a large median keel in between, and the lunate caudal fin are driven by powerful muscles for fast and sustained swimming. Sustained swimming depends on red muscle (comparatively thin muscle fibers containing large amounts of myoglobin and mitochondria), and the body temperature of tunas may be several degrees above water temperature. Tunas feed on a wide variety of fishes, squids, and crustaceans. *See* MUSCULAR SYSTEM; PERCIFORMES.

In American waters there are nine species of scombrids in four genera which bear the name tuna. However, there are other scombrids that are just as much tuna fishes as those that bear the name. Following is a brief description of "tunas" that occur in the western North Atlantic, the eastern North Pacific, and beyond. Practically all tunas are pelagic (that is, living in the open sea) and oceanodromous (migratory in salt water), with most species being highly migratory. Most species are very important commercial food fishes as well as game fishes.

Tribe Thunnini. The upper part of the body lacks distinct dark oblique stripes; the surface of the tongue has two longitudinal ridges; the maxilla can reach to, but does not go past, the middle of the eye; and there are 9–16 spines in the first dorsal fin.

Slender tuna (Allothunnus fallai). The body is slender, with a fork length (the length of a fish from its mouth to the fork in its tail) about five times the depth of the body; the dorsum is bluish to deep purple; and the belly is white without stripes or spots. Slender tuna are circumglobal and of minor commercial importance. The fork length is 105 cm (41 in.).

Genus Auxis. The first and second dorsal fins are widely separated, and the interpelvic process (a fleshy projection between the inner edges of the pelvic fins) is a large, single-pointed flap.

Bullet tuna (Auxis rochei). Bullet tuna are also called bullet mackerel. They inhabit the Pacific, Indian, and Atlantic oceans, including the Mediterranean Sea. The fork length is 50 cm (20 in.).

Frigate tuna (Auxis thazard). Frigate tuna are also called frigate mackerel. They inhabit the Atlantic, Indian, and Pacific oceans. The fork length is 65 cm (26 in.).

Genus Euthynnus. The body is naked beyond the corselet, and usually has black spots between the pectoral and pelvic fin bases.

Little tunny (Euthynnus alletteratus). Little tunny are reef-associated and oceanodromous. They inhabit the tropical and subtropical Atlantic, including the Mediterranean and Black seas, Caribbean Sea, and Gulf of Mexico. The total length is 122 cm (48 in.).

Kawakawa (Euthynnus affinis). The posterior portion of the back has a pattern of broken oblique stripes. They inhabit the Indo–West Pacific and eastern central Pacific and are highly migratory. The fork length is 100 cm (39 in.).

Black skipjack (Euthynnus lineatus). The black skipjack are generally iridescent blue with black dorsal markings composed of three to five horizontal stripes and variable black or dark gray spots above the pelvic fins; they occasionally have extensive longitudinal stripes of light gray on the belly. They are of minor commercial importance. They inhabit the eastern Pacific from California to Peru and the Galápagos Islands. The fork length is 84 cm (33 in.).

Skipjack tuna (Katssuwonus pelamis). These have three to five longitudinal stripes on the belly. Although highly important commercially, the skipjack tuna have been reported for ciguatera poisoning. They are highly migratory in all tropical and warm-temperate seas, and are absent from the Mediterranean and Black seas. The fork length is 108 cm (42 in.).

Genus Thunnus. There are no dark stripes or black spots on the body, which is covered with very small scales beyond the corselet.

Yellowfin tuna (Thunnus albacares). The second dorsal and anal fins are exceptionally long and falcate

(sickle-shaped) and yellow. They are found worldwide but are absent from the Mediterranean Sea. The fork length is 239 cm (94 in.).

Blackfin tuna (Thunnus atlanticus). Blackfin tuna have relatively few (19 to 25) gill rakers (bony or cartilaginous projections that point forward and inward from the gill arches), whereas other species of genus *Thunnus* have more. Their backs are metallic dark blue, their sides are silvery, and their bellies are milky white. They are found in the western Atlantic from Massachusetts to the Trinidad Islands to Rio de Janeiro, Brazil. The fork length is 108 cm (42 in.).

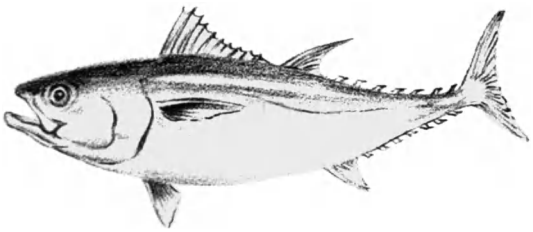
Bigeye tuna (Thunnus obesus). The corselet of bigeye tuna is not very distinct; the pectoral fin is moderately long; the first dorsal fin is deep yellow; the second dorsal and anal fins are light yellow; and the finlets are bright yellow edged with black. They inhabit tropical and subtropical waters of the Atlantic, Indian, and Pacific oceans, and are absent from the Mediterranean Sea. The total length is 250 cm (98 in.).

Atlantic bluefin tuna (Thunnus thynnus). These tuna have backs that are dark steel blue or nearly black; their sides and belly are silvery gray, with large silvery spots and bands; their cheeks are silvery (see **illustration**). They inhabit the western Atlantic from Canada to the Gulf of Mexico, the Caribbean Sea to Venezuela, and south to Brazil; and Norway to the Canary Islands, including the Mediterranean and Black seas. These are the largest tunas. The all-tackle angling record is a 679 kg (1494 lb) specimen caught off Nova Scotia in 1979. The total length is 458 cm (180 in.).

Pacific bluefin tuna (Thunnus orientalis). Once considered as a subspecies, the Atlantic bluefin and the Pacific bluefin are very similar, with the latter reaching a much smaller maximum size. They inhabit the Gulf of Alaska to southern California, and the Sea of Okhotska south to the northern Philippines. The fork length is 300 cm (118 in.).

Albacore (Thunnus alalunga). Albacore constitute the most distinctive species of *Thunnus*, with a long paddle-shaped pectoral fin that reaches past the second dorsal fin and often to the second finlet. They are cosmopolitan in range, inhabiting the tropical and temperate waters of all oceans, including the Mediterranean Sea. The fork length is 140 cm (55 in.).

Tribe Sardini. The upper part of the body has 5 to 10 dark oblique stripes; the surface of the tongue lacks a pair of cartilaginous longitudinal ridges; the mouth is large, with the maxilla reaching past the middle of the eye; and there are 20 to 22 spines in the first dorsal fin.



Bluefin tuna (*Thunnus thynnus*).

Atlantic bonito (Sarda sarda). The mouth of Atlantic bonito is large, with the maxilla reaching beyond the posterior margin of the eye. The body exhibits oblique dorsal stripes with a greater angle than in other species of *Sarda*. Atlantic bonito have been reported for ciguatera poisoning. They inhabit the eastern Atlantic from Norway to South Africa, including the Mediterranean and Black seas; the western Atlantic from Nova Scotia to Florida, the northern Gulf of Mexico, and Colombia to northern Argentina; and are apparently absent from most of the Caribbean Sea. The fork length is 96 cm (38 in.).

Pacific bonito (Sarda chiliensis lineolata). These exhibit oblique stripes on the back; the mouth is of moderate size. They inhabit the eastern Pacific from Alaska to Baja California. The fork length is 102 cm (40 in.).

Striped bonito (Sarda orientalis). These have virtually horizontal stripes on the back. They inhabit the Indo-Pacific, the Hawaiian Islands, and the eastern Pacific from the coast of the United States to Baja California to Peru and the Galápagos Islands. The fork length is 102 cm (40 in.).

Herbert Boschung

Bibliography. B. B. Collette, Mackerels, Family Scombridae, pp. 516–536 in B. B. Collette and G. Klein-MacPhee (eds.), *Bigelow and Schroder's Fishes of the Gulf of Maine*, 3d ed., Smithsonian Institution Press, Washington, DC, 2002; B. B. Collette, Mackerels, molecules, and morphology, pp. 149–164 in B. Séret and J.-Y. Sire (eds.), *Proceedings of the 5th Indo-Pacific Fish Conference*, Noumea, Paris, 1999; B. B. Collette, Scombridae, in W. Fisher (ed.), *FAO Species Identification Sheets for Fishery Purposes*, vol. 4: *Western Central Atlantic*, Rome, 1978; B. B. Collette and C. R. Aadland, Revision of the frigate tunas (Scombridae, *Auxis*), with descriptions of two new subspecies from the eastern Pacific, *Fish. Bull.*, 94:423–441, 1996; B. B. Collette and C. E. Nauen, *FAO Species Catalogue*, vol. 2: *Scombrids of the World*, an annotated and illustrated catalogue of tunas, mackerels, bonitos and related species known to date, FAO Fish. Synopsis, 125(2), 1983; B. B. Collette and B. R. Smith, Bluefin tuna, *Thunnus thynnus orientalis* from the Gulf of Papua, *Jap. J. Ichthyol.*, 28(2):166–168, 1981; B. W. Halstead, P. S. Auerbach, and D. R. Campbell, *A Colour Atlas of Dangerous Marine Animals*, Wolfe Medical Publications, W. S. Cowell, Ipswich, England, 1990; D. A. Olsen, D. W. Nellis, and R. S. Wood, Ciguatera in the Eastern Caribbean, *Mar. Fish. Rev.*, 46(1):13–18, 1984.

Tundra

An area supporting some vegetation beyond the northern limit of trees, between the upper limit of trees and the lower limit of perennial snow on

mountains, and on the fringes of the Antarctic continent and its neighboring islands. The term is of Lapp or Russian origin, signifying treeless plains of northern regions. Biologists, and particularly plant ecologists, sometimes use the term tundra in the sense of the vegetation of the tundra landscape. Tundra has distinctive characteristics as a kind of landscape and as a biotic community, but these are expressed with great differences according to the geographic region.

Patterns. Characteristically tundra has gentle topographic relief, and the cover consists of perennial plants a few centimeters to a meter or a little more in height. The general appearance during the growing season is that of a grassy sward in the wetter areas, a matted spongy turf on mesic sites, and a thin or sparsely tufted lawn or lichen heath on dry sites. In winter, snow mantles most of the surface with drifts shaped by topography and surface objects including plants; vegetation patterns are largely determined by protecting drifts and local areas exposed to drying and scouring effects of winter winds. By far, most tundra occurs where the mean annual temperature is below the freezing point of water, and perennial frost (permafrost) accumulates in the ground below the depth of annual thaw and to depths at least as great as 1600 ft (500 m). A substratum of permafrost, preventing downward percolation of water, and the slow decay of water-retaining humus at the soil surface serve to make the tundra surface moister during the thaw season than the precipitation on the area would suggest. Retention of water in the surface soils causes them to be subject to various disturbances during freezing and thawing, as occurs at the beginning and end of, and even during, the growing season. Where the annual thaw reaches depths of less than about 20 in. (50 cm), the soils undergo "swelling," frost heaving, frost cracking, and other processes that result in hummocks, polygonal ridges or cracks, or "soil flows" that slowly creep down slopes. As the soils are under this perennial disturbance, plant communities are unremittingly disrupted and kept actively recolonizing the same area. Thus topography, snow cover, soils, and vegetation interact to produce patterns of intricate complexity when viewed at close range.

Plant species, life-forms, and adaptations. The plants of tundra vegetation are almost exclusively perennial. A large proportion have their perennating buds less than 8 in. (20 cm) above the soil (chamaephytes in the Raunkiaer life-form system), especially among the abundant mosses and lichens. Another large group has the perennating organs at the surface of the soil (hemicryptophytes in the Raunkiaer system). Vegetative reproduction is common—by rhizomes (many of the sedges), stolons (certain grasses and the cloudberry, *Rubus chamaemorus*), or bulbils near the inflorescence (*Polygonum viviparum*, *Poa vivipara*, *Saxifraga hirculis*); thus clone formation is common in plant populations. Apomixis, the short-circuiting of the sexual reproduction process, is found frequently among flowering plants of tundra. Seed is set regularly by agamospermy, for example, in many dandelions (*Taraxacum* sp.), hawkweeds (*Hieracium* sp.), and grasses (*Calamogrostis* sp., *Poa* sp., *Festuca* sp.). The high incidence of apomixis in tundra flowering plants is coincident with high frequency of polyploidy, or multiple sets of chromosomes, in some circumstances a mechanical cause of failure of the union of gametes by the regular sexual process. Asexual reproduction and polyploidy tend to cause minor variations in plant species populations to become fixed to a greater extent than in populations at lower latitudes, and evolution tends to operate more at infraspecific levels without achieving major divergences. Adaptations are more commonly in response to physical factors of the stressful cold environment rather than to biotic factors, such as pollinators or dispersal agents, of the kinds that exert such control in the congenial warm, moist climates.

Soil conditions. Tundra soils are azonal, without distinct horizons, or weakly zonal. Soils on all but very dry and windswept sites tend to accumulate vegetable humus because low temperatures and waterlogging of soils inhibit processes of decay normally carried out by bacteria, fungi, and minute animals. Where permafrost or other impervious layers are several meters or more beneath the surface in soils with some fine-grained materials, leaching produces an Arctic Brown Soil in which there is moderately good drainage and cycling of mineral nutrients. In the greater part of tundra regions not mantled by coarse, rocky "fjell-field" materials (**Fig. 1**), the soils are more of the nature of half-bog or bog soils. These are characterized by heavy accumulations of raw or weakly decayed humus at the surface overlying a waterlogged or perennially frozen mineral horizon that is in a strongly reduced state from lack of aeration. Such boggy tundra soils are notoriously unproductive from the standpoint of cultivated plants, but they are moderately productive from the standpoint of shallowly rooted native plants. In

Fig. 1. Fjell-field tundra of the high Arctic. Sedges, mosses, and lichens form a thin and discontinuous sod. Late-persisting snowbanks are withdrawing from surfaces that are lighter in color because they lack many of the common plants, including dark-colored species of lichens. (*W. S. Benninghoff, U.S. Geological Survey*)

Fig. 2. Alpine tundra in French Alps. Altitudinal limit of trees occurs in valley behind building in middle distance. Although similar in vegetation structure to tundra of polar regions, Alpine tundras of lower latitudes are usually richer in vascular plant species than tundras of polar regions, and structure and composition of the vegetation have been modified by pasturing. (*W. S. Benninghoff, U.S. Geological Survey*)

Finland, forest plantations are being made increasingly productive on such soils by means of nutrient feeding to aerial parts. *See* SOIL.

Productivity. By reason of its occurrence where the growing season is short and where cloudiness and periods of freezing temperatures can reduce growth during the most favorable season, tundra vegetation has low annual production. Net radiation received at the Earth's surface is less than 20 kg-cal/(cm^2 · year) [84 kilojoules/(cm^2 · year)] for all Arctic and Antarctic tundra regions. Assuming a 2-month growing season and 2% efficiency for accumulation of green plant biomass, 1 cm^2 could accumulate biomass equivalent to 66.6 g-cal/year (279 J/year). This best value for tundra is not quite one-half the world average for wheat production and about one-eighth of high-yield wheat production. The tundra ecosystem as a whole runs on a lower energy budget than ecosystems in lower latitudes; in addition, with decomposer and reducer organisms working at lower efficiency in cold, wet soils, litter, and humus accumulate, further modifying the site in unfavorable ways. Grazing is one of the promising management techniques (**Fig. 2**) because of its assistance in speeding up the recycling of nutrients and reducing accumulation of raw humus. *See* BIOMASS; ECOLOGICAL ENERGETICS; ECOSYSTEM.

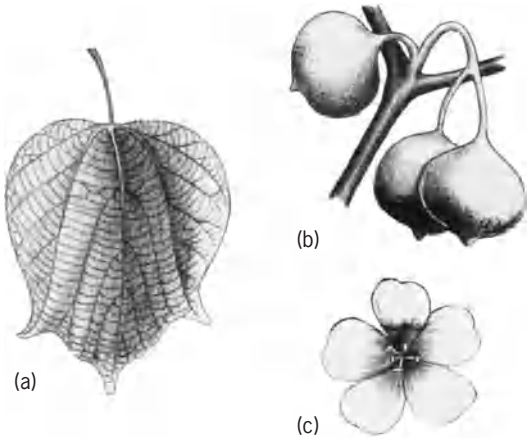
Fauna. The Arctic tundras support a considerable variety of animal life. The vertebrate herbivores consist primarily of microtine mammals (notably lemmings), hares, the grouselike ptarmigan, and caribou (or the smaller but similar reindeer of Eurasia). Microtine and hare populations undergo cyclic and wide fluctuations of numbers; these fluctuations affect the dependent populations of predators, the foxes, weasels, hawks, jaegers, and eagles. Alpine tundras generally have fewer kinds of vertebrate animals in a given area because of greater discontinuity of the habitats. Arctic and Alpine tundras have distinctive migrant bird faunas during the nesting season. Tundras of the Aleutian Islands and other oceanic islands are similar to Alpine tundras with respect to individuality of their vertebrate faunas, but the islands support more moorlike matted vegetation over peaty soils under the wetter oceanic climate. Tundras of the Antarctic continent have no vertebrate fauna strictly associated with it. Penguins and other sea birds establish breeding grounds locally on ice-free as well as fringing ice-covered areas. The only connection those birds have with the tundra ecosystem is the contribution of nutrients from the sea through their droppings. All tundras, including even those of the Antarctic, support a considerable variety of invertebrate animals, notably nematode worms, mites, and collembola on and in the soils, but some other insects as well. Soil surfaces and mosses of moist or wet tundras in the Arctic often teem with nematodes and collembola. Collembola, mites, and spiders have been found above 20,000 ft (6000 km) in the Himalayas along with certain molds, all dependent upon organic debris imported by winds from richer communities at lower altitudes. *See* COLLEMBOLA; TAIGA.

William S. Benninghoff

Bibliography. Arctic Institute of North America, *Arctic Bibliography*, 16 vols., 1953–1975; M. J. Dunbar, *Ecological Development in Polar Regions*, 1968; R. E. English, *World Regional Geography*, 1990; J. D. Ives and R. G. Barry, *Arctic and Alpine Environments*, 1974; M. C. Kellman, *Plant Geography*, 2d ed., 1980; G. A. Llano (ed.), *Antarctic Terrestrial Biology*, American Geophysical Union, Antarctic Research Series, vol. 20, 1972; J. C. F. Tedrow (ed.), *Antarctic Soils and Soil Forming Processes*, American Geophysical Union, Antarctic Research Series, vol. 8, 1966; H. E. Wright, Jr., and W. H. Osburn, *Arctic and Alpine Environments*, 1968.

Tung tree

The plant *Aleurites fordii*, a species of the spurge family (Euphorbiaceae). The tree, native to central and western China, is the source of tung oil.



Tung tree. (*a*) Leaf. (*b*) Fruits. (*c*) Flower.

It has been grown successfully in the southern United States. The globular fruit (see **illus.**) has three to seven large, hard, rough-coated seeds containing the oil, which is expressed after the seeds have been roasted. Tung oil is used to produce a hard, quick-drying, superior varnish, which is less apt to crack than other kinds. The foliage, sap, fruit, and commercial tung meal contain a toxic saponin, which causes gastroenteritis in animals that eat it. *See* DRYING OIL; EUPHORBIALES; VARNISH.

Perry D. Strausbaugh; Earl L. Core

Tungsten

A chemical element, W, atomic number 74, and atomic weight 183.85. Naturally occurring tungsten consists of five stable isotopes having the following mass numbers and relative abundances: 180 (0.14%), 182 (26.4%), 183 (14.4%), 184 (30.6%), and 186 (28.4%). Twelve radioactive isotopes ranging from 173 to 189 also have been characterized. *See* PERIODIC TABLE.

1												13	14	15	16	17	18
1 H	2																2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Tungsten crystallizes in a body-centered cubic structure in which the shortest interatomic distance is 274.1 picometers at 25°C (77°F). The pure metal has a lustrous, silver-white appearance. It possesses the highest melting point, lowest vapor pressure, and the highest tensile strength at elevated temperature of all metals. Some important physical properties of tungsten are compiled in the **table**.

At room temperature tungsten is chemically resistant to water, oxygen, most acids, and aqueous alkaline solutions, but it is attacked by fluorine or a mixture of concentrated nitric and hydrofluoric acids.

Tungsten is used widely as a constituent in the alloys of other metals, since it generally enhances high-temperature strength. Several types of tool steels and some stainless steels contain tungsten. Heat-resistant alloys, also termed superalloys, are nickel-, cobalt-, or iron-based systems containing varying amounts (typically 1.5–25 wt. %) of tungsten. Wear-resistant alloys having the trade name Stellites are composed mainly of cobalt, chromium, and tungsten. *See* ALLOY; HIGH-TEMPERATURE MATERIALS.

The major use of tungsten in the United States is in the production of cutting and wear-resistant materials. Tungsten carbides (representing 60% of total tungsten consumption) are used for cutting tools, mining and drilling tools, dies, bearings, and armor-piercing projectiles.

Unalloyed tungsten (25% of tungsten consumption) in the form of wire is used as filaments in incandescent and fluorescent lamps, and as heating elements for furnaces and heaters. Because of its high electron emissivity, thorium-doped (thoriated) tungsten wire is employed for direct cathode electronic filaments. Tungsten rods find use as lamp filament supports, electrical contacts, and electrodes for arc lamps.

Tungsten compounds (5% of tungsten consumption) have a number of industrial applications. Calcium and magnesium tungstates are used as phosphors in fluorescent lights and television tubes. Sodium tungstate is employed in the fireproofing of fabrics and in the preparation of tungsten-containing dyes and pigments used in paints and printing inks. Compounds such as WO_3 and WS_2 are catalysts for various chemical processes in the petroleum industry. Both WS_2 and WSe_2 are dry, high-temperature lubricants. Other applications of tungsten compounds have been made in the glass, ceramics, and tanning industries.

Miscellaneous uses of tungsten account for the remainder (2%) of the metal consumed.

Charles Kutal

Physical properties of tungsten

Property	Value
Melting point	3410 ± 20°C (6170 ± 36°F)
Boiling point	5700 ± 200°C (10,300 ± 360°F)
Density, 27°C (81°F)	19.3 g/cm^3 (11.2 oz/in.^{3})
Specific heat, 25°C (77°F)	0.032 cal/g-°C (0.13 J/g-°C)
Heat of fusion	52.2 ± 8.7 cal/g (218 ± 36 J/g)
Vapor pressure 2027°C (3681°F)	6.4×10^{-12} atm (6.5×10^{-7} Pa)
3382°C (6120°F)	2.3×10^{-5} atm (2.3 Pa)
5470°C (9878°F)	0.53 atm (5.4×10^4 Pa)
Electrical resistivity 27°C (81°F)	5.65 microhm-cm
1027°C (1881°F)	34.1
3027°C (5481°F)	103.3
Thermal conductivity 27°C (81°F)	0.43 cal/cm-s-°C (1.8 J/cm-s-°C)
1027°C (1881°F)	0.27 (1.1)
Absorption cross section, 0.025-eV neutrons	18.5 ± 0.5 barns (18.5 ± 0.5 × 10^{-24} cm^2)

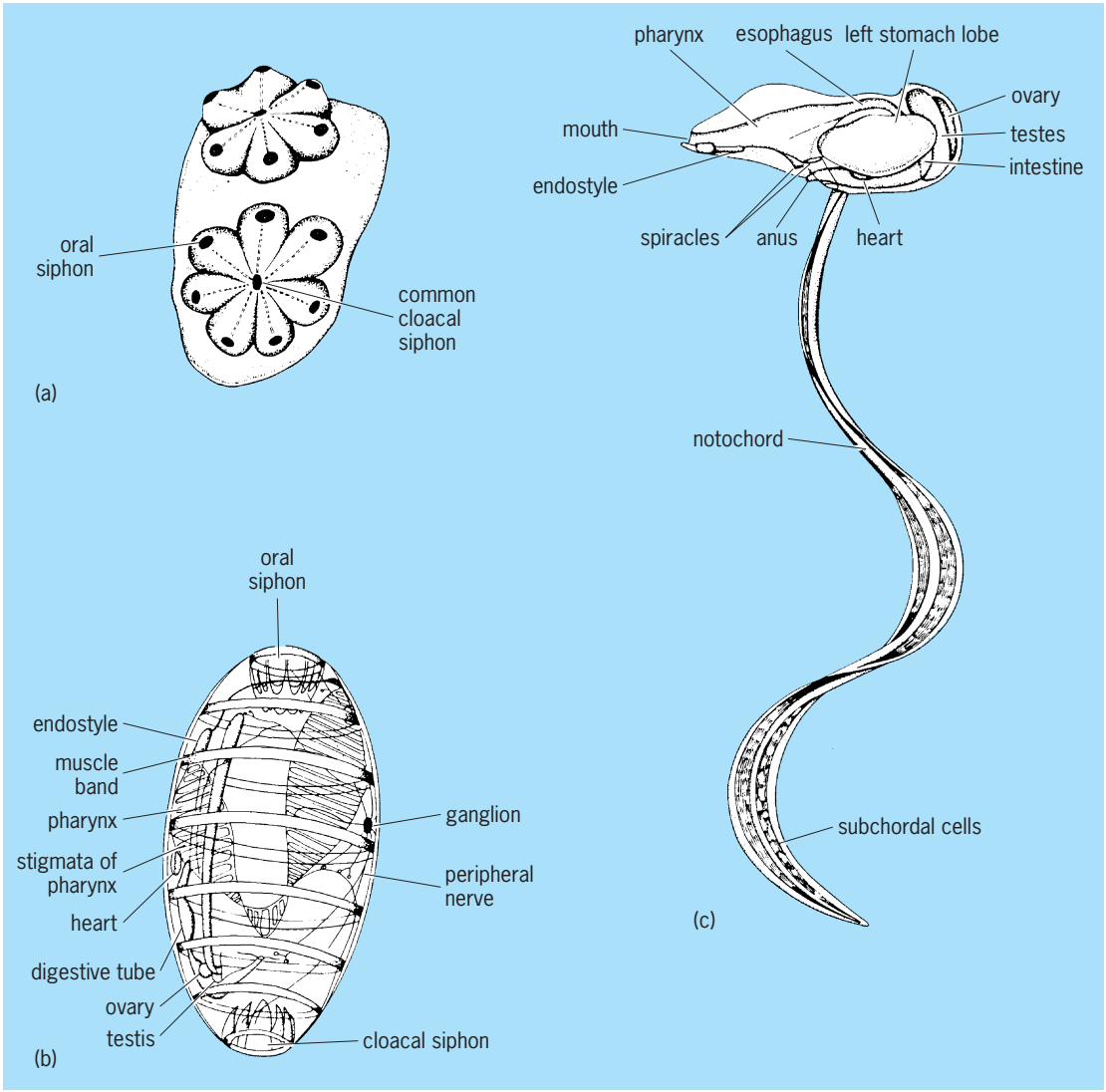
Bibliography. F. A. Cotton and G. Wilkinson, *Advanced Inorganic Chemistry*, 6th ed., 1999; E. Lassner and W.-D. Schubert, *Tungsten: Properties, Chemistry, Technology of the Element, Alloys, and Chemical Compounds*, 1999; E. Pink and L. Bartha (eds.), *The Metallurgy of Doped Non-Sag Tungsten*, 1989.

Tunicata (Urochordata)

A subphylum of marine animals of the Chordata. They are characterized by a perforated pharynx or branchial sac used for food collection, a dorsal notochord restricted to the tail of the larva (and the adult in one class), absence of mesodermal segmentation or a recognizable coelom, and secretion of an outer covering (the test or tunic) which contains large amounts of polysaccharides related to cellulose (see **illustration**).

Three classes are usually recognized: the sessile Ascidiacea (sea squirts or ascidians); planktonic Thaliacea (salps, doliolids, and pyrosomids); and Appendicularia, minute planktonic forms with tails living inside a specialized test or house adapted for filtering and food gathering. Approximately 2000 species of Tunicata are recognizable. The group is found in all parts of the ocean.

Except for a few deep-sea carnivorous forms, tunicates feed on minute plankton and finely divided organic detritus. Food drawn into the pharynx in water currents created by ciliary or muscular activity is filtered on a mucous sheet secreted in the floor of the pharynx and rolled into a cord on the dorsal side. The food cord is passed into the esophagus and from there to the rest of the alimentary canal. Digestion is extracellular; feces are discharged into the outgoing current of water. There are no excretory organs, and most nitrogenous wastes are removed in soluble form. Concentrations of urates



Representative Tunicata. (*a*) *Botryllus*, a compound ascidian; (*b*) *Doliolum*, a planktonic thaliacean (*after P. A. Meglitsch, Invertebrate Zoology, Oxford University Press, 1967*). (c) *Oikopleura*, a planktonic larvacean, removed from its house (*after A. Alldredge, Appendicularians, Sci. Amer., 235:94–102, 1976*).

may also accumulate throughout life in closed vesicles. The heart is tubular and reverses the direction of its beat at intervals, alternately driving blood in opposite directions through the body organs and through the pharyngeal wall. Heavy metals, usually vanadium or iron, accumulate in specialized cells in the blood; these are probably concerned with test formation or chemical defense, and are not respiratory carriers. The nervous system is simple and consists of a solid dorsal ganglion and a few peripheral nerves. The ganglion is closely associated with a neural gland discharging into the entrance of the pharynx.

Most tunicates are hermaphroditic and may be oviparous or viviparous. The larva is a minute (about 0.04 in. or 1 mm) tailed tadpole with a notochord, dorsal nerve cord, and sense organs. Larvae are free swimming for a short period and, except in Appendicularia, lose the tail and notochord at metamorphosis into an adult. Ascidians may be solitary or form colonies; Thaliacea form aggregates or colonies by asexual budding and may alternate between solitary and aggregate forms. Appendicularia are always solitary.

Tunicates have little economic importance except as fouling organisms. A few species have pharmacological properties, and a few larger ascidians are used for food. *See* APPENDICULARIA (LARVACEA); ASCIDIACEA; CHORDATA; THALIACEA. Ivan Goodbody

Bibliography. A. Alldredge, Appendicularians, *Sci. Amer.*, 235:94–102, 1976; N. J. Berrill, *The Tunicata*, 1950; P. Brien and W. Harant, Embranchement des tuniciers, in P. P. Grassé (ed.), *Traité de Zoologie*, vol. 11, 1948; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982.

Tuning

The process of adjusting the frequency of a vibrating system to obtain a desired result. The term is applicable to a wide variety of such systems, but is most commonly used in connection with musical instruments, electronic circuits, and machinery.

Musical instruments. The frequency (pitch) of stringed instruments is determined by the length, mass, and tension of vibrating strings. In members of the violin family, for example, tuning is accomplished by adjusting the tension of each string individually. In keyboard instruments, the process of tension adjustment must achieve a tempered scale in each octave throughout the range of the keyboard. The pitch of wind instruments depends on the volume and shape of a vibrating column of air. Tuning is accomplished by adjusting the shape of the column, usually its length. *See* MUSICAL ACOUSTICS; MUSICAL INSTRUMENTS.

Electronic circuits. In electronic circuits, there are a variety of frequency-determining elements. The most widely used is a combination of an inductance L (which stores energy in a magnetic field) and a capacitance C (which stores it in an electric field). The frequency of oscillation is determined by the rate of exchange of the energy between the two fields, and is inversely proportional to $\sqrt{LC}$. Tuning is accomplished by adjusting the capacitor or the inductor until the desired frequency is reached. The desired frequency may be one that matches (resonates with) another frequency. This occurs when a piano tuner adjusts a string's tension until its vibration agrees with that of a tuning fork, and when a radio receiver is tuned to a desired station. Another purpose of tuning may be to match a frequency standard, as when setting an electronic watch to keep accurate time. The frequency-determining element in such watches, as well as in radio transmitters, digital computers, and other equipment requiring precise frequency adjustment, is a vibrating quartz crystal. The frequency of vibration of such crystals can be changed over a narrow range by adjusting a capacitor connected to it. *See* QUARTZ CLOCK; RADIO RECEIVER; RESONANCE (ALTERNATING-CURRENT CIRCUITS).

At extremely high radio frequencies (in the range above a gigahertz, or 10^9 vibrations per second), the inductance and capacitance are not discrete elements, but are distributed over the surface of the apparatus to be tuned. An example is the microwave cavity, in which electromagnetic waves are trapped and reflected. The cavity is tuned by a plunger or other movable element that changes the volume of the cavity. *See* CAVITY RESONATOR; MICROWAVE.

Another function of tuning in electronics is the elimination of undesired signals. Filters for this purpose employ inductors and capacitors, or crystals. The filter is tuned to the frequency of the undesired vibration, causing it to be absorbed elsewhere in the circuit. *See* ELECTRIC FILTER.

Automatic tuning by electrical control is accomplished by a varactor diode. This is a capacitor whose capacitance depends on the direct-current (dc) voltage applied to it. The varactor serves as a portion of the capacitance of the tuned circuit. Its capacitance is controlled by a dc voltage applied to it by an associated circuit, the voltage and its polarity depending on the extent and direction of the mismatch between the desired frequency and the actual frequency. *See* VARACTOR.

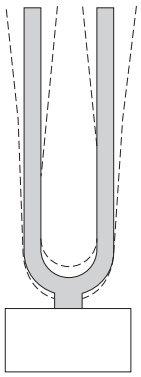
The extremely high-frequency electromagnetic waves associated with visible, infrared, and ultraviolet radiation could not be tuned in the manner of the longer radio waves until the advent of the laser. The very narrow spectrum regions of radiation produced by lasers can be tuned by various means, such as temperature control or the insertion of prisms or gratings in the laser cavity. *See* LASER.

Machinery. Tuning is a technique also applied to the vibrations of machinery. The vibrating elements are typically weights associated with springs, or their equivalent inertial mass and restoring force. Suppression of undesired vibration is the usual aim of mechanical tuning, by the mechanical equivalent of an electrical filter. *See* RESONANCE (ACOUSTICS AND MECHANICS). Donald G. Fink

Tuning fork

A steel instrument consisting of two prongs and a handle which, when struck, emits a tone of fixed pitch. Because of their simple mechanical structure, purity of tone, and constant frequency, tuning forks are widely used as standards of frequency in musical acoustics. In its electrically driven form, a tuning fork serves to control electric circuits by producing frequency standards of high accuracy and stability.

A tuning fork is essentially a transverse vibrator. The amplitude of longitudinal vibration at the end of the stem is small compared with the amplitude of the transverse vibrations at the ends of the prongs



Tuning fork vibrating at its fundamental frequency.

(see **illus.**). Thus, when the stem of the fork is pressed on a sounding board or a resonance box, the vibrations persist for a considerable time since the small-amplitude vibrations transfer energy to the sounding board at a low rate.

Tuning forks are constructed to cover the entire range of audible frequencies from 20 to 20,000 Hz. The frequency of a fork varies approximately as the inverse square of the length and directly as the thickness of the prongs. The sound output of the fundamental frequency of a fork may be reinforced by attaching the stem to an air column type of resonance box having the same fundamental frequency. *See* MUSICAL ACOUSTICS; VIBRATION.

Lawrence E. Kinsler

Tunnel

An underground space of substantial length, usually having a tubular shape. Tunnels can be either constructed or natural and are used as passageways, storage areas, carriageways, and utility ducts. They may also be used for mining, water supply, sewerage, flood prevention, and civil defense.

Construction. Tunnels are constructed in numerous ways. Shallow tunnels are usually constructed by burying sections of tunnel structures in trenches dug from the surface. This is a preferred method of tunneling as long as space is available and the operation will not cause disturbance to surface activities. Otherwise, tunnels can be constructed by boring underground. Short tunnels are usually bored manually or by using light machines (such as a roadheader or backhoe). If the ground is too hard to bore, a drill-and-blast method is frequently used. For long tunnels, it is more economical and much faster to use tunneling boring machines which work on the full face (complete diameter of the opening) at the same time. In uniform massive rock formations without fissures or joints, tunnels can be bored without any temporary supports to hold up the tunnel crowns. However, temporary supports are usually required because of the presence of destabilizing fissures and joints in the rock mass (**Fig. 1**). A layer of shotcrete serves as the primary lining to protect the newly exposed surface and to support the tunnel crowns as well. The shotcrete is frequently reinforced by steel meshes and, if necessary, braced by steel lattices. *See* DRILLING, GEOTECHNICAL.

In soft ground, it has become popular to use shield machines for boring and reinforced concrete segments for lining. The largest shield machine ever, 14 m (46 ft) in diameter, was used in constructing the Tokyo Trans-Bay Highway (Tokyo Wan Aqua-Line) between Kawasaki City and Kisarazu City of Japan (completed in 1997). There are various types of shield machines available to serve different purposes. The multiface shield machine (**Fig. 2**) was first used in constructing the Osaka Business Park subway station in Japan (completed in 1995). The machine is 17 m (56 ft) in width and 7.5 m (25 ft) in height and has three cutters which operate independently. Theoretically, the two side cutters can be detached from the center cutter upon the completion of station excavation, leaving the center cutter to continue boring toward the next station. However, this technique was not applied in this case, and the machine was used for the Osaka Business Park station only.

Immersed tunnels. For tunnels to be constructed across bodies of water, an alternative to boring is to lay tunnel boxes directly on the prepared seabed. These boxes, made of either steel or reinforced concrete, are prepared in dry docks and sealed at their ends by the use of bulkheads. They float as the docks are flooded, and are towed to the site by tugboats. The boxes are then flooded to allow them to sink to the seabed after they are properly positioned. Immersed tunnels are usually buried in shallow trenches dug for this purpose and covered by ballast so they will not be affected by the movement of the water. The joints between tunnel sections are made watertight by using rubber gaskets, and water is pumped out of the tunnel to make it ready for service. Among the numerous immersed tunnels the Øresund Link, completed in 1999 [3.5 km (2.2 mi)], between Denmark and Sweden is second in length to the cross-bay tunnel [5.8 km (3.6 mi)] for the Bay Area Rapid Transit system of San Francisco, California. *See* CONCRETE; STEEL.

Microtunnels. Small tunnels, such as sewer lines and water mains, are usually installed by jacking steel or concrete pipes into the ground (**Fig. 3**). The soil core inside the tubes can be removed manually or by

using "moles," which are essentially small shield tunneling machines. The alignment of the pipes is continuously monitored and adjusted. With the moles guided by a computerized navigation system, it is possible to align pipes to a precision within 100 mm (4 in.) regardless of length. This technique has been used for jacking pipes as large as 2 m (7 ft) or so in diameter for distances more than 100 m (330 ft). Pipes with smaller sizes can be jacked to distances of more than 300 m (990 ft).

Special tunneling techniques. Auxiliary measures are frequently required to ensure the safety of tunnels during boring in soft ground. Compressed air was used in the past, but is it is seldom used anymore because improper decompression may cause aeroembolism (diver's disease) in workers. Instead, grouting and ground freezing are now preferred. In the Central Artery Project in Boston, Massachusetts, ground freezing is carried out to permit three tunnels to be bored under the railway tracks leading to the South Station Railway Terminal. Roughly 1600 freezing pipes are installed to depths varying from 13.7 to 16.8 m (45 to 55 ft), and the volume of frozen soil is more than 60,000 m^3 (2,100,000 ft^3). This could be the largest undertaking of this nature.

An underpass, scheduled to be completed in 2001, is being constructed beneath the Taipei International Airport by using the Endless-Self-Advancing method. To minimize ground settlements and ensure the safety of air traffic, interlocked steel pipes are first jacked into the ground to form a protective shelter. The soil core inside the shelter is excavated at the rate of 400 mm (16 in.) per lift. Concrete segments are moved one by one into the space created, by jacking the segments behind. Each time only the last segment is jacked forward, and the jacking force is taken by the frictional resistance acting on all the rest of the segments. Cables are anchored to the first and last segments so the force acting on the last segment can be transmitted to the segments in front of it. The movements of these segments resemble the movements of centipedes. In theory, there is no limit on the length of tunnels installed by this method.

Longest tunnels. The longest tunnel of any kind is the New York City/West Delaware water supply tunnel (completed in 1944). It runs for 169 km (105 mi) from the Rondout Reservoir into the Hillview Reservoir in Yonkers, New York. The longest rail tunnel is Seikan Tunnel (53.6 km; 33.3 mi) in Japan (completed in 1985). The longest road tunnel was Saint Gotthard Tunnel (16.9 km; 10.5 mi) in Switzerland (completed in 1980). In mid2001 the title was taken over by the Laerdal Tunnel (24.5 km; 15.2 mi).

The longest undersea tunnel is the Channel Tunnel [49.4 km (30.7 mi), of which 38 km (23.6 mi) is undersea] across the English Strait. It runs from Folkestone in Britain to Calais in France. There are two running tunnels plus one service tunnel in the center, 7.6 and 4.8 m (25 and 16 ft) in internal diameter, respectively. The tunnel was officially inaugurated on May 6, 1994, when the Queen of England and President Mitterrand of France became the first official passengers to pass by train between the two countries. Eurotunnel has a concession from the British and French governments to run the tunnel until 2052.

Fig. 1. Temporary supports for a four-lane road tunnel. (*RESA Engineering Corporation*)

Fig. 2. Multiface shield tunneling machine. (*Hitachi Zosen Corporation*)

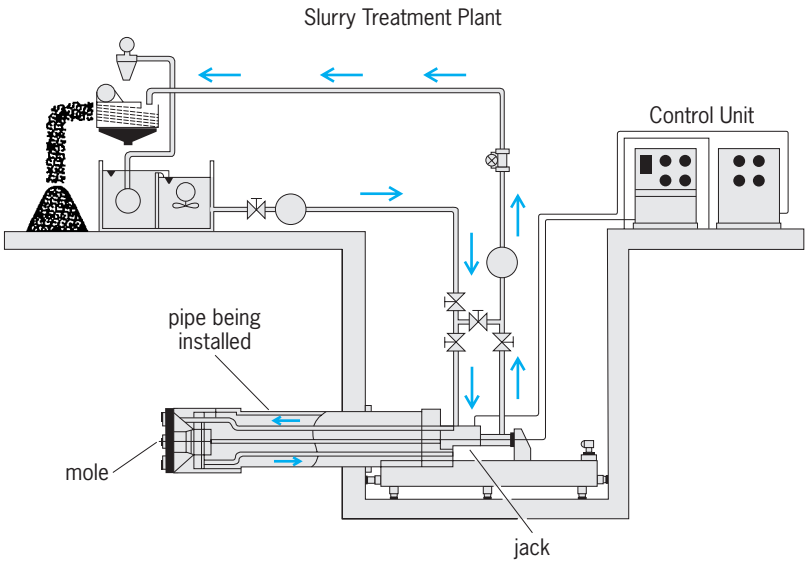


Fig. 3. Microtunneling and pipe jacking technique.

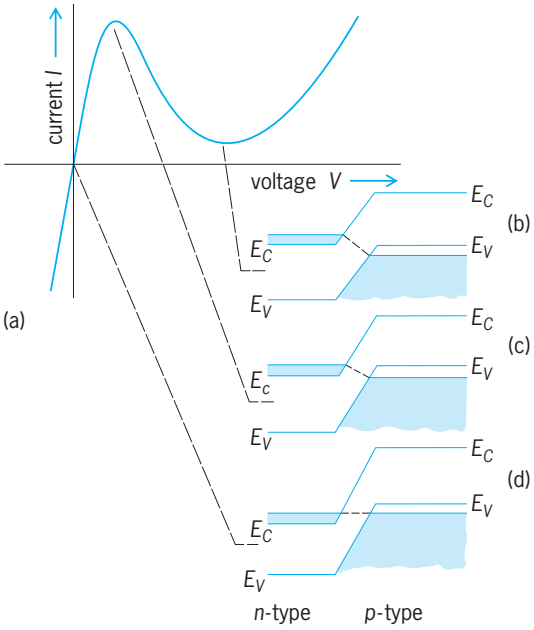
Shuttle trains, carrying up to 180 cars, and freight shuttles, carrying 28 lorries, will take 35 minutes to cross the strait. Za-Chieh Moh; Richard N. Hwang

Bibliography. C. J. Kirkland (ed.), *Eurotunnel: Engineering the Channel Tunnel*, 1995; J. O. Bickel, T. R. Kuesel, and E. H. King (eds.), *Tunnel Engineering Handbook*, 1995; Z. C. Moh et al., Underpass beneath Taipei International Airport, *Proceedings of the Conference on New Frontiers and Challenges*, Bangkok, Thailand, November 8–12, 1999; A. M. M. Wood and A. M. Wood, *Tunnelling: Management by Design*, 2000.

Tunnel diode

A two-terminal semiconductor junction device (also called the Esaki diode) which does not show rectification in the usual sense, but exhibits a negative resistance region at very low voltage in the forward-bias characteristic and a short circuit in the negative-bias direction.

The short-circuit condition exists because both the *p* and *n* regions of the device are doped with such high concentrations of the appropriate impurities that the normal barrier is rendered sufficiently thin to allow the free passage of current at zero and all negative-bias conditions. The forward-bias characteristic (**illus.** *a*) shows a maximum and a minimum in the current with a negative-resistance region between. Band-potential diagrams show the internal electronic situation existing at the current minimum (illus. *b*), the current maximum (illus. *c*), and zero bias (illus. *d*). The top of the shaded regions of the band-potential diagrams shows the level to which electrons fill the available energy levels in the valence and conduction bands of the materials forming the *pn* junction. The bottom of the conduction band is designated E_c and the top of the valence band E_v. No electrons can penetrate the forbidden energy gap between E_c and E_v except in the barrier region, where it is thin enough to allow electron transit by tunneling.



Tunnel diode characteristic. (*a*) Forward-bias voltage-current plot. (*b*) Band-potential diagram for the current minimum. (*c*) Band-potential diagram for the current maximum. (*d*) Band-potential diagram for zero bias.

The observed characteristic may be accounted for as follows. In illus. *d* the electron level is the same on both sides of the junction (zero bias). No net current flows because there is no difference in electronic energy across the junction. As forward bias is applied, tunneling current will flow since now the electrons in the *n*-type material on the left will rise to a level above those on the right (illus. *c*). As long as these electrons are still below the top of the valence band E_v on the right, current will increase. When the top of the elevated electron distribution exceeds the level of E_v, current will begin to decrease and the diode has entered the negative-resistance region. In illus. *c* the top of the electron distribution on the left is even with E_v on the right, and any further increase in bias will reduce the number of electrons available for the tunneling current. Therefore, illus. *c* corresponds to the current maximum. As bias continues to increase, the point is reached where the bottom of the conduction band E_c on the left is even with E_v on the right, and the entire electron distribution, being above E_v, is removed from the tunneling process. At this point, only normal forward-bias diffusion current flows. This current is composed of energetic electrons diffusing over the top of the barrier while remaining in the conduction band. Illustration *b* shows the point at which tunneling is no longer possible and thus corresponds to the current minimum at the end of the negative-resistance region. For further discussion of the properties of junction diodes *see* JUNCTION DIODE; SEMICONDUCTOR; ZENER DIODE. Lloyd P. Hunter

Bibliography. H. Mizuta and T. Tanoue, *The Physics and Applications of Resonant Tunnelling Diodes*, 1995; J. Singh, *Semiconductor Devices: An Introduction*, 1994; E. S. Yang, *Microelectronic Devices*, 1988.

Tunneling in solids

A quantum-mechanical process which permits electrons to penetrate from one side to the other through an extremely thin potential barrier to electron flow. The barrier would be a forbidden region if the electron were treated as a classical particle. A two-terminal electronic device in which such a barrier exists and primarily governs the transport characteristic (current-voltage curve) is called a tunnel junction. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

During the infancy of the quantum theory, L. de Broglie introduced the fundamental hypothesis that matter may be endowed with a dualistic nature—particles such as electrons, particles, and so on, may also have the characteristics of waves.

This hypothesis found expression in the definite form now known as the Schrödinger wave equation, whereby an electron or an alpha particle is represented by a solution to this equation. The nature of such solutions implies an ability to penetrate classically forbidden regions of negative kinetic energy and a probability of tunneling from one classically allowed region to another. The concept of tunneling, indeed, arises from this quantum-mechanical result. The subsequent experimental manifestations of this concept, such as high-field electron emission from cold metals, alpha decay, and so on, in the 1920s, can be regarded as one of the early triumphs of the quantum theory. *See* FIELD EMISSION; RADIOACTIVITY; SCHRÖDINGER'S WAVE EQUATION.

In the 1930s, attempts were made to understand the mechanism of electrical transport in resistive contacts between metals and rectifying metal-semiconductor contacts in terms of electron tunneling in solids. In the latter case, since a proposed theoretical model did not properly represent the actual situation, the theory predicted the wrong direction of rectification. In many cases, however, conclusive experimental evidence of tunneling was lacking, primarily because of the rudimentary stage of material science.

Tunnel diode. The invention of the transistor in 1947 spurred the progress of semiconductor technology. By the 1950s, materials technology for semiconductors such as Ge and Si was sufficiently advanced to permit the construction of well-defined semiconductor structures. The tunnel diode (also called the Esaki diode) was discovered in 1957 by L. Esaki. This discovery demonstrated the first convincing evidence of electron tunneling in solids, a phenomenon which had been clouded by questions for decades. This device is a version of the semiconductor *pn* junction diode which is made of a *p*-type semiconductor, containing mobile positive charges called holes (which correspond to the vacant electron sites), and an *n*-type semiconductor, containing mobile electrons (the electron has a negative charge). Esaki succeeded in making the densities of holes and electrons in the respective regions extremely high by doping a large amount of the appropriate impurities with an abrupt transition from one region to the other. Now, in semiconductors, the conduction band for mobile electrons is separated from the valence band for mobile holes by an energy gap, which corresponds to a forbidden region. Therefore, a narrow transition layer from *n*-type to *p*-type, 5 to 15 nm thick, consisting of the forbidden region of the energy gap, provided a tunneling barrier. Since the tunnel diode exhibits a negative incremental resistance with a rapid response, it is capable of serving as an active element for amplification, oscillation, and switching in electronic circuits at high frequencies. The discovery of the diode, however, is probably more significant from the scientific aspect because it has opened up a new field of research—tunneling in solids. *See* BAND THEORY OF SOLIDS; CIRCUIT (ELECTRONICS); HOLE STATES IN SOLIDS; JUNCTION DIODE; NEGATIVE-RESISTANCE CIRCUITS; SEMICONDUCTOR; SEMICONDUCTOR DIODE; TUNNEL DIODE.

Esaki and colleagues have explored negative resistance phenomena in semiconductors which can be observed in novel tunnel structures. One obvious question is: What would happen if two tunnel barriers are placed close together, or if a periodic barrier structure—a series of equally spaced potential barriers—is made in solids? It has been known that there is a phenomenon called the resonant transmission. Historically, resonant transmission was first demonstrated in the scattering of electrons by atoms of noble gases and is known as the Ramsauer effect. In the above-mentioned tunnel structures, it is clear that the resonant tunneling should be observed. In preparing double tunnel barriers and periodic structures with a combination of semiconductors, the resonant tunneling was experimentally demonstrated and negative resistance effects were observed. *See* SEMICONDUCTOR HETEROSTRUCTURES.

Tunnel junctions between metals. As discussed above, tunneling had been considered to be a possible electron transport mechanism between metal electrodes separated by either a narrow vacuum or a thin insulating film usually made of metal oxides. In 1960, I. Giaever demonstrated for the first time that, if one or both of the metals were in a superconducting state, the current-voltage curve in such metal tunnel junctions revealed many details of that state. At the time of Giaever's work, the first satisfactory microscopic theory of superconductivity had just been developed by J. Bardeen, L. N. Cooper, and J. R. Schrieffer (BCS theory). Giaever's technique was sensitive enough to measure the most important feature of the BCS theory—the energy gap which forms when the electrons condense into correlated, bound pairs (called Cooper pairs).

The tunneling phenomenon has been exploited in many fields. For example, small-area tunnel junctions are used for mixing and synthesis of frequencies ranging from dc to the infrared region of the spectrum. This leads to absolute frequency measurement in the infrared and provides the most accurate determination of the speed of light. *See* PARAMETRIC AMPLIFIER.

To study nonequilibrium superconducting properties, two tunnel junctions, one on top of the other sharing the middle electrode, are used. One junction seems to inject quasi-particles, while the other detects their effects on the important parameters. Tunnel junctions are also used as a spectroscopic tool to study the phonon and plasmon spectra of the metals and the vibrational spectra of complex organic molecules introduced inside the insulating barriers (tunneling spectroscopy). *See* SPECTROSCOPY; SUPERCONDUCTIVITY.

Josephson effects. Giaever's work opened the door to more detailed experimental investigations—it pioneered a new spectroscopy of high accuracy to study the superconducting state. In 1962, B. Josephson made a penetrating theoretical analysis of tunneling between two superconductors by treating the two superconductors and the coupling process

as a single system, which would be valid if the insulating oxide were sufficiently thin, say 2 nanometers. His theory predicted, in addition to the Giaever current, the existence of a supercurrent, arising from tunneling of the bound electron pairs.

This led to two startling conclusions: the dc and ac Josephson effects. The dc effect implies that a supercurrent may flow even if no voltage is applied to the junction. The ac effect implies that, at finite voltage V, there is an alternating component of the supercurrent which oscillates at a frequency of 483.6 MHz per microvolt of voltage across the junction, and is typically in the microwave range. The dc Josephson effect was soon identified among existing experimental results, while the direct observation of the ac effect eluded experimentalists for a few years. The effects are indeed quantum phenomena on a macroscopic scale. Extraordinary sensitivity of the supercurrents to applied electric and magnetic fields has led to the development of a rich variety of devices with application in wide areas of science and technology. Superconducting quantum interference devices (SQUIDs) are made of one or more Josephson junctions connected to form one or more closed superconducting loops. Owing to their unprecedented sensitivity, SQUIDs are the main building blocks of many sensitive instruments such as magnetometers, power meters, voltmeters, gradiometers, and low-temperature thermometers. These are finding wide-range application in the fields of solid-state physics, medicine, mineral exploration, oceanography, geophysics, and electronics. Josephson junction and SQUIDs are used as switches for digital applications. They are the basic elements found in the picosecond-resolution sampling oscilloscope, as well as memory and logic circuits featuring high switching speed and ultralow power dissipation, in the order of 1 microwatt. In the communication field, they are used in analog applications, such as high-frequency local oscillators, detectors, mixers, and parametric amplifiers. Furthermore, the ac Josephson effect is now used to define the volt in terms of frequency in standards laboratories, eliminating the antiquated standard cell. *See* FUNDAMENTAL CONSTANTS; JOSEPHSON EFFECT; SQUID; SUPERCONDUCTING DEVICES.

Leo Esaki

Bibliography. L. L. Chang (ed.), *Resonant Tunneling in Semiconductors: Physics and Applications*, 1992; L. Esaki, Long journey into tunneling, *Science*, 183:1149–1155, 1974; I. Giaever, Electron tunneling and superconductivity, *Science*, 183:1253–1258, 1974; H. Grahn, *Semiconductor Superlattices: Growth and Electronic Properties*, 1995; B. D. Josephson, The discovery of tunneling supercurrents, *Science*, 184:527–530, 1974.

Tupelo

A tree belonging to the genus *Nyssa* of the sour gum family, Nyssaceae. The most common species is *N. sylvatica*, variously called pepperidge, black gum, or sour gum, the authorized name being black tupelo. Tupelo grows in the easternmost third of the United States; southern Ontario, Canada; and Mexico. In moist soil this tree usually ranges from 60 to 80 ft (18–24 m) in height and 2 to 3 ft (0.6–0.9 m) in diameter, but some may be 110 ft (33 m) tall and 5 ft (1.5 m) in diameter. *See* MYRTALES.

Leaf, fruit cluster, bud, and twig of black tupelo.

The tree can be identified by the comparatively small, obovate, shiny leaves (see **illus.**); by branches that develop at a wide angle from the axis; and by a chambered pith. The fruit is a small blue-black drupe, a popular food for birds. The wood is yellow to light-brown and hard to split because of the twisted grain. Tupelo wood has been used for boxes, baskets, and berry crates, and as backing on which veneers of rarer and more expensive woods are glued. It is also used for flooring, rollers in glass factories, hatters' blocks, and gunstocks. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth P. Davis

Turbellaria

A class of the phylum Platyhelminthes commonly known as the flatworms. These animals are chiefly free-living and have simple life histories. The bodies are elongate and flat to oval or circular in cross section. Their length ranges from less than 0.04 in. (1 mm) to several centimeters, but may exceed 20 in. (50 cm) in land planaria. Large forms are often brightly colored. Smaller forms may have black, gray, or brown parenchymal pigment or may be white or transparent except for the color of ingested food and symbiotic algae. This class, which numbers some 3400 described species, is ordinarily subdivided into 12 orders: Acoela, Catenulida, Haplopharyngida, Lecithoepitheliata, Macrostomida,

Nemertodermatida, Neorhabdocoela, Polycladida, Prolecithophora, Proseriata, Temnocephalida, and Tricladida.

Economics and ecology. Turbellaria are not economically important but have proved valuable in the study of such fundamental biological problems as regeneration, metabolism, axial gradients, evolution, and adaptations to parasitism. Although widely distributed in fresh and salt water and moist soil, they are usually overlooked because of their generally small size, secretive habits, and inconspicuous color. They are seldom eaten by other animals but frequently feed on one another and may harbor commensals and parasites, chiefly Protozoa and Nematoda. Some species are themselves parasites or commensals on other aquatic invertebrates.

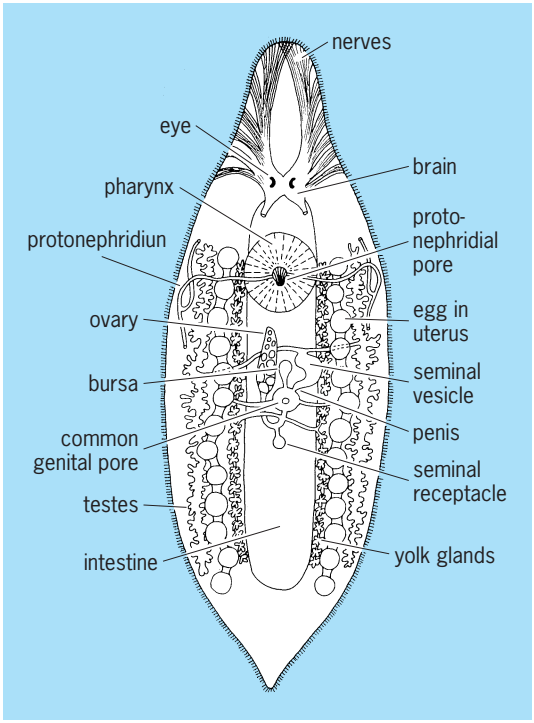
Morphology and physiology. Locomotion is by gliding, swimming, or muscular movements of the body wall. Respiration takes place by diffusion through the body wall since respiratory organs are lacking. The cellular or syncytial epidermis is usually covered with cilia and may contain the nematocysts of ingested coelenterates. Gland cells are of frequent occurrence, producing adhesive substances, mucus, and two types of rod-shaped secretions or rhabdoids. These are the commoner rhabdites and the longer and slenderer rhammites. Usually there is a basement membrane beneath the epidermis, inside which lies an outer circular and an inner longitudinal muscle layer, sometimes with a diagonal layer between. The muscular system also includes both parenchymal and organ muscles. Between the body wall and internal organs lies the parenchyma, a more or less compact network of mesenchyme cells. Plasma in its interstices may function as a circulatory fluid.

Digestive system. The mouth, on the midventral surface, is generally followed by the pharynx which occurs in several forms. The simple pharynx has no limiting membrane and is only slightly eversible or protrusible. The plicate pharynx lies in a deep pouch, has a limiting membrane except at its base, and is highly protrusible. The bulbous pharynx lies in a shallow pouch and is surrounded by a limiting membrane. It is eversible and occurs in two chief forms, the cask-shaped or doliiform located at the anterior end of the intestine parallel to the main body axis, and the globular or rosulate lying ventral to the intestine and perpendicular to the main body axis. A short esophagus is often present and polypharyngy may occur. The intestine may be unbranched or a two, three-, or many-branched sac with or without diverticula. It is lined with tall epithelial cells and generally has no anus. Turbellaria are carnivorous; digestion is largely intracellular.

Excretion. The protonephridia are elongated tubules, usually paired, with ciliated flame bulbs on the lateral branches and one or more external openings. They are often lacking in marine forms and are probably primarily concerned with elimination of water.

Nervous system. In primitive forms the nervous system is an epidermal network with five pairs of longitudinal nerves connected by a nerve ring. Swellings at the intersection of nerve ring and nerves represent the beginnings of a brain. With further development, a pair of cerebral ganglia is formed by fusion of these swellings and the entire system sinks into the parenchyma. Posteriorly the longitudinal nerves are reduced to one well-developed ventral pair and often a much smaller dorsal pair. Intermediate conditions between these two extremes are common. Sensory receptors are located chiefly in the head region, although tactile bristles arise from widely scattered sensory cells. Chemoreceptors consist of depressed areas of epidermis with cilia for circulating water over the sensory surface. These are located in auricles, frontal organs, and ciliated rings, pits, and furrows. Statocysts or organs of equilibrium occur chiefly in primitive marine forms. Many species have one or two pairs of photoreceptors or eyes, and in land planarians and polyclads these may become numerous.

Reproductive system and reproduction. Virtually all Turbellaria are hermaphroditic (see **illus.**). Male and female systems may be separate throughout or may have a common antrum and pore. Genital pores are usually on the midventral surface, sometimes combined with the mouth opening. The saclike gonads are compact or follicular depending upon whether they consist of a few relatively large or several to many small bodies, but some acoeles have only scattered clusters of germ cells in the parenchyma. The male system has a single pair of sperm ducts which may fuse to form a seminal duct, enlarge to form spermiducal or storage vesicles, or empty into a true seminal vesicle. The copulatory organ is usually muscular, often encloses the seminal vesicle, may be armed



Typical hermaphroditic turbellarian, *Mesostoma ehrenbergii wardii*.

with a cuticular apparatus, and contains the ejaculatory duct. When prostate glands are present, their secretions may be stored in a prostatic vesicle but are mixed with the sperm in the seminal vesicle or in the male genital canal. Some Turbellaria produce entolecithal eggs which contain yolk, but generally the eggs are ectolecithal and yolk is derived from yolk cells. These yolk cells are probably degenerate ovocytes which may be produced throughout the gonad, in a specialized section of the gonad, or in special yolk glands which are distinct from the ovary. The female system usually has its own ducts through which its products reach the female antrum and ultimately the exterior. In the absence of ducts, fertilized eggs escape by rupture of the body wall or by way of the digestive system and the mouth. Accessory structures such as vaginae, copulatory bursae, uteri, seminal receptacles, and specialized glands may also occur. Asexual reproduction by fragmentation or by binary fission with the formation of temporary chains of zooids occurs in some rhabdocoeles and triclads. *See* ACOELA; MACROSTOMORPHA; POLYCLADIDA; TRICLADIDA.

E. Ruffin Jones

Bibliography. E. Bresslau, *Turbellaria*, in W. Kükenthal (T. Krumbach, ed.), *Handbuch der Zoologie*, vol. 2, pt. 1, 1933; L. H. Hyman, *The Invertebrates*, vol. 2, 1951; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; W. D. Russell-Hunter, *A Life of Invertebrates*, 1979.

Turbidite

A bed of sediment or sedimentary rock that was deposited from a turbidity current. A turbidity current is an underwater flow produced by movement of a turbid mass of water downslope as a result of the excess weight of the turbid water as compared with the surrounding clear water. Turbidity currents are therefore a type of gravity (or density) current; such currents are kept in motion by gravity acting on relatively small differences in density between different fluids (gases or liquids) or between different parts of the same fluid mass. In some gravity currents the density difference results from a difference in temperature or salinity (for example, fresh water flowing above salt water), but in turbidity currents it results from the presence of dispersed sediment.

Turbidity currents. Many geologists would restrict the term turbidity current to flows in which the sediment dispersion is maintained by fluid turbulence (that is, in which the sediment is in true suspension). In this case, there is a need for terms to describe other types of gravity flows resulting from sediment which is in dispersion, but not as a result of the action of turbulence. The general class of such flows has been called sediment gravity flows, and there appear to be four main kinds: turbidity currents, in which sediment suspension is maintained by turbulence; subaqueous debris flows, in which the sediment grains are held in dispersion by the strength of the (generally muddy) matrix; subaqueous grain flows, in which sediment is dispersed by forces acting between the grains themselves (dispersive pressure) produced by shearing of a concentrated mass of grains; and liquefied sediment flows, in which the sediment is supported by a transient upward flow of pore fluids. Deposits of all four types of flows are generally closely associated, and it is not easy to reconstruct the exact mechanism of deposition from examination of the deposit. The term turbidite may therefore be used by some geologists in a broad sense to include deposits from all four types of sediment gravity flows.

The concept of turbidity currents as agents for the transportation and deposition of sandy (or coarser) sediment was introduced and developed mainly by P. H. Kuenen in the 1950s. Most geologists accept that a large part of the sediments both in the modern ocean basins and in the ancient stratigraphic record is composed of turbidites, even though direct observation of natural turbidity currents is extremely difficult. Evidence derived both from detailed topographic surveys of the modern continental slopes and rises, and from core samples of the surface sediments, indicates that sediment gravity flows are generated from time to time by failure of masses of sediment that have been deposited (in relatively shallow water) on the upper parts of submarine slopes. Slumping or liquefaction of sediments produces a mass of moving sediment which mixes with the surrounding water, accelerates as it moves downslope, and generally becomes a fully turbulent turbidity current. Most such currents move downslope in well-defined submarine canyons or channels, which have probably been eroded or constructed by a long series of previous flows. At the base of the slope, the turbidity currents leave the main channel to construct large submarine fans composed largely of turbidites, alternating with finer-grain sediments deposited by slow settling of mud from suspension (hemipelagic sediments). The most powerful turbidity currents may carry the finer sand and mud for many hundreds of kilometers along the sea floor, to construct the great flat submarine plains known as abyssal plains.

The most striking example of a turbidity current whose effects were actually recorded was the flow triggered by the earthquake that in 1929 struck the continental slope just south of the Grand Banks of Newfoundland. Cables crossing the North Atlantic were broken not only at the time of the earthquake near the epicenter, but also in sequence over a period of 12 h for a distance of several hundred kilometers to the south. The sequence of breaks indicates that the cables south of the epicenter were broken not by the effect of the earthquake itself, but by a large turbidity current which traveled downslope away from the epicenter at speeds of between 10 and 45 mi/h (5 and 20 m/s). Probably most turbidity currents are not of this magnitude, but it is not unusual to find individual turbidite beds that are more than a meter in thickness, and in several well-documented cases such beds have been traced laterally for distances of several tens of kilometers. *See* TURBIDITY CURRENT.

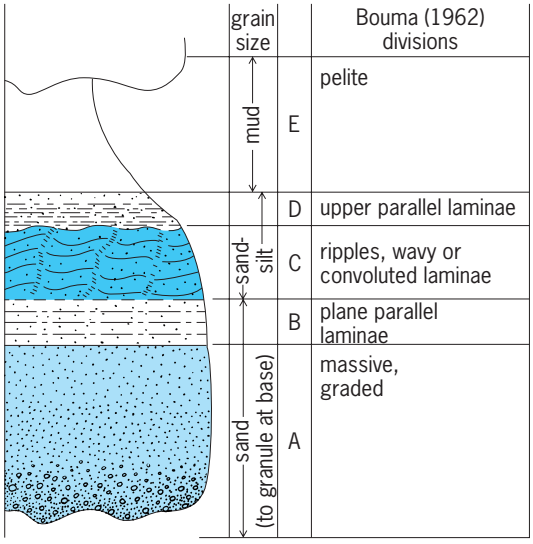
Characteristics. The term turbidite is fundamentally genetic and interpretive in nature, rather than being a descriptive term (like common rock names). Turbidites are clastic sedimentary rocks, but they may be composed of silicic grains (quartz, feldspar, rock fragments) and therefore be a type of sandstone, or they may be composed of carbonate grains and therefore be a type of limestone. A geologist's description of a rock as a turbidite is actually an expression of an opinion that the rock was deposited by a turbidity current, rather than being a description of a particular type of rock.

A combination of inference from laboratory experiments and field observations suggests that many turbidites display several, if rarely all, of a group of characteristics:

They show a sharp base and a gradational top, and there is not infrequently a progressive decline in mean grain size from the base to the top (graded bedding).

Examination of the exposed base of the bed (known as the sole) reveals a variety of structures (sole marks) that were originally produced by erosion of the mud over which the turbidity current moved. The marks cut by the head of the flow were then rapidly covered by sediment deposited by the body of the same current, and in some cases the marks have been distorted by sinking of sediment into the originally soft muds below the turbidite (to form load casts). Common types of sole marks include those scoured by the flow (scour marks), and of these the most characteristic form is scooplike, elongated parallel to the current that cut it, with the blunt end on the upcurrent side. Such structures are called flutes (or flute casts where preserved as molds on the sole of the turbidite that was deposited above them), and they are very useful for indicating the direction of flow of ancient turbidity currents. Also, presence of abundant scour marks suggests that the current which formed them was fully turbulent, and that the bed above was deposited by a turbidity current and not by some other type of (nonturbulent) sediment gravity flow. A second main type of sole marks are those that are cut by large fragments carried in the lower part of the flow (tool marks); the most common are long straight sets of grooves, parallel to the flow direction.

There is commonly a characteristic (Bouma) sequence of sedimentary structures within the bed (see **illus.**). Each characteristic structure (or group of structures) defines a division in the sequence and is identified by a letter: (a) massive (that is, no internal structures) except for some size grading; (b) plane lamination, parallel to the base of the bed (in many beds this lamination is faint and indistinct); (c) cross-lamination formed by migration of small-scale ripples, or convolute lamination formed by the distortion of ripple cross-lamination during, or soon after, its deposition; (d) an upper division of plane lamination, grading up into the next division; (e) muddy sediment (pelite). In most turbidites, only some of the Bouma divisions are represented, but the order in which the divisions occur is only very rarely reversed. Where one turbidite rests directly on another, several of the upper divisions of the first turbidite may have been eroded by the second turbidity current before the second turbidite was deposited, so that it appears at first as though there is only a single bed. Careful examination shows the erosion surface (and generally a size break between the turbidite below and the base of the one above). Such turbidites are said to be amalgamated into a single (composite) bed.



Ideal sequence of structures in a turbidite. (*After H. Blatt, G. Middleton, and R. Murray, Origin of Sedimentary Rocks, 2d ed., Prentice-Hall, 1980*)

Because many (if not all) turbidites have been very rapidly deposited, perhaps over a period of a few hours or days, they represent catastrophic events. Organisms originally living on the mud bottom may be swept away or buried. If buried, they may burrow to the surface (forming escape burrows), but thick turbidites are rarely extensively disturbed by burrowing organisms, even if the muds between them indicate that such organisms were common in the environment of deposition.

Sediment deposited rapidly is liable to liquefaction and subsequent expulsion of water to form fluid escape structures, and these are more commonly formed in turbidites than in other types of sediments.

The fossils found in turbidites are commonly broken, abraded, and displaced from shallower water, as compared with the fauna indigenous to the environment of deposition. In the case of geologically recent turbidites (Mesozoic and Cenozoic), the microfossils in the muds interbedded with turbidite beds not infrequently indicate great depths of deposition (several thousands of meters). Thick sand beds would normally be taken to indicate deposition in shallow water, so identification of the beds as turbidites is important for interpretation of the geological history of the deposits.

Identification. No single feature of a deposit is sufficient to identify it as a turbidite; all of the features

above have been described from other types of beds. A correct interpretation can be made only after analyzing all of the evidence together. Furthermore, it must be emphasized that not all turbidites are marine; there are well-documented examples of modern turbidity currents and turbidites described from lakes. Although probably most turbidites were originally deposited in water of considerable depth (hundreds to thousands of meters), it is generally difficult to be specific about estimating the depth of deposition. The most that can be said is that (in most cases) there is no sign of sedimentary structures formed by the action of waves.

Flysch. Turbidites and interbedded shales form large parts of the stratigraphic succession exposed on the continents. Many such thick turbidite-shale units were called flysch by geologists working in the era before plate tectonics (roughly before the mid-1960s). Some flysch units probably represent the deposits of ancient continental slopes and rises, but many are related to smaller basins of deposition formed during the early stages of plate subduction or continental collision, and a few may be fillings of deep-sea trenches. Many flysch formations have been strongly deformed and even metamorphosed subsequent to deposition. A few turbidite formations, however, were deposited along with suitable source beds in basins which were later raised to moderate temperatures 120–300°F (50–150°C), but which escaped strong deformation or metamorphism, and these have subsequently become prolific producers of oil or gas. Examples include several fields that produce from Cretaceous or Cenozoic sandstones in the Ventura and Great basins of California. *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; SEDIMENTARY ROCKS; SEDIMENTOLOGY.

G. V. Middleton

Bibliography. B. J. Skinner and S. C. Porter, *The Dynamic Earth: An Introduction to Physical Geology*, 5th ed., 2003; D. A. V. Stow (ed.), *Deep Water Turbidite Systems*, 1991; R. Tinterri et al. (eds.), *An Introduction to the Analysis of Ancient Turbidite Basins from an Outcrop Perspective*, 1999.

Turbidity current

A flow of water laden with sediment that moves downslope in an otherwise still body of water. The driving force of a turbidity current is obtained from the sediment, which renders the turbid water heavier than the clear water above. Turbidity currents occur in oceans, lakes, and reservoirs. They may be triggered by the direct inflow of turbid water, by wave action, by subaqueous slumps, or by anthropogenic activities such as dumping of mining tailings and dredging operations.

Turbidity currents are characterized by a well-defined front, also known as head, followed by a thinner layer known as the body of the current. They are members of a larger class of stratified flows known as gravity or density currents. A simple density current is driven by density differences produced by salinity or temperature. Salt and heat are conservative contaminants; they may be advected or diffused, but their total amount in the flow is conserved. Sediment is in general a nonconservative contaminant; it can be entrained from or deposited on the bed, thus changing the total amount of sediment in suspension. *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS.

A turbidity current must generate enough turbulence to hold its sediment in suspension. If it is not able to do so, the sediment deposits and the current dies. The extra degree of freedom, that is, sediment entrainment and deposition, is what makes turbidity currents an interesting phenomenon. Under certain conditions, a turbidity current might erode its bed, pick up sediment, become heavier, accelerate, and pick up even more sediment, increasing its driving force in a self-reinforcing cycle akin to the formation of a snow avalanche.

Observations of turbidity currents were made in lakes and anthropogenic reservoirs long before their occurrence in the ocean became apparent. For many years, the only direct evidence of turbidity current activity in the ocean was the failure of submarine cables, such as the one associated with the Grand Banks earthquake of 1929. Following the earthquake, a large number of transoceanic telegraph cables were broken. At first, it was believed that the earthquake itself had caused the damage. However, 23 years after the earthquake an examination of the timing of the failures indicated that a turbidity current triggered by the earthquake had been responsible for the breakage.

Turbidity currents constitute a major mechanism for the transport of fluvial, littoral, and shelf sediments onto the ocean floor. These flows are considered to be responsible for the scouring of submarine and sublacustrine canyons. These canyons are often of massive proportions and rival the Grand Canyon in scale; they may be eroded directly into deltaic deposits, as in the case of the Rhone River delta in Lake Geneva, Switzerland, or into the continental shelf as in the case of Scripps and La Jolla submarine canyons in California. Only flows of substantial velocity could accomplish this excavation. The conditions under which the generation of such swift, canyon-scouring currents becomes possible have intrigued marine geologists and oceanographers for many years. Below the mouths of most canyons, turbidity currents form vast depositional fans that have many of the features of alluvial fans built by rivers and constitute major hydrocarbon reservoirs. The sedimentary deposits created by turbidity currents, known as turbidites, are a major constituent of the geological record. *See* MARINE GEOLOGY; MARINE SEDIMENTS; SUBMARINE CANYON; SUBMARINE FAN; TURBIDITE.

Marcelo H. Garcia

Bibliography. D. A. Edwards, *Turbidity Currents: Dynamics, Deposits and Reversals*, 1993; B. McCaffrey, B. Kneller, and J. Peakall, *Particulate Gravity Currents*, 2001; D. A. V. Stow (ed.), *Deep Water Turbidite Systems*, 1991.

Turbine

A machine for generating rotary mechanical power from the energy in a stream of fluid. The energy, originally in the form of head or pressure energy, is converted to velocity energy by passing through a system of stationary and moving blades in the turbine. Changes in the magnitude and direction of the fluid velocity are made to cause tangential forces on the rotating blades, producing mechanical power via the turning rotor.

The fluids most commonly used in turbines are steam, hot air or combustion products, and water. Steam raised in fossil fuel-fired boilers or nuclear reactor systems is widely used in turbines for electrical power generation, ship propulsion, and mechanical drives. The combustion gas turbine has these applications in addition to important uses in aircraft propulsion. Water turbines are used for electrical power generation. Collectively, turbines drive over 95% of the electrical generating capacity in the world. *See* ELECTRIC POWER GENERATION; GAS TURBINE; HYDRAULIC TURBINE; STEAM TURBINE; TURBINE PROPULSION; TURBOJET.

Turbines effect the conversion of fluid to mechanical energy through the principles of impulse, reaction, or a mixture of the two. For the impulse turbine, high-pressure fluid at low velocity in the boiler is expanded through the stationary nozzle to low pressure and high velocity. The blades of the turning rotor reduce the velocity of the fluid jet at constant pressure, converting kinetic energy (velocity) to mechanical energy. *See* IMPULSE TURBINE.

For the reaction turbine, the nozzles are attached to the moving rotor. The acceleration of the fluid with respect to the nozzle causes a reaction force of opposite direction to be applied to the rotor. The combination of force and velocity in the rotor produces mechanical power. *See* REACTION TURBINE.

Frederick G. Baily

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; H. P. Bloch, *A Practical Guide to Steam Turbine Technology*, 1995; M. P. Boyce, *Gas Turbine Engineering Handbook*, 3d ed., 2006; R. H. Kehlhofer, *Combined Cycle Gas and Steam Turbine Power Plants*, 2d ed., 1999; J. L. Kerrebrock, *Aircraft Engines and Gas Turbines*, 2d ed., 1992; E. Logan (ed.), *Handbook of Turbomachinery*, 2d ed., 2003.

Turbine engine subsystems

A typical aircraft gas turbine engine is composed of an assemblage of individual turbine engine subsystems and components which function together to provide thrust to propel the aircraft and power for its onboard functioning (**Fig. 1**).

Inlets. Virtually all operational aircraft gas turbine engines work with subsonic axial inlet Mach numbers at the fan or compressor inlet face (although there are experimental fans and compressors that involve supersonic flow into the compressor). It is the function of the inlet to intercept the stream tube that contains the airstream that will be ingested by the engine at any flight speed, accommodating any power level of the engine (that is, airflow requirement) and any flight attitude (such as a high yaw or pitch flight angle of attack), and to accelerate or diffuse that airstream with a minimum of pressure drop and distortion to the Mach number accepted by the fan or compressor or fan face (**Fig. 2**).

The inlet may involve a complex transition geometry where the intake is not a simple circle concentric with the engine; such a geometry is often used in turboprop engines with offset gearboxes. An additional complication is encountered in applications of engines embedded in aircraft fuselages with "cheek" intakes on either side of the aircraft. In this case, the two inlets must be joined to supply a single stable flow of air to the engine, even in the case of an aircraft in a highly yawed flight condition where one inlet may be partially blanked by separated flow from the aircraft's forebody. *See* TURBOPROP.

For low-speed aircraft such as helicopters, the inlet must accelerate the intake flow, so that the inlet is essentially a bellmouth. This generally allows the designer to provide a generous radius to the intake lip (Fig. 2*a*), which makes the inlet very tolerant of high pitch or yaw angles of attack. For aircraft with a very large range of flight speeds, ranging from takeoff to transonic cruise, the inlet must diffuse the high-speed intake flow down to the lower Mach number required by the fan or compressor. This generally requires a relatively smaller intake lip radius (Fig. 2*b*), which makes the system much more sensitive to off-design variation in airflow and to aircraft attitude.

The ultimate complication is encountered in supersonic flight, which requires a convergent/divergent diffuser to decelerate the supersonic intake flow to subsonic velocity. This may require a variable-area throat to swallow the high-loss strong shock that forms in front of the intake as the aircraft reaches supersonic flight speed. *See* SUPERSONIC DIFFUSER.

Inlet particle separators. Helicopters often land on, take off from, and hover over unprepared spaces and can churn up a dust storm that is sucked into the engine inlet. If allowed to enter the engine, this sand and dust will severely erode and damage the turbomachinery blading and will clog small air holes that supply vital cooling air to hot parts. Modern helicopter engines include inlet particle separators that pass the air through curvilinear passages in which the sand and dust particles are centrifuged and concentrated in a collection zone where they are extracted with a portion of the airflow by an ejector or a blower (**Fig. 3**). Turboprop engines may include primitive separator systems to capture ingested birds, hailstones, or other foreign objects that might damage the engine. *See* HELICOPTER.

Fans and compressors. The vast majority of fans and compressors (whose function is to increase the

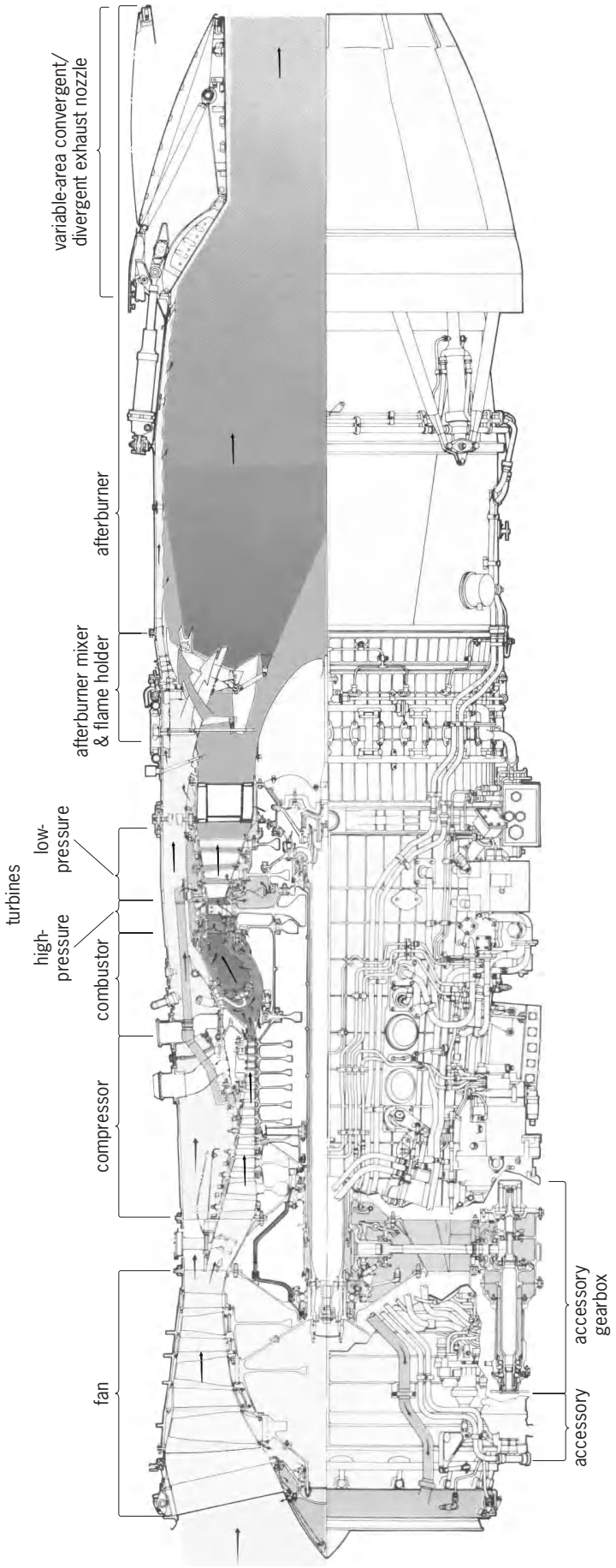


Fig. 1. Turbine engine subsystems and components of a typical aircraft gas turbine engine (afterburning low-bypass turbofan). (*Courtesy of GE Aircraft Engines*)

pressure of the incoming flow of air) of aircraft gas turbine engines are axial-flow turbo-compressors (Fig. 1). These devices consist of a series of individual stages, each comprising a rotating row (or cascade) of radially oriented airfoils (or rotor blades) followed by a stationary row of radially oriented airfoils (or stator blades). The rotating row of blades acts to increase the pressure and tangential velocity of the flow. The flow exiting the rotating row of blades is directed onto the stationary blade row which converts some of the tangential component of velocity imparted by the preceding rotor blade into a further increase in stage pressure rise.

Although the compressors of early jet engines were radial or centrifugal flow components, axial-flow compressors have been developed to achieve higher efficiency levels and lower frontal areas. They are therefore the prevailing approach in all jet engines except for the smallest sizes, where the axial-flow airfoils are not rugged enough for engine duty. In these small machines, single- and two-stage centrifugal compressors are found, as well as hybrid axi-centrifugal components consisting of one or more axial stages followed by a centrifugal stage.

Multistage compressors with pressure ratios above 5:1 or 10:1 (Fig. 1) involve extremely wide variation of flow conditions to the individual rows of airfoils in the front and the rear stages of the compressor as the speed and pressure ratios vary through the operating range. Such mismatch is aggravated by other operational circumstances such as nonuniform pressure distributions delivered by the inlet duct or rapid changes in the spool speed. Ultimately, the large angles of attack seen by individual airfoils will cause them to stall and result in unstable operation of the engine called surge. Such surge may cause flameout or large pressure pulsations that result in physical damage to the engine. Various approaches have evolved to afford high-pressure compressors the required amounts of stall margin. A very high pressure compressor may be split into two sections—the intermediate section and the high-pressure section—each operating on a separate spool at its own speed and driven by its own turbine. An alternative or supplementary approach to design of high-pressure compressors is to provide pivots for each stator vane in forward stages of the compressor along with actuators that will vary the pitch angle of the vanes as a function of the spool speed and compressor inlet conditions to avoid the large angles of attack and the stall and surge that might ensue. *See* COMPRESSOR.

The fans of low-bypass turbofans (sometimes referred to as low-pressure compressors) are generally multistage axial-flow components, which do not differ materially from high-pressure compressors in concept or design (Fig. 1). On the other hand, medium- and high-bypass turbofans usually have a single-stage axial-flow compression stage, and may differ considerably in important respects from multistage compressors. In order to minimize the frontal area of the engines, fans are designed with a very low hub radius of the blades (compared to the blades' tip radius). They must also be designed to be much more rugged than other compressor blading in order to absorb the impact of birds, hailstones, and other foreign objects that may be drawn into the engine inlet. Fan blades in ultrahigh-bypass engines may incorporate provisions for varying their pitch in order to achieve thrust-reversing capability. *See* TURBOFAN.

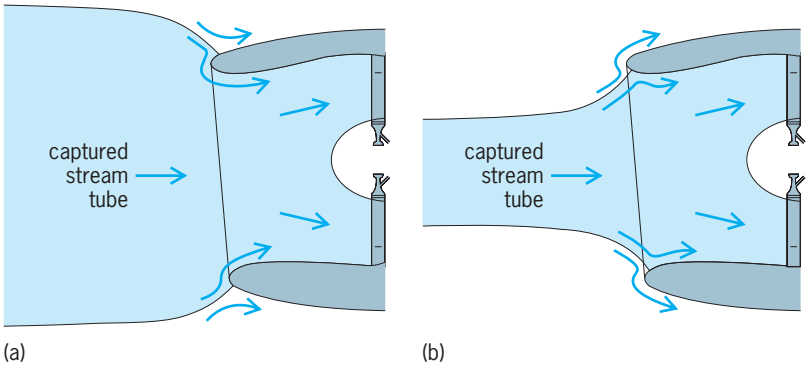


Fig. 2. Operation of a subsonic inlet at low and high flight speeds. (*a*) Flight speed lower than fan entry air velocity. (*b*) Flight speed higher than fan entry air velocity.

Combustors. The combustion chamber or combustor of the aircraft gas turbine engine (Fig. 1) provides for the burning of fuel in the airflow exiting from the compressor, and for supplying the resulting stream of high-temperature, high-pressure products of combustion to the turbine. In modern engines the combustor is usually an annular chamber, in the middle of the core between the compressor exit and the turbine entry, surrounding the center shaft of the rotor. Fuel is introduced at the upstream end through injectors, which may exploit any of several mechanisms to condition the fuel for the combustion process.

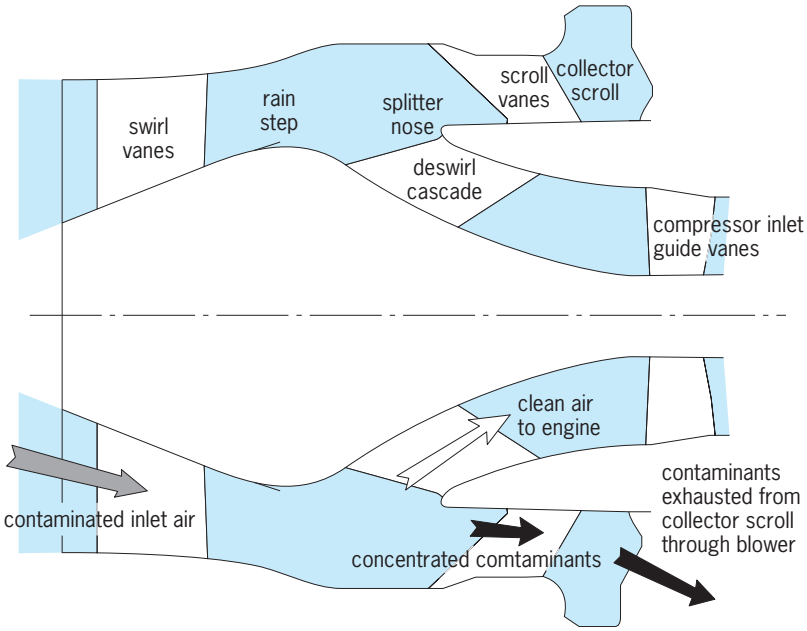


Fig. 3. Inlet particle separator. (*After M. G. Ray and J. L. Browne, T700 Engine Integral Inlet Separator, 6th European Rotorcraft and Powered Lift Forum, No. 37, University of Bristol, United Kingdom, 1980*)

With a high-pressure fuel or atomizing system, the fuel is introduced through very small orifices in a nozzle, generating a very fine spray. With an intermediate fuel pressure system, sometimes referred to as the air-blast injection system, the droplets may not be fine enough to permit efficient combustion, so various types of shear layers are introduced into the airstream in proximity to the injector to further break up the droplets. With a low-pressure fuel system and a vaporizing system, the injectors may be located in a hot region of the combustor so that the fuel is preheated and completely vaporized before entering the combustor. The combustion process is confined in a liner or shell, cooled by unburned air, which serves not only to contain the flame, but also to provide passages for air jets, lateral to the main direction of the flow, which promote and stabilize the flame. These lateral jets of unburned air also dilute its high-temperature zones to a lower and more uniform level of temperature that can be tolerated by the downstream turbine. One or more electric spark igniters are placed strategically in the chamber to light the flame in the course of starting the engine. Provision is made in the design for minimizing production of pollutants—carbon monoxide, oxides of nitrogen, hydrocarbon residues, particulates and smoke. *See* ATOMIZATION.

Turbines. Turbines in aircraft gas turbine engines extract mechanical energy from the stream of high-pressure, high-temperature airflow exiting the combustor. There may be several turbines in series in a typical engine, each driving an individual spool of the compressor ahead of it, or a fan component, or an external load like a propeller or helicopter rotor.

As shown typically in Fig. 1, each stage of turbines in modern aircraft gas turbine engines generally includes a stationary row (or cascade) of radially oriented blades or nozzle passages which accelerate and turn the incoming flow and direct it onto the following rotating row of radially oriented airfoils (or rotor blades). *See* TURBINE.

The modern gas turbine engine places enormous demands on virtually all the mechanical, aerodynamic, heat-transfer, metallurgical, and manufacturing process arts. The blades of the turbine rotate at velocities near the speed of sound, in a centrifugal acceleration field thousands of times that of gravity, in a chemically active transonic gas stream which is filled with abrasive particles and whose temperature may be well above the melting temperature of the material of which the blades are made. The blades and other parts of the turbine must also endure thousands of cycles of large changes in temperature and stress as the engines are started and shut down and as power is varied during the course of their normal operation.

There are two approaches to coping with the extremely high temperatures: by means of heat-resistant materials, and by cooling the blade with lower-temperature air. Jet engine needs have provided the incentive for a continuing progression of improved high-temperature materials, starting with steel alloys, followed by generations of cobalt and nickel alloys, and enhanced by treatments such as dispersion strengthening and by directional solidification and monocrystal casting processes. More recent developments include intermetallic alloys and even nonmetallic materials such as ceramics and composite compositions of high-temperature phases of carbon. *See* HIGH-TEMPERATURE MATERIALS.

The turbine blades are also cooled by air extracted from the compressor discharge flowpath, bypassing the combustor, and brought aboard the turbine rotor. The air is introduced into small circuitous passages cast into the turbine blades which are designed to promote conductive, convective, impingement, or transpiration cooling of the blades. Some of the spent cooling air is then conducted through small passages through the surface of the blade, and it is directed to provide film cooling over critical surface areas of the blade (**Fig. 4**). In spite of this challenging environment and duty, the refinement of all these technologies has made possible durable turbines that operate safely and reliably, and last thousands of hours

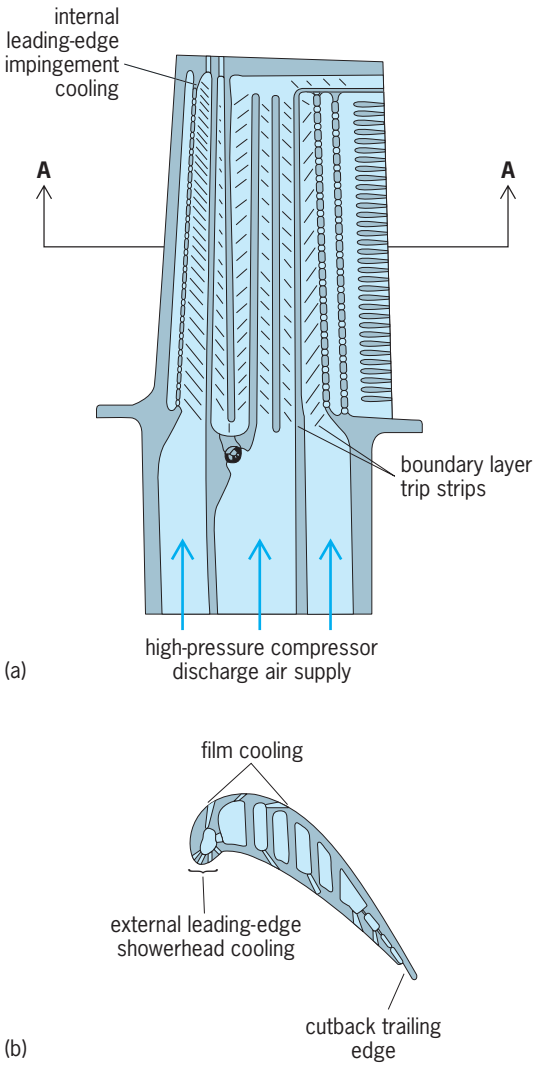


Fig. 4. Typical cooling scheme for a high-temperature turbine blade. (*a*) Plan view. (*b*) Cross-section view A–A. (*Courtesy of Pratt and Whitney*)

without overhaul or replacement, in many cases, several years "on-the-wing."

Regenerators and recuperators. A regenerator is a heat exchanger used in a gas turbine engine to transfer heat from the waste heat in the exhaust stream to the air being discharged from the compressor, which enters the combustor. The net effect is to reduce the amount of fuel required in the combustor with a net improvement in the fuel economy of the engine. Although regenerators are common in stationary and surface vehicle gas turbines, they have not yet proven useful in aircraft propulsion because the weight and bulk of current regenerators cause more deterioration in the mission performance of the aircraft than can be compensated by the possible reduction in fuel consumption.

Jet nozzles and variable exhaust nozzles. Simple jet engines and turbofan engines, which do not have an external propulsor such as a propeller, derive their thrust from the acceleration of the airflow through the turbine engine. This dictates that the energy generated in the engine as high-pressure, high-temperature airflow must be used to accelerate the airflow through an exhaust nozzle at the rear of the engine. The shape of the nozzle is very much a function of the pressure ratio across the nozzle (that is, the ratio of the pressure delivered by the engine to the nozzle, to the ambient pressure to which the exhaust is delivered). When the pressure ratio is less than the critical value (the value required to drive the stream to sonic velocity, approximately 1.9), the nozzle may have a simple convergent or conical shape. For engine operating conditions where the nozzle pressure ratio is greater than the critical value, the nozzle must be configured with an area distribution that initially converges and then diverges to accommodate the transition to supersonic conditions in the exhaust stream. In circumstances where there is a large variation of volumetric airflow from the engine in the course of its range of power level and environmental operation, the nozzle throat area may be designed to be variable to accommodate the range of operation most efficiently. *See* NOZZLE.

Afterburners. As illustrated in Fig. 1, an afterburner may be added to low-bypass turbofans and turbojets as a means of achieving thrust augmentation by burning fuel in the exhaust of the engine before it exits the engine through the exhaust nozzle. Although afterburners are relatively lightweight and are capable of augmenting thrust from 30% to 50%, they are extremely inefficient in fuel usage and are applied only in circumstances requiring short-duration bursts of thrust augmentation for extreme conditions such as takeoff, transonic acceleration in supersonic aircraft, and combat maneuvers in military aircraft.

The main components of an afterburner are a mixer (in turbofan engines) to mix the low-temperature bypass stream with the high-temperature core engine exhaust; the fuel injectors, which spray the liquid hydrocarbon fuel in the upstream end of the burner; the flame holders, bluff bodies whose wakes stabilize the location of the flame front; a shell liner, which provides containment of a film of cooler air between itself and the afterburner casing, and may also include a perforated panel whose purpose is to suppress combustion instability known as screech; and a variable exhaust nozzle, which is capable of adjusting the nozzle throat area and the exhaust area to accommodate the considerable changes in volumetric flow associated with the variation of the

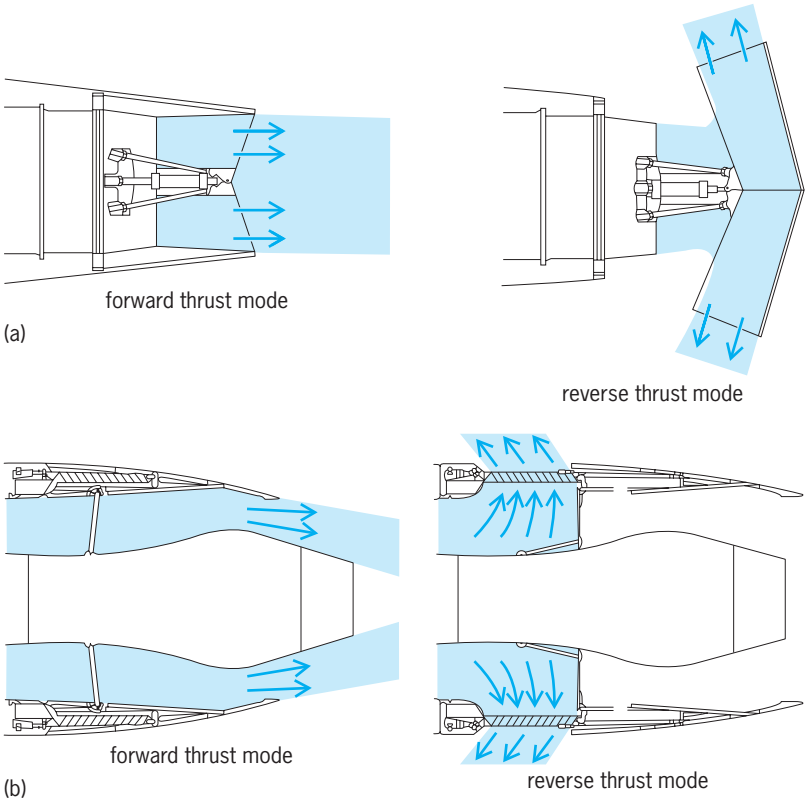


Fig. 5. Typical aircraft gas turbine thrust reversers. (*a*) Target-type reverser, typically applied to jet engines. (*b*) Blocker-door-type reverser, typically applied to turbofan engines. (*After M. J. T. Smith, Aircraft Noise, 2d ed., Cambridge University Press, Cambridge, United Kingdom, 2004*)

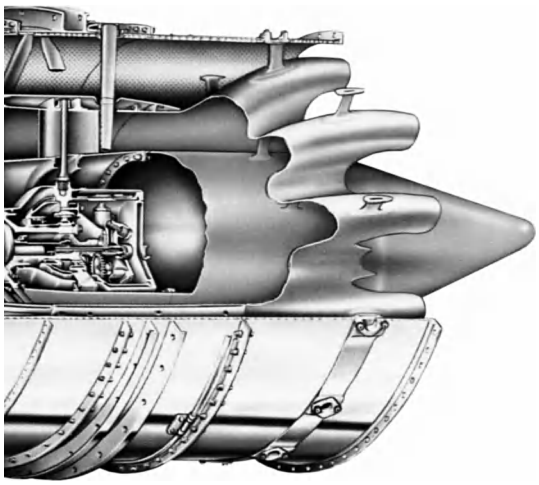


Fig. 6. Exhaust mixer in low-bypass turbofan for noise suppression and improved propulsive efficiency. (*Courtesy of Pratt & Whitney*)

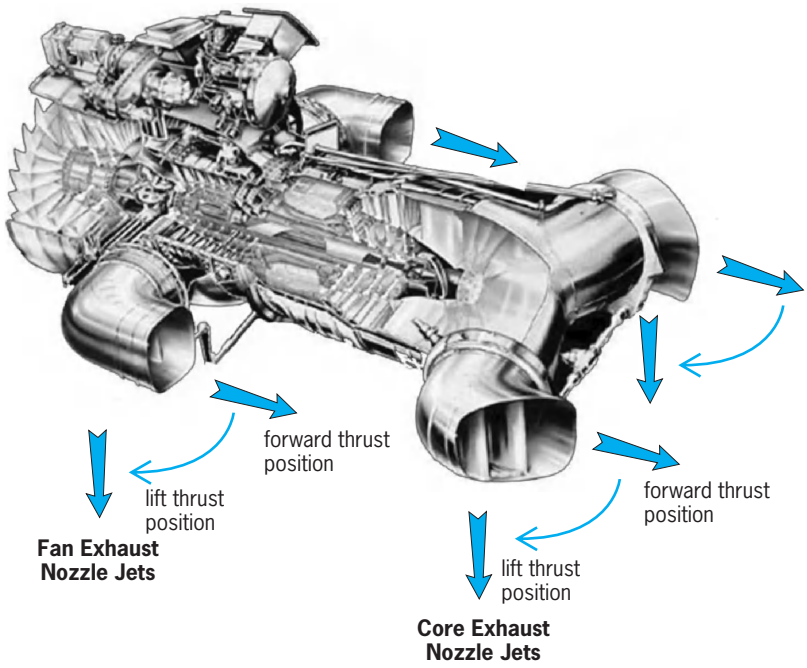


Fig. 7. Turbofan engine with swivel nozzles in both fan and core exhaust for vertical takeoff and landing. (*Courtesy of Rolls Royce*)

gas temperature rise in the afterburner (Fig. 1). *See* AFTERBURNER.

Thrust reversers. Aircraft gas turbine engines may include provision for reversing the direction of flow of the airstream handled by the propulsor in order to provide braking thrust for the aircraft, primarily during its landing roll on very short runways or runways with icy surfaces. For turboprops and ultrahigh-bypass engines, this is usually accomplished by varying the pitch angle of the propeller or of the fan blading, so that the propulsor pumps the air in the direction opposite to that required for forward flight. Simple jet engines may include a target reverser, in which a barrier is deployed in the jet stream exiting from the rear of the engine to redirect the jet in a direction with a forward component of velocity (**Fig. 5***a*). For high- and medium-bypass turbofans, this is more easily accomplished by introducing blocker doors into the fan discharge bypass stream, and simultaneously opening up alternative passages that reverse the air to a direction opposite to the flight direction (Fig. 5*b*). A retarding thrust of the order of 40% of the normal engine thrust is typically obtained by these devices.

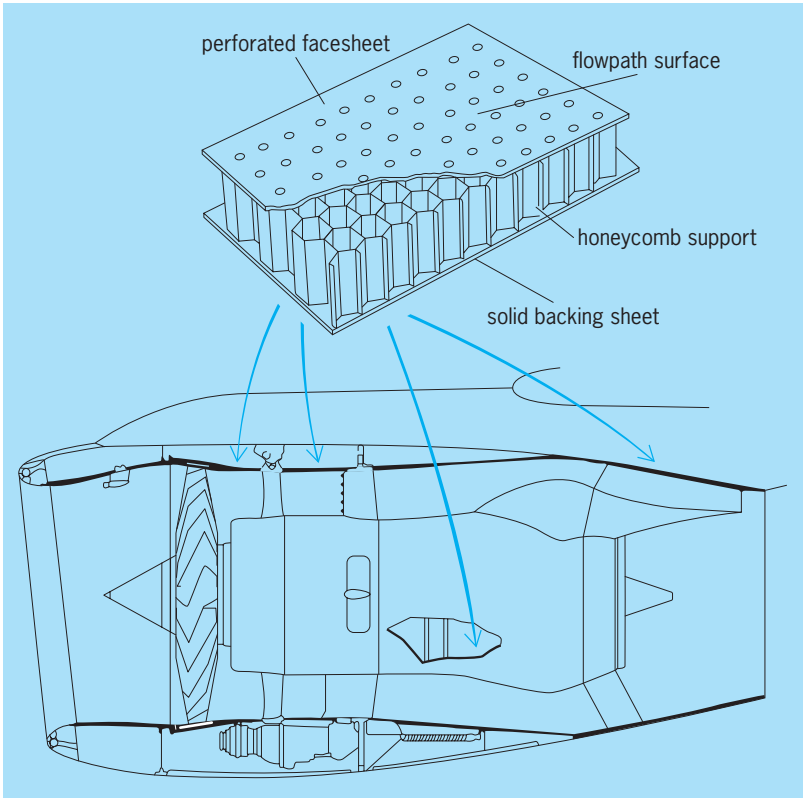


Fig. 8. Duct lined with hexagonal-cell honeycomb Helmholtz resonator cavities for turbomachinery noise suppression. (*After M. J. T. Smith, Aircraft Noise, 2d ed., Cambridge University Press, United Kingdom, 2004*)

Mixers and noise suppressors. Jet noise generally involves a broad frequency band and is dealt with by providing a highly convoluted exhaust nozzle, making a transition from the annular duct of the turbomachinery exit to a "daisy" or a multijet configuration with a extended perimeter (**Fig. 6**). The action of the nozzle's extended perimeter is to intensify the mixing and hence to shorten the length of the noise-generating shear layer between the jet and the lower-speed airflow external to the nozzle. Such suppressors were very common in earlier generations of subsonic transports that used turbojet engines. The action and the configurations are quite similar to the mixers that are used to mix the core jet with the bypass jet in mixed-flow turbofan engines to provide for noise suppression as well as the propulsive efficiency increase associated with eliminating any substantial difference between the velocities of the core and the bypass streams. They are no longer required on the very high bypass turbofans, with their much lower jet velocities, used in recent generations of subsonic transports. Supersonic transport engines use low-bypass turbofans or turbojet engines with much higher jet velocities that require such suppression devices. Mixers are also used in low-bypass afterburning turbofan engines to mix most of the bypass stream with the core exhaust to assure that the bypass stream will be burned in the afterburner (Fig. 1). *See* AERODYNAMIC SOUND; TURBOJET.

Thrust deflectors. To achieve increased agility and maneuverability, the low-bypass turbofans or turbojet engines of military combat aircraft may include provision for mechanically deflecting the jet nozzle from the direction of flight to a small angle upward or downward to provide a component of thrust in the direction of climb or descent. The nozzle may also be provided with an additional degree of freedom to deflect the exhaust to either side to provide the aircraft with increased capability to execute sharp turns. In a more extreme application, the thrust deflectors may be designed to deflect the exhaust stream to a full 90° or more to facilitate vertical takeoff and landing (**Fig. 7**). *See* VERTICAL TAKEOFF AND LANDING (VTOL).

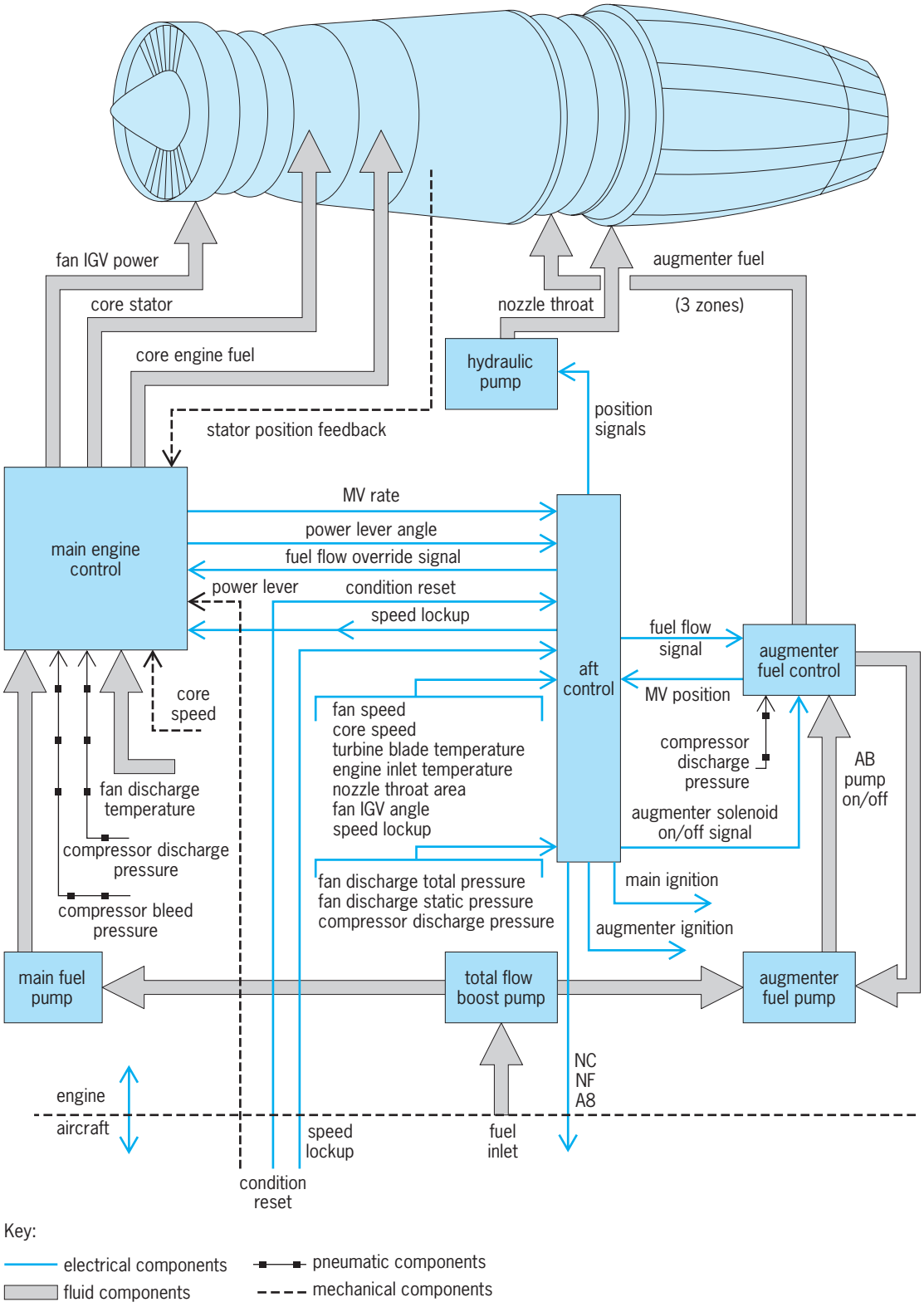


Fig. 9. Schematic of typical aircraft gas turbine engine control system. IGV = inlet guide vane; MV = motorized valve; AB = afterburner; NC = core spool speed signal; NF = fan spool speed signal; A8 = exhaust nozzle area signal. (*After T. W. Fowler, Jet Engine Propulsion Systems for Engineers, GE Aircraft Engines, Cincinnati, OH, 1989*)

Turbomachinery noise suppressors. In high-bypass-ratio turbofan engines where the exhaust jet noise is no longer dominant, turbomachinery noise becomes intrusive and must be suppressed. The generated noise is generally dominated by pure tones (blade passing frequency and its whole-number multiples). It may be suppressed by panels which include a large number of tuned Helmholtz resonating cavities that

are embedded in the walls of inlet and exhaust ducts (**Fig. 8**). *See* HELMHOLTZ RESONATOR.

Infrared suppressors and radar cross-section reduction. A key requirement of modern military aircraft engines is that they minimize their emission of infrared energy from their hot parts, and their reflection of radar microwave energy by which the aircraft might be detected and tracked by hostile missiles. Infrared suppressors are basically convoluted exhaust ducts that shield the line-of-sight of the hot turbine from any view, and provide for cooling the duct walls enough so that they themselves do not radiate detectable levels of energy.

Radar cross-section reduction also may involve convoluted ducts that shield the line-of-sight of the face of the engines' turbomachinery, which might otherwise act as a large reflector of radar microwave energy. The ducts are coated with substances that absorb microwaves so that they themselves do not contribute to the radar reflectivity of the aircraft. *See* MILITARY AIRCRAFT; RADAR-ABSORBING MATERIALS.

Accessory drives and accessories. In order to provide power to rotate an engine's accessories, a bevel gear set is generally provided at the front end of the core engine which drives a radial shaft through one of the front frame struts to an external gearbox (Fig. 1). The external gearbox includes pads for mounting and driving the accessories. Typical driven accessories might include an electric generator-starter, tachometer, fuel pump, hydraulic pump, and lubrication supply and scavenge pumps.

Control systems. A key auxiliary function that must be performed in any engine is control. The control system (**Fig. 9**) is centered on the main fuel control, which accepts a variety of signals from sensors situated around the engine that provide data indicative of the status of the engine's operation (such as measurements of key pressures and temperatures within the engine, rotative speeds of the engine's spools, actual positions of the engine's variable geometry, torque in the engine's output shaft, and rates of fuel flow to the engine's combustion systems); data indicative of the aircraft's operating conditions; and a signal from the pilot's throttle or power demand lever. On the basis of these data and of a complex program, the control energizes actuators that manipulate fuel valves and other variable geometry of the engine (such as variable-pitch stators, variable-area and vectoring exhaust nozzles, and thrust reversers) to provide stable power at the demanded level and, when requested, provide smooth and fast changes in power level. The control must also provide for starting and shutdown of the engine; must protect the engine from surge, overspeed, overtemperature, and overtorque; and in the event of any failure, must provide for residual power or for safe shutdown. Early generations of engines used hydromechanical controls, but modern systems are generally based on digital electronic devices. A starting system (a hydraulic, pneumatic, or electrical motor) is mounted on the core spool's accessory gearbox to rotate the spool up to a speed where it is pumping enough air to permit combustion and is able to generate enough energy to be self-sustaining. The fuel system includes the fuel pumps, regulating valves, flow dividers, and fuel nozzles. The ignition system includes a high voltage unit and igniter. The anti-icing system provides hot air or electrical heating to parts in the engine inlet which might otherwise become clogged with ice during flight through supercooled moisture in the atmosphere. Electrical, hydraulic, or pneumatic actuators drive the engine's variable geometry such as variable-pitch compressor stator blades and variable-area exhaust nozzles. *See* CONTROL SYSTEMS.

Fredric F. Ehrich

Bibliography. J. H. Horlock, *Axial Flow Compressors: Fluid Mechanics and Thermodynamics, Butterworths*, London, 1958, Krieger Publishing, Malabar, FL, 1973; J. H. Horlock, *Axial Flow Turbines: Fluid Mechanics and Thermodynamics, Butterworths*, London, 1966, Krieger Publishing, Malabar, FL, 1973; J. L. Kerrebrock, *Aircraft Engines and Gas Turbines*, 2d ed., 1992; M. Kroes et al., *Aircraft Powerplants*, 7th ed., 1995; J. D. Mattingly, W. H. Heiser, and D. T. Pratt, *Aircraft Engine Design*, 2d ed., AIAA, Reston, VA, 2002; G. C. Oates, *Aircraft Propulsion Systems: Technology and Design*, 1989; Rolls Royce plc, *The Jet Engine*, 5th ed., Derby, United Kingdom, 2005.

Turbine propulsion

Propulsion of a vehicle by means of a gas turbine. While gas turbines have found significant applications in stationary and mobile power plants, their light weight, compact volume, low frontal area, and long life (in comparison to reciprocating diesel or Otto-cycle internal combustion engines) make them ideal for primary propulsion of vehicles.

Since about 1940, gas turbines have come to dominate most areas of common carrier aircraft propulsion, have made significant inroads into the propulsion of surface ships, and have been incorporated into military tanks. Turbine propulsion has been a subject of development activity for application to railroad locomotives and to buses, trucks, and automobiles.

Core or gas generator. The primary power producer common to all gas turbines used for propulsion is the core or gas generator, operating on a continuous flow of air as working fluid. The air is compressed in a rotating compressor, heated at constant pressure in a combustion chamber burning a liquid hydrocarbon fuel, and expanded through a core turbine which drives the compressor. This manifestation of the Brayton thermodynamic cycle generates a continuous flow of high-pressure, high-temperature gas which is the primary source of power for a large variety of propulsion schemes. The turbine is generally run as an open cycle; that is, the airflow is ultimately exhausted to the atmosphere rather than being recycled to the inlet. *See* BRAYTON CYCLE.

For gas generators using a very high pressure, the compressor and turbine may each be divided into two subunits to make a two-spool gas generator.

A heat exchanger may be incorporated in a midstation in the compressor as an intercooler to improve the cycle efficiency. A heat exchanger may be installed as a regenerator or recuperator to transfer heat from the airflow ultimately exhausted from the engine to the airflow entering the combustor, thereby decreasing the amount of hydrocarbon fuel needed to produce a specific temperature of the gas. This also improves the cycle efficiency. *See* HEAT EXCHANGER.

Water, or a water-methanol mixture, may be injected in the airstream upstream of the combustor to augment the available power from the gas generator.

Propulsion. The residual energy available in the high-temperature, high-pressure airstream exiting from the core is used for propulsion in a variety of ways.

For traction-propelled vehicles (buses, trucks, automobiles, military tanks, and most railroad locomotives), the core feeds a power turbine which extracts the available energy from the core exhaust and provides torque to a high-speed drive shaft as motive power for the vehicle. With a free-turbine arrangement, this power turbine is a separate shaft, driving at a speed not mechanically linked to the core speed. With a fixed turbine, this power turbine is on the same shaft as the core turbine, and must drive at the same speed as the core spool. In traction vehicles the power turbine generally drives through a transmission system (that is, a gear system or hydraulic pump-motor set or electrical generator-motor set) which affords a constant- or a variable-speed reduction to provide the necessary torque-speed characteristics to the traction wheels. *See* AUTOMOTIVE TRANSMISSION.

Aircraft, ships, and high-speed land vehicles, which cannot be driven by traction, are propelled by reaction devices. Some of the ambient fluid around the vehicle (that is, the water for most ships, and the air for all other vehicles) is accelerated by some turbomachinery (a ship propeller, aircraft propeller, helicopter rotor, or a fan integrated with the core to constitute a turbofan engine). The reaction forces on this propulsion turbomachinery, induced in the process of accelerating the ambient flow, provide the propulsion thrust to the vehicle. In all these cases, motive power to the propeller or fan is provided by a power turbine extracting power from the gas generator exhaust. In the case of a jet engine, exhaust from the gas generator is accelerated through a jet nozzle, so that the reaction thrust is evolved in the gas generator rather than in an auxiliary propeller or fan. Indeed, in turboprop and turbofan engines, both forms of reaction thrust (from the stream accelerated by the propeller or fan and from the stream accelerated by the core and not fully extracted by the power turbine) are used for propulsion (**Fig. 1**). *See* JET PROPULSION; TURBOFAN; TURBOJET; TURBOPROP.

Propulsive efficiency. The efficiency of the energy generation in a turbine power plant is a function of the pressure ratio of the compression process (including the ram pressure induced in a high-speed vehicle by decelerating the ambient air to bring it aboard the vehicle), the maximum temperature of the cycle, the efficiency of compression and expansion of the turbomachinery components, and the losses associated with pressure drops in static components, leakages, and parasitic air losses.

In reaction propulsion the efficiency of the propulsion process is also a strong function of the velocity at which the propulsive exhaust jet exits from the engine. The propulsive efficiency is given approximately by Eq. (1), where a is the ratio of exhaust

$$\eta_p = \frac{2}{1 + a} \tag{1}$$

velocity (relative to the vehicle) to the vehicle's air speed. Maximum propulsive efficiency is achieved when the exhaust velocity is equal to the air speed or $a = 1$. This ideal condition can be approached, but only at the expense of making the propulsor very large and heavy, since the thrust F_n per unit airflow W_a is found to be approximately as given by Eq. (2),

$$\frac{F_n}{W_a} = V_0(a - 1) \tag{2}$$

implying that values of a near unity give vanishingly small thrust per unit airflow. Equation (2) indicates that the thrust per unit airflow is proportional to flight speed V_0, so that a relatively large-mass flow must be handled at low speed. *See* THRUST.

In the design of a turbine engine for reaction propulsion, this balance of considerations is handled by a spectrum of propulsion systems illustrated in Fig. 1. For very low flight speeds, very large propulsors are used to accelerate large amounts of air through small velocity increments, only slightly above that low flight speed, as typified by the helicopter. High-speed propulsion systems are designed to accelerate smaller amounts of air to much larger velocities, as typified by the pure jet engine. Turboprops and turbofans constitute intermediate devices in this propulsion spectrum.

Aircraft propulsion. Except for reciprocating engines for small general aviation airplanes and helicopters with less than 700 hp (520 kW), turbine engines dominate the entire spectrum of aircraft propulsion. *See* AIRCRAFT ENGINE; AIRCRAFT PROPULSION; RECIPROCATING AIRCRAFT ENGINE.

Helicopters are generally powered by one, two, or three turboshaft engines, with the output power shafts feeding a combining gearbox which drives the helicopter rotor. The engines generally have free turbines, since the rotor must be able to operate at constant rotating speed, independent of the amount of power required. The engines are fitted with special controls which automatically hold output speed and share power evenly among the engines in multiple-engine installations. *See* HELICOPTER.

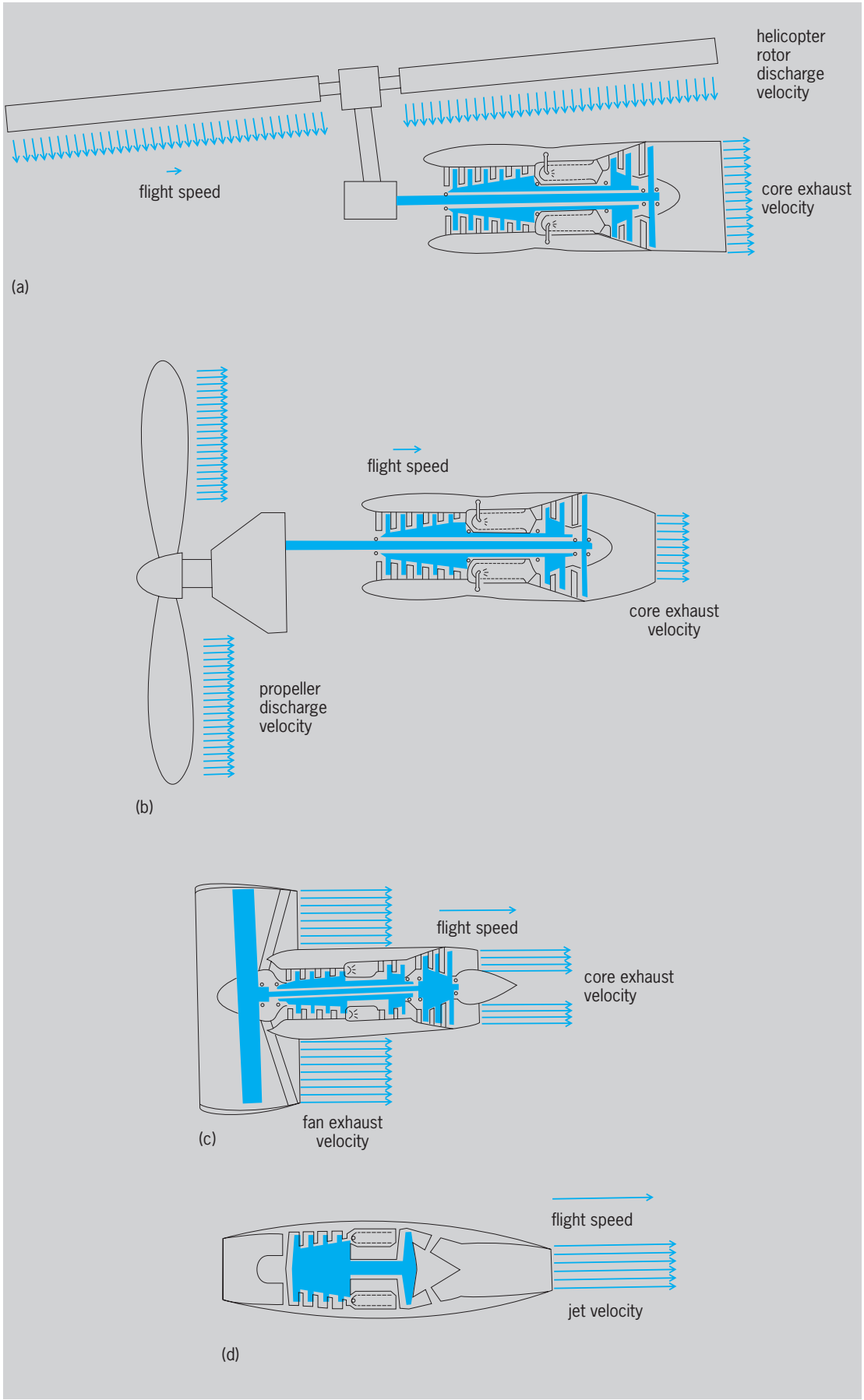


Fig. 1. Aircraft turbine propulsion systems. (*a*) Helicopter turboshaft. (*b*) Turboprop. (*c*) High-bypass turbofan. (*d*) Turbojet.

Vertical/short takeoff and landing (V/STOL) aircraft constitute a relatively new and evolving class of aircraft with an enormous variety of operational, experimental, and proposed propulsion systems. They have two categories of propulsion: conventional thrusters to propel the aircraft in forward flight, and vertical thrusters providing lift to support the aircraft in vertical and low-speed flight when the wings are not providing lift. The vertical thrust may be provided by one or a combination of schemes: deflection of the jet from the conventional thrusters; physical rotation of the nacelle or a tilt wing in which the conventional thruster is mounted; addition of a helicopter rotor to a conventional aircraft (in a convertible aircraft); addition of auxiliary vertically mounted direct-lift engines; addition of auxiliary vertically mounted fans, powered by separate engines or by shafting or compressed air drawn from the forward thrusters; and so forth. In the vertical flight mode the aircraft control and stability are dependent on differential modulated thrust from the several lifting thrusters, since the conventional aircraft control surfaces are ineffective without forward flight speed. *See* CONVERTIPLANE; SHORT TAKEOFF AND LANDING (STOL); VERTICAL TAKEOFF AND LANDING (VTOL).

Conventional takeoff and landing aircraft are generally equipped with turboprops for low subsonic flight speeds up to a Mach number of 0.7. For higher flight speeds through the transonic range, high-bypass turbofans are generally used. Hybrid engines, encompassing the advantages of the lower weight and size of the turbofan and the higher propulsive efficiency of the turboprop, are another alternative. One category is the prop-fan, where high-speed, low-diameter multibladed, counterrotating propellers are used, often in a pusher configuration appropriate to tail mounting. Although most proposed configurations use a gearbox to drive the propeller, the unique unducted fan type manages to accomplish propulsion without the use of a gearbox. Another configuration in the class of ultrahigh-bypass engines employs a shrouded fan in the conventional position before the power producer, but with a bypass ratio of 10 or more, which generally requires turboproplike features such as a gearbox drive and a variable-pitch fan to achieve thrust reversal. For propulsion at very high transonic and low supersonic flight speeds, very low-pass turbofans are generally used in which engine thrust augmentation is provided for short-duration acceleration or, in military aircraft, for combat maneuvers. Additional developments in advanced combat aircraft include variable-cycle features to accommodate extreme mixes of subsonic and supersonic duty, thrust vectoring to achieve extremes of aircraft maneuverability, and features to suppress the infrared and radar cross sections of the engine. Aircraft such as supersonic transports that require sustained flight at high supersonic speeds have been equipped with afterburning turbojet engines. *See* AFTERBURNER; MILITARY AIRCRAFT.

Ship propulsion. Turbines have also found usage in high-speed ships, where their light weight, compact envelope, fast starting, transient response characteristics, adaptability to a wide spectrum of liquid fuels, and long life are particularly suited.

Virtually all types of ships are powered by turboshaft engines which drive propellers through geared speed reducers. Many of the turbines are derived from aircraft engines: helicopter turboshafts, or turbofan engines with the fan removed, or jet engines with the addition of a power turbine. These engines are made suitable for the marine environment by rerating for the appropriate duty cycle, substitution of materials and coatings for protection from the seawater environment, addition of inlet systems to exclude ingestion of seawater, addition of appropriate exhaust stacks, provision of bases for floor mounting, addition of enclosures for acoustic isolation and thermal environmental control, and addition of starting and control systems unique to ship propulsion. *See* MARINE ENGINE.

In conventional hull- or buoyancy-borne ships, turbines are found in the high-speed end of the vehicle spectrum and, indeed, are fast coming to dominate the modern naval fleets of the world. *See* NAVAL SURFACE SHIP.

Hydrofoil ships generally have two separate power plants: small units to power the ship propellers at low forward speeds when the craft is hull-borne, and larger units to power the ship propellers or water jet propulsion units during high-speed, foil-borne operation. *See* HYDROFOIL CRAFT.

Water-borne surface-effect ships or amphibious air-cushion vehicles have two classes of turbine propulsion: prime movers to power fans which provide the air cushion to support the vehicle, and prime movers (effectively turboprops) which power air propellers to propel the vehicle. In certain installations the same turbine may power both functions. *See* AIR-CUSHION VEHICLE; SHIP POWERING, MANEUVERING, AND SEAKEEPING.

Locomotive propulsion. Over the years there has been considerable experimental and development work, and a modest amount of actual rolling stock experience, on gas-turbine-powered, traction-propelled locomotives in France, Japan, Germany, and the United States. As with ship applications, many of the gas turbines are turboshafts, adapted from shaft, fan, or jet engines developed for aeronautical use.

Various power conditioning schemes have been employed. The Sikorsky Turbotrain I, with a free turbine engine, drives directly through a gear reduction to the drive wheels, as does the Japanese National Railways KI-HA391. The Rohr/Frangeco turbine train developed for, and still in use by, Amtrak drives through a hydraulic torque converter to the drive wheels. General Electric built a series of powerful gas turbine-electric locomotives for the Union Pacific Railroad. In these units a fixed-turbine prime mover drove an electric generator through a gear reduction system, and the generated electric power drove motors at the drive wheels. One of these locomotives was modified to burn powdered coal rather than liquid petroleum.

Unique variants of turbine-powered locomotives were the very high-speed air-cushion track vehicles or aerotrains developed in the 1960s, most notably by Jean Bertin in France. Like the amphibious air-cushion vehicle, the aerotrain was supported by a pressurized air cushion, in this case developed between the vehicle and an inverted-T guideway or track. In addition to gas turbines to supply the pressurized air for the air cushion, aircraft-type turbofan or turboprop engines were used for reaction propulsion of the vehicles. Speeds on the order of 250 mi/h (400 km/h) were obtained on experimental vehicles. *See* LOCOMOTIVE.

Automotive vehicle propulsion. Beginning in the late 1940s and early 1950s, the United States auto industry ran a number of modified gas turbine engines, and in 1953 a standard passenger car was manufactured with a specially designed gas turbine engine. Over the next 15 years, five generations of gas turbine engines were installed in passenger automobiles. Although these gas-turbine-powered vehicles did not result in a marketable product capable of competing with the conventional spark-ignition internal combustion engine, they did demonstrate the gas turbine engine's smooth operation and ability to start in cold weather. *See* INTERNAL COMBUSTION ENGINE.

The two main characteristic disadvantages of the gas turbine engine for automotive operation, which must be overcome before it can displace the internal combustion engine, are its poor efficiency at part load and idle conditions, and its poor acceleration from idle condition. Another negative factor, the high production cost, relates to the use of expensive and scarce cobalt and nickel in the hot turbine sections. However, technological advances have been made in the use of ceramic components. In addition to lowering the production costs, ceramic components permit increased engine operating temperatures, around 2500°F (1370°C), which improve engine cycle efficiency. The conventional piston engine vehicle has become more expensive as a result of efforts to meet federal car mileage requirements and exhaust emission standards, causing the gas turbine engine to become more cost-competitive. *See* CERAMICS.

A free-turbine gas turbine engine, designed for 100 hp (75 kW), was developed (**Fig. 2**), which demonstrates that technology is available to achieve fuel economies equivalent to those of spark ignition

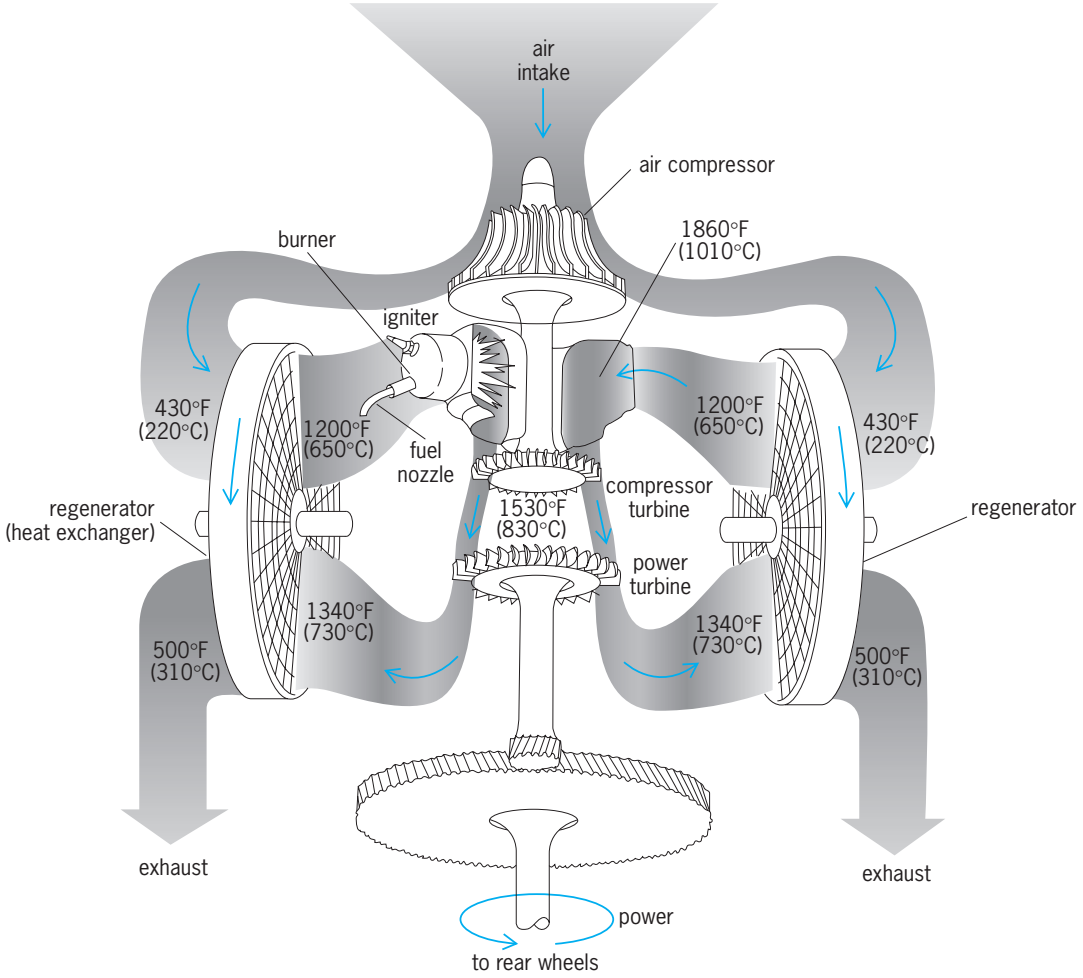


Fig. 2. Chrysler upgraded turbine engine. (*Chrysler Corp./U.S. Department of Energy*)

engines and also achieve lower exhaust emissions, easier starting, and less noise and vibration. The engine can operate on various kinds of fuel, such as coal-derived fuels, alcohols, and blends.

The leading automotive companies have also experimented with more powerful versions of automobile gas turbines for large buses and trucks. In typical long-range intercity applications, the gas turbine should retain all its intrinsic advantages and be less subject to the disability of its high idle fuel consumption. Regenerative, free-turbine engines have been developed especially to meet the requirements of heavy-duty engine applications in ground vehicles and equipment in the general range of 300–600 hp (225–450 kW). The regeneration system reduces fuel consumption and lowers exhaust temperatures.

The power transfer system is an automatic transmission without any torque converter. It offers controlled engine braking and high torque-rise performance. Gas turbine power results in less weight in the power section, the absence of a radiator, and a clean exhaust obtained with standard diesel fuels.

In 1976 the U.S. Army selected a 1500-hp (1120-kW) regenerative engine with a two-spool core and a free turbine to power its XM1 main battle tank. For cross-country duty, the gas turbine inlet is supplied with a combination inertial separator and barrier filter system to extract sand and dust from the air entering the engine. The engine runs on a wide range of fuels, including gasoline, diesel, and jet aircraft fuels. The power turbine is supplied with variable inlet guide vanes to optimize fuel consumption at all output powers. The automatic transmission provides four speeds forward and two in reverse, contains integral brakes, has continuously variable hydrostatic steering, and provides pivot steering while the transmission is in neutral. *See* ARMY ARMAMENT; GAS TURBINE; PROPULSION.

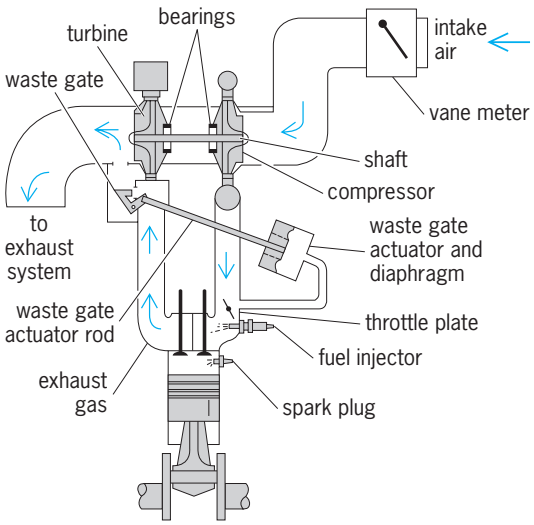
Fredric F. Ehrich

Bibliography. C. F. Foss, *Jane's Armour and Artillery*, annually; *Gas Turbine World Handbook*, annually; P. Jackson, *Jane's All the World's Aircraft*, annually; R. L. Trillo, *Jane's High-Speed Marine Craft*, annually.

Turbocharger

An air compressor or supercharger on an internal combustion piston engine that is driven by the engine exhaust gas to increase or boost the amount of fuel that can be burned in the cylinder, thereby increasing engine power and performance. On an aircraft piston engine, the turbocharger allows the engine to retain its sea-level power rating at higher altitudes despite a decrease in atmospheric pressure. *See* RECIPROCATING AIRCRAFT ENGINE; SUPERCHARGER.

Construction and operation. The turbocharger is a turbine-powered centrifugal supercharger. It consists of a radial-flow compressor and turbine mounted on a common shaft (see **illus.**). The turbine uses the energy in the exhaust gas to drive the compressor, which draws in outside air, precompresses it, and supplies it to the cylinders at a pressure above atmospheric pressure. Turbocharger speed, which may be upward of 130,000 revolutions/min, is not dependent upon engine speed, but is determined by the power balance between the turbine and the compressor.



Typical turbocharger installation on a fuel-injected spark-ignition engine, using a waste gate to limit boost pressure. Charge-air cooler is not shown. (*Ford Motor Co.*)

Common turbocharger components include the rotor assembly, bearing housing, and compressor housing. The turbine housing and control devices frequently differ according to application. The compressor housing is generally fabricated of cast aluminum, and it may contain an integral bypass valve. Turbine housings are cast from nodular graphite iron and other materials, depending on exhaust temperature. The shaft bearings usually receive oil from the engine lubricating system. Engine coolant may circulate through the housing to aid in cooling. *See* ENGINE COOLING.

Some turbochargers have twin-flow housings in which the two flows do not join until just ahead of the turbine wheel inlet. This allows pulse turbocharging in which the kinetic energy of the exhaust gas is used in addition to its pressure energy. In the usual constant-pressure turbocharging, only the pressure energy of the exhaust gas is used, along with single-flow turbine housings. Such turbocharger installations are used for automotive engines.

Boost-pressure control. The faster the engine runs, the faster the turbine spins the compressor. This causes a pressure buildup that must be controlled in order to avoid engine damage. One method is to allow a portion of the exhaust gas to bypass the turbine by opening a valve or flap known as a waste gate. It can be a separate component (see illus.) or integrated into the turbine housing. As much exhaust gas is diverted as is necessary to slow the turbine down to the proper speed.

The waste gate is actuated pneumatically by control pressure that is tapped off at the pressure end of the turbocharger. Instead of a waste gate, regulation may be provided by variable turbine geometry, as in a variable-nozzle turbocharger. This allows the constant-pressure characteristics of the turbine to change continuously, more efficiently using the exhaust-gas energy.

Charge-air cooling. As the air is compressed, its temperature rises, thus reducing the efficiency of turbocharging. The use of a heat exchanger as a charge-air cooler (also known as aftercooler or intercooler) helps overcome this problem. Before entering the cylinders, the hot air from the compressor is sent through the charge-air cooler. This cools the air, thereby increasing its density. *See* HEAT EXCHANGER; INTERNAL COMBUSTION ENGINE.

Donald L. Anglin

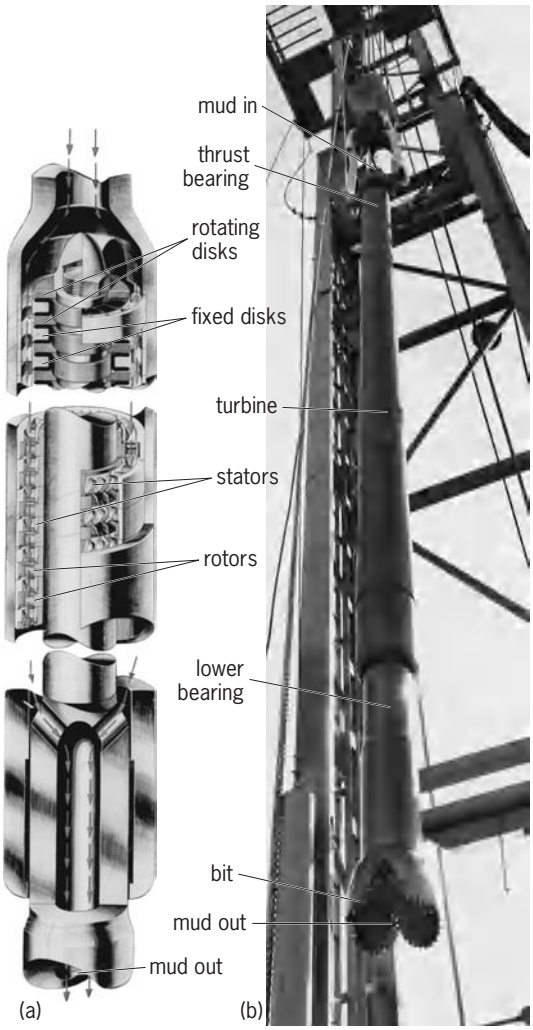
Bibliography. *Bosch Automotive Handbook*, 1987; J. Humphries, *Automotive Supercharging and Turbocharging Systems*, 1995; Society of Automotive Engineers, *Turbocharger Performance and Application*, 1982; Society of Automotive Engineers, *Turbochargers and Turbocharged Engines*, 1979; K. Zinner, *Supercharging of Internal Combustion Engines*, 1978.

Turbodrill

A rotary tool used in drilling oil or gas wells in which the bit is rotated by a turbine motor inside the well. The principal difference between rotary and turbo drilling lies in the manner that power is applied to the rotating bit or cutting tool. In the rotary method, the bit is attached to a drill pipe, which is rotated through power supplied on the surface. In the turbodrill method, power is generated at the bottom of the hole by means of a mud-operated turbine.

The turbodrill (see **illus.**) consists of four basic components: the upper, or thrust, bearing; the turbine; the lower bearing; and the bit. Most turbodrills are about 30 ft (9 m) long, with shafts about 20 ft (6 m) long. The turbodrill is attached at its top to a drill collar, or heavy length of steel pipe, that makes up the bottom end of the drill pipe extending to the surface. Once the turbodrill passes below the well head, operations on the rig floor are the same as for rotary drilling. Rotation of the drill pipe is not necessary for turbodrilling, because rotation of the bit develops through the turbine on the lower end of the drill string. It is usual practice, however, to rotate the drill pipe above the turbine slowly, at 6–8 revolutions/minute, either by means of the rotary table on the derrick floor or through torque of the turbine on the bottom. Rotation of the bit is much faster than in rotary drilling, and is usually between 500 and 1000 revolutions/minute.

In operation, mud is pumped through the drill pipe, passing through the thrust bearing and into the turbine. In the turbine, stators attached to the body of the tool divert the mud flow onto rotors attached to the shaft. This causes the shaft, which is connected to the bit, to rotate. The mud passes through a hollow part of the shaft in the lower bearing and through the bit, as in rotary drilling, to remove cuttings, cool the bit, and perform the other functions of drilling fluid. Capacity of the mud pump, which is the power source, determines rotational speed.



Components of a turbodrill. (*a*) Cutaway view of turbine, bearings, and bit. (*b*) Drill string suspended above hole. (*Dresser Industries*)

Two basic types of turbodrills are used in the United States. One is a standard 100-stage unit (one rotor and one stator comprise a stage); the other is a tandem turbodrill, made up of two or three standard sections. Although turbodrills have been in wide use in Russia, they are still relatively rare in the United States, where the emphasis has been on the rotary tool method of drilling. Despite faster penetration with the turbodrill, and several other advantages, widespread use of the turbodrill in the United States has been limited principally because of the faster wear of the bits, necessitating time-consuming and costly round trips to remove the drill string from the

hole and change the bit. *See* DRILLING, GEOTECHNICAL; OIL AND GAS WELL DRILLING. Ade L. Ponikvar

Turbofan

An air-breathing aircraft gas turbine engine (**Fig. 1**) with operational characteristics between those of the turbojet and the turboprop. Like the turboprop, the turbofan consists of a compressor-combustor-turbine unit, called a core or gas generator, and a power turbine. This power turbine drives a low- or medium-pressure-ratio compressor, called a fan, some or most of whose discharge bypasses the core. *See* TURBOJET; TURBOPROP.

Operating principle. The gas generator produces useful energy in the form of hot gas under pressure. Part of this energy is converted by the power turbine and the fan it drives into increased pressure of the fan airflow. This airflow is accelerated to ambient pressure through a fan jet nozzle and is thereby converted into kinetic energy. The residual core energy is converted into kinetic energy by being accelerated to ambient pressure through a separate core jet nozzle (**Fig. 2***a*). The reaction in the turbomachinery in producing both streams produces useful thrust.

Bypass ratio. Turbofans are generally characterized by their bypass ratio, the ratio of the airflow which flows around (bypasses) the core to the airflow which passes through the core. Airplanes which cruise at a flight Mach number less than 0.6 are generally propeller-driven (if they are turboprop-driven, they may be considered extremely high-bypass-ratio turbofan aircraft). In the range of Mach 0.6 to transonic flight speeds, bypass ratios of 3 to 7, where the fan imparts a very modest propulsive velocity to a very large bypass flow, are found to be most efficient and dominate aircraft applications (Fig. 1*a*). Their outstanding fuel economy also assures a low rate of exhaust emissions because of their relatively low fuel consumption. The low velocity in their exhaust jets also makes them inherently quieter than jet engines or low-bypass turbofans. For aircraft which fly supersonically or have mixed supersonic-subsonic missions, turbofans with higher exhaust jet velocities and lower bypass ratios of 0.1 to 3 are used. Such low-bypass turbofans are often equipped with afterburners and variable exhaust nozzles (Fig. 1*b*) to provide thrust augmentation (on the order of 50% greater than the thrust otherwise obtainable). Since afterburning operation involves very high specific fuel consumption, its usage is generally restricted to very short duration, typically for transonic acceleration or combat maneuver. *See* AFTERBURNER; SPECIFIC FUEL CONSUMPTION.

Where the bypass ratio is designed to be less than 0.1, the bypass air has a very minor role in affecting the propulsive efficiency of the engine. The usual

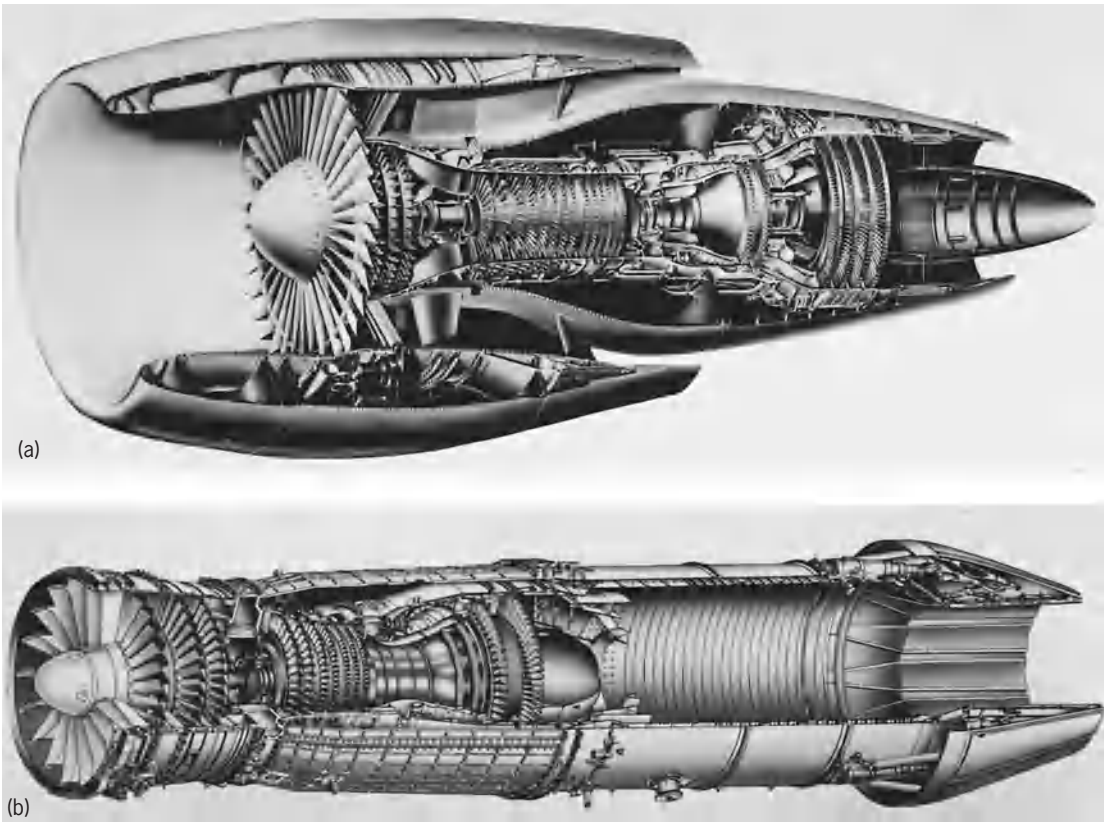


Fig. 1. Modern turbofan engines. (*a*) High-bypass engine for high subsonic flight speed. (*b*) Low-bypass afterburning engine for transonic and low-supersonic flight speeds. (*GE Aircraft Engines*)

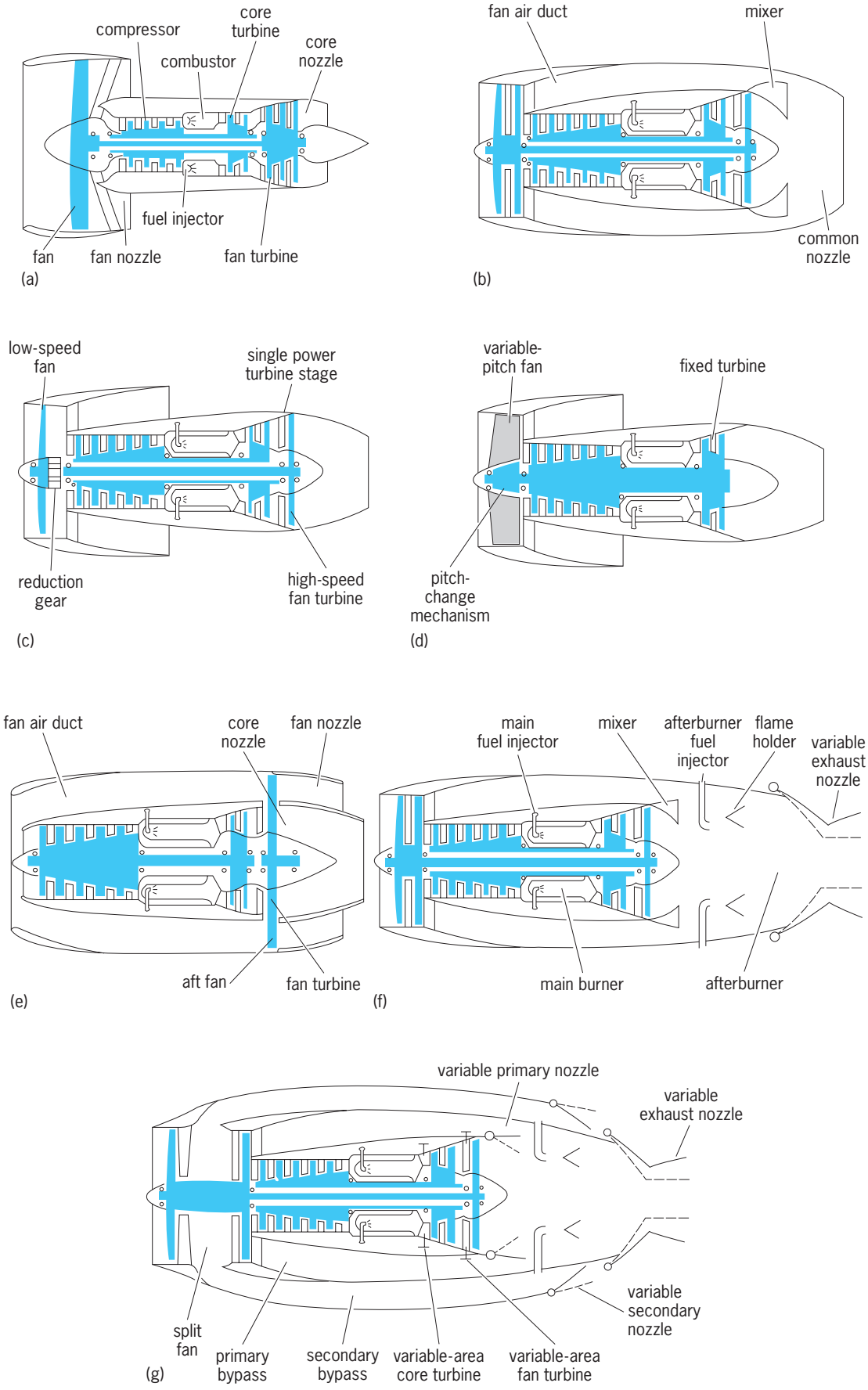


Fig. 2. Turbofan configurations. (*a*) High-bypass, separate-flow turbofan. (*b*) Mixed-flow turbofan. (*c*) Fixed-turbine, variable-pitch fan. (*d*) Ultrahigh-bypass turbofan. (*e*) Aft fan. (*f*) Afterburning turbofan. (*g*) Variable-cycle turbofan.

intent of this approach to engine design is to use the bypass air to blanket the hot end of the engine (the combustor casing, the turbine casings, and the afterburner casing) with cool air to protect the surrounding aircraft structure. In thermodynamic terms, the very low bypass turbofan engine is indistinguishable from a jet engine, where a small fraction of the compressor airflow is usually extracted to cool the structure and casings of the hot end of the engine.

Configurations. Ideally the fan and core exhaust velocities should be nearly equal. If they are not, some benefit is derived from incorporation of a mixer between the two streams and exhausting the mixed flow (Fig. 2*b*) through a common nozzle.

Variable-pitch fan blading, such as is found in propeller blading, is being applied in turbofan blading. An early application was in turbofan engines with fixed turbines (that is, where the power turbine driving the fan was part of the same rotor as the core), where the variability was used for unloading the fan during the engine starting sequence (Fig. 2*c*). *See* PROPELLER (AIRCRAFT).

A more subsequent application is in the ultrahigh-bypass turbofan, where the more conventional blocker-door thrust-reverser systems become impractical, and where variable (that is, reversible) pitch fan blading is used to provide reverse thrust (Fig. 2*d*). Another feature of such ultrahigh-bypass engines is a reduction gear, which permits the driving power turbine to rotate at a much higher speed than the large-diameter, low-speed fan. This feature allows the power turbine to be designed with fewer stages and with a lower diameter (and hence to be lighter and smaller) than would otherwise be required without a gearbox.

An early manifestation of the turbofan was the aft fan, where a row of fan blades is mounted in the rear of the engine, astride the power turbine blading (Fig. 2*e*).

Turbofans intended for service in mixed subsonic-supersonic missions are often designed with afterburners in which an additional combustor is located in the exhaust streams (Fig. 2*f*) of the turbofan to provide thrust augmentation for any thrust-limited segment of the mission. The burning may be arranged in the core stream or the bypass stream, in both, or in a mixed stream. A variable-area exhaust nozzle is generally also necessary. *See* AFTERBURNER.

Variable-cycle turbofans exploit the advantages of high-bypass operation for subsonic flight, and provide low-bypass operation for efficient supersonic cruise for mixed-mission aircraft. These involve a variety of variable-geometry components and elements in the basic turbofan: additional variable stators in the fan, variable stators in turbines, valving elements in the bypass streams, addition of a second bypass stream, more than one augmentation burner, and so forth (Fig. 2*g*).

To shorten the landing of aircraft, turbofans are often provided with reversers, devices which block the exhaust jet and partially reverse its direction, thereby reversing the thrust vector.

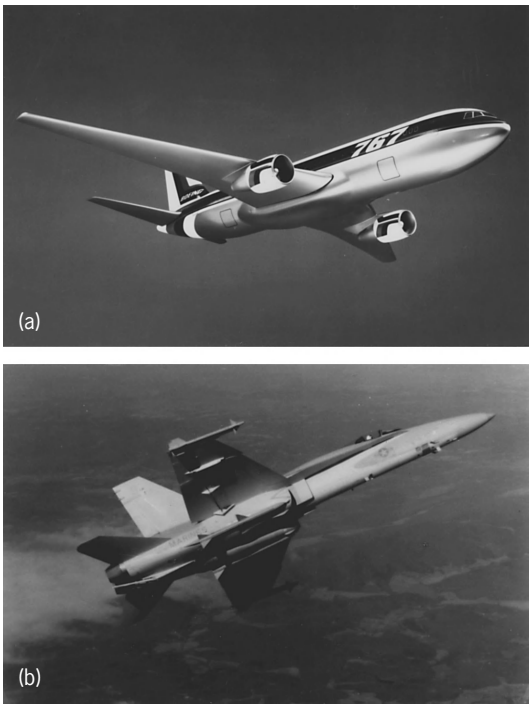


Fig. 3. Turbofan installations. (*a*) Twin-turbofan, under-the-wing installation in a modern transport aircraft (*Boeing*). (*b*) Military combat aircraft with twin low-bypass afterburning turbofan engines embedded in fuselage (*GE Aircraft Engines*).

To shorten the takeoff field length required, particularly on hot days or at high altitudes, provision can be made for water or water-methanol to be injected into the air upstream of the combustor.

Installation. Turbofan engines are most commonly installed in a pod or nacelle which provides an efficient air inlet system and a streamlined, low-drag external configuration. Such pods are commonly hung below the aircraft's wings (**Fig. 3***a*) or less commonly above the wings, or at either side of the aft end of the fuselage. A third engine can be installed in a pod at the root of the vertical tail assembly. In single- or twin-engined military aircraft, the engines may be embedded in the fuselage (Fig. 3*b*). *See* JET PROPULSION; TURBINE PROPULSION.

Fredric F. Ehrich

Bibliography. P. Hill and C. Peterson, *Mechanics and Thermodynamics of Propulsion*, 2d ed., 1992; P. Jackson (ed.), *Jane's All the World's Aircraft*, annually; M. Kroes et al., *Aircraft Powerplants,* 7th ed., 1995; Rolls-Royce plc, *The Jet Engine,* 5th ed., 2005; I. E. Treager, *Aircraft Gas Turbine Engine Technology*, 3d ed., 1995.

Turbojet

A gas turbine power plant used to propel aircraft, where the thrust is derived within the turbo- machinery in the process of accelerating the air and products of combustion out an exhaust jet nozzle (**Fig. 1**). *See* GAS TURBINE.

Operating principle. In its most elementary form (**Fig. 2***a*), the turbojet operates on the gas turbine or Brayton thermodynamic cycle. The working fluid,

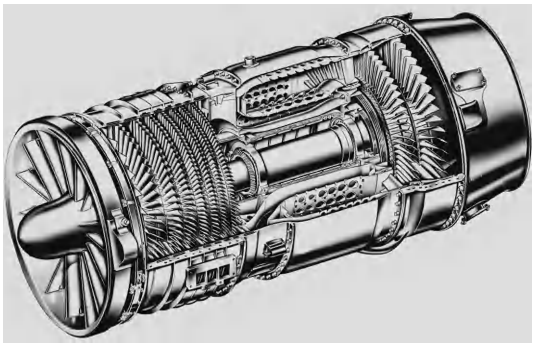


Fig. 1. Typical turbojet engine. (*GE Aircraft Engines*)

air drawn into the inlet of the engine, is first compressed in a turbo-compressor with a pressure ratio of typically 10:1 to 20:1. The high-pressure air then enters a combustion chamber, where a steady flow of a hydrocarbon fuel is introduced in either spray or vapor form and burned continuously at constant pressure. The exiting stream of hot high-pressure air, at an average temperature whose maximum value may range typically from 1800 to 2800°F (980 to 1540°C), is then expanded through a turbine, where energy is extracted to power the compressor. Because heat had been added to the air at high pressure, there is a surplus of energy left in the stream of combustion products that exits from the turbine and that can be harnessed for propulsion. *See* BRAYTON CYCLE; GAS TURBINE.

The turbojet engine derives propulsive thrust from the stream of hot high-pressure gas by ejecting the flow at high velocity through a jet nozzle. The principle of action was characterized by Isaac Newton in his second law of motion, in which he generalized the fact that the force F required to accelerate an object was simply proportional to the product of the object's mass m and the acceleration a, as in Eq. (1).

$$F = ma \tag{1}$$

In the case of the turbojet engine, rather than simply accelerate a single lump of mass, the engine acts to accelerate a continuous stream of air by raising its velocity from that at which it enters the engine, V_0, taken to be the flight speed of the aircraft, to that which leaves the exhaust nozzle, V_j. The force that the components of the engine must then exert on the airflow to achieve this acceleration is given by Eq. (2), where M is the rate of mass flow through

$$F = M(V_j - V_0) \tag{2}$$

the engine. The reaction to this force exerted by the engine's components on the airstream is then felt as thrust on the engine's mounts, which connect it to the aircraft. *See* FORCE.

There is a significant inefficiency in this process of thrust production that is manifest in the considerable amount of energy tied up in the high-temperature, high-velocity exhaust stream that leaves the engine without having been usefully harnessed. The propulsive efficiency of the thrust generation is found to be approximately given by Eq. (3), where $V_j > V_0$. The

$$\eta_p = \frac{2V_0}{V_0 + V_j} \tag{3}$$

relationship implies that high propulsive efficiency requires matching the jet velocity as closely as possible with the flight speed. In actual modern practice, therefore, simple turbojet engines are applied most often to aircraft that have very large portions of their flight profile at supersonic flight speed in order to most efficiently exploit the supersonic jet velocity and avoid prohibitively low propulsive efficiencies.

Configurations. The basic configuration of the turbojet (Fig. 2*a*) includes an inlet system, where the inlet stream of air, approaching the engine at a relative velocity equal to the flight speed of the aircraft, is decelerated to a lower velocity at which it can be drawn into the compressor. The compressor may be a multistaged axial-flow unit (Fig. 1), a centrifugal (radial-flow) compressor, or a combination axi-centrifugal component.

In modern usage, large multistage axial-flow compressors with airflows greater than 12–15 lb/s (5–7 kg/s) are typically 4–6% more efficient than

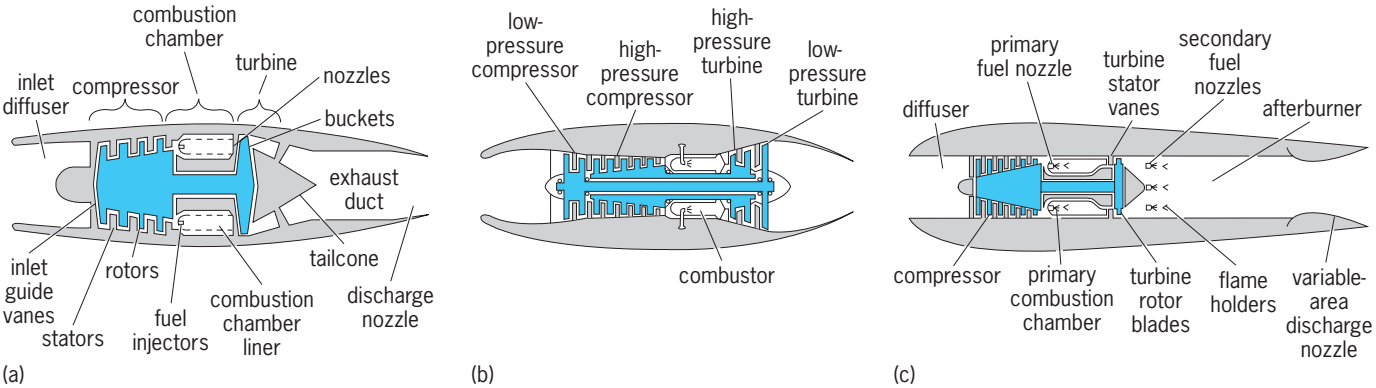


Fig. 2. Turbojet engine configurations. (*a*) Basic turbojet engine with axial-flow components. (*b*) Dual-rotor turbojet engine. (*c*) Afterburning turbojet engine.

the equivalent single-stage centrifugal compressor. Moreover, the requirement for very high tip speeds and for radially extended diffusion systems generally dictates frontal areas for centrifugal compressors that may be 30–50% greater than for the equivalent set of axial-flow stages. The large frontal area is particularly disadvantageous in fitting the engine into a low-drag nacelle for high flight speed, and in fitting a power producer within the confines of the bypass stream of a turbofan. The higher tip speeds of the centrifugal machines also dictate that they be considerably heavier than their axial-flow counterparts. For these reasons, axial-flow compressors are invariably used in modern engines except for very small units with airflows less than 12–15 lb/s (5–7 kg/s), where the small size of axial-flow blading involves disproportionate sacrifices in compressor efficiency and manufacturing complexity so that a centrifugal or axi-centrifugal unit becomes the configuration of choice.

Multistage compressors designed for high pressure ratios, above 6:1 or 7:1, have included cascades of stator vanes, each of which is pivoted so that its stagger angles may be varied at different power levels of operation to match individual-stage characteristics and avoid the phenomenon known as compressor stall or surge. In very high-pressure compressors, it may be necessary to split the compressor into two sections, each operating at its own independent rotative speed and driven by its own turbine stages (Fig. 2). *See* COMPRESSOR.

The combustors in early machines were often composed of an annular array of separate cans, each with its own fuel injector, but modern practice generally uses full annular chambers. The turbines are most often axial-flow in configuration, and have involved the application of the most advanced developments and fabrication techniques of high-temperature, high-strength alloys and, subsequently, nonmetallic materials, as well as complex internal passages to introduce cooling air, all to permit continuous operation at the very high gas temperatures that are required for high thermodynamic efficiency. The final component, the exhaust nozzle, is generally a fixed convergent unit, or, for very high values of engine pressure ratios and flight speeds, may include some divergence after the throat of the convergent section to most efficiently achieve supersonic jet velocities. *See* NOZZLE.

The thrust of a turbojet may be augmented by the addition of an afterburner (Fig. 2*c*). The afterburner also entails the use of a variable-area exhaust nozzle, since the area must be increased during afterburning to accommodate the additional heating of the exhaust air to a temperature in the range of 3400°F (1870°C), which greatly increases it volume. Although it permits the achievement of up to 50% more thrust from an engine that would otherwise be limited by the maximum temperature that can be withstood by the turbine, that thrust increment is achieved very inefficiently, that is, with very high fuel consumption. Afterburners are therefore included in engines only for short-term or emergency usage, such as in combat by military fighter aircraft. *See* AFTERBURNER.

Application. Turbojets were first used to power military combat aircraft in World War II. In the first decades after the war, turbojets were successfully applied to commercial transport aircraft, but were soon displaced by turboprop and turbofan aircraft, which were substantially more efficient at the subsonic and transonic flight speeds used, while exploiting the same principles of propulsion and enjoying much the same advantages of simplicity, light weight, and durability as the turbojet. *See* TURBOFAN; TURBOPROP.

Turbofans have also displaced turbojets in propulsion of military aircraft. Although combat aircraft are designed with supersonic flight capability where turbojets might be expected to find application, the aircraft must also spend a considerable amount of time and fuel at lower flight speeds for purposes of takeoff, climb, loiter, cruise, loiter, and approach and landing, so that low-bypass turbofans have evolved as the most efficient design arrangement. The single major current application of the turbojet has been to the supersonic commercial transport. But the next generation of supersonic transports may very well be powered by some type of turbofan, because of the need to fly substantial portions of its mission at subsonic speed and because of the difficult noise problem that a simple turbojet imposes.

Turbojets have retained a small niche in the aircraft propulsion spectrum, where their simplicity and low cost are of paramount importance, such as in short-range expendable military missiles, or where their light weight may be an overriding consideration, such as for lift jets in prospective vertical takeoff and landing aircraft. *See* AIRCRAFT PROPULSION; JET PROPULSION; TURBINE PROPULSION; VERTICAL TAKEOFF AND LANDING (VTOL).

Fredric F. Ehrich

Bibliography. P. Hill and C. Peterson, *Mechanics and Thermodynamics of Propulsion*, 2d ed., 1992; P. Jackson (ed.), *Jane's All the World's Aircraft*, annually; M. Kroes et al., *Aircraft Powerplants,* 7th ed., 1995; Rolls-Royce plc, *The Jet Engine,* 5th ed., 2005.

Turboprop

A gas turbine power plant producing shaft power to drive a propeller or propellers for aircraft propulsion. Because of its high propulsive efficiency at low flight speeds, it is the power plant of choice for short-haul and low-speed transport aircraft where the flight speeds do not exceed Mach 0.5–0.6. Developments in high-speed, highly loaded propellers have extended the range of propellers to flight speeds up to Mach 0.8–0.9, and there are prospects of these extremely efficient prop-fans assuming a much larger role in powering high-speed transport aircraft. *See* GAS TURBINE.

Power producer. As with all gas turbine engines, the basic power production in the turboprop is

accomplished in the gas generator or core of the engine, where a steady stream of air drawn into the engine inlet is compressed by a turbocompressor. The high-pressure air is next heated in a combustion chamber by burning a steady stream of hydrocarbon fuel injected in spray or vapor form. The hot, high-pressure air is then expanded in a turbine that is mounted on the same rotating shaft as the compressor and supplies the energy to drive the compressor. By virtue of the air having been heated at higher pressure, there is a surplus of energy in the turbine that may be extracted in additional turbine stages to drive

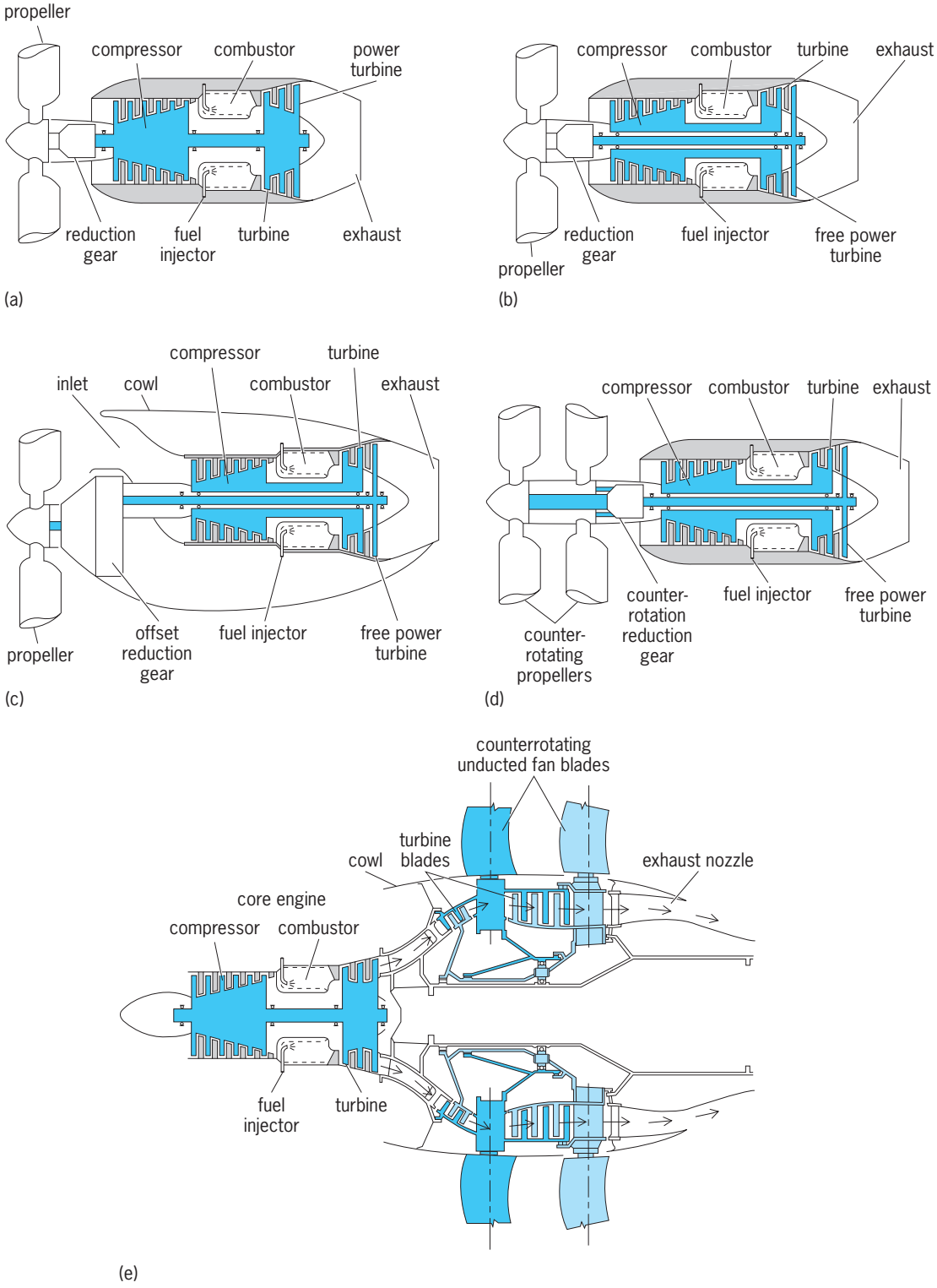


Fig. 1. Typical turboprop configurations. (*a*) Fixed-turbine turboprop. (*b*) Free-turbine turboprop. (*c*) Offset-gearbox turboprop. (*d*) Counterrotating turboprop. (*e*) Unducted fan, an ungeared, counterrotating, pusher turboprop. (*After R. C. Hawkins, Unducted fan for tomorrow's subsonic propulsion, Aerospace America, October 1984*)

a useful load, in this case a propeller or propellers.

A large variety of detailed variations are possible within the core. The compressor may be an axial-flow type, a centrifugal (that is, radial-flow) type, or a combination of stages of both types (that is, an axi-centrifugal compressor). In modern machines, the compressor may be split in two sections (a low-pressure unit followed by a high-pressure unit), each driven by its own turbine through concentric shafting, in order to achieve very high compression ratios otherwise impossible in a single spool. *See* COMPRESSOR.

Engine efficiency, measured as specific fuel consumption, is the rate of fuel flow (mass of fuel per unit time) per shaft power. A value of specific fuel consumption of 0.4 lb/(h)(hp) or 0.24 kg/(h)(kW) is typical of modern engines. The compressor pressure ratio is an important parameter in designing for a low specific fuel consumption. Pressure ratios of 12:1 to 20:1 are typical in current engines. Future engines may have pressure ratios as high as 40:1. The maximum value of the average temperature of the hot gases entering the turbine has a major effect on the specific power of the engine, the amount of horsepower generated per unit of airflow passing through the engine. The specific power is of fundamental importance since the size and weight of the engine depend strongly on the amount of airflow through the engine. A typical value of specific power in a modern engine might be 200 hp/(lb)(s) or 329 kW/(kg)(s). *See* SPECIFIC FUEL CONSUMPTION.

Configurations. The earliest and simplest turboprops had the turbine that extracted useful output power integrated (or fixed) on the same spool as the turbine that drove the compressor, forming a single dual-purpose turbine (**Fig 1***a*). In this configuration, the power producer spool drives the propeller shaft through a speed-reduction gearbox to accommodate the large-diameter, low-rotational-speed propeller.

There are several major disadvantages to this design having to do with starting, part-power efficiency, and windmilling operation of inoperative engines. These disadvantages are overcome by the free-turbine configuration (Fig. 1*b*). In this case, the load turbine is mounted on a spool separate from the gas generator and is free to drive the propeller through the propeller gearbox at a speed independent of the gas generator, and it is controlled to optimize propeller operation.

A common variant of these engines involves an offset reduction gear that calls for an offset inlet at one side of the gearbox (Fig. 1*c*).

A turboprop may also be used to drive two concentric counterrotating propellers (Fig. 1*d*). Counterrotating propellers are capable of providing significantly higher propulsive efficiency than single propellers.

Turboprops may also be designed in different configurations characterized as tractor or pusher types. In the tractor type the propeller is forward of the engine, which makes it most amenable to mounting with the propeller forward of the wing (**Fig. 2***a*).

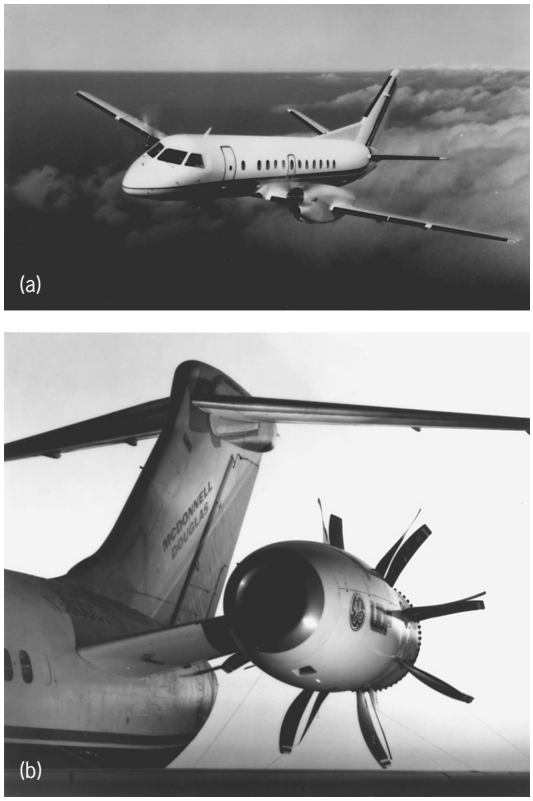


Fig. 2. Typical turboprop installations. (*a*) Conventional tractor-type single-rotation turboprop (*Saab Corp.*). (*b*) Unducted fan, a pusher-type counterrotating turboprop (*GE Aircraft Engines*).

The pusher type has the propeller mounted aft of the engine, which makes this engine amenable to being mounted aft of the wing (Fig. 2*b*).

An ingenious engine type, referred to as the unducted fan, involves the pusher configuration, combined with very efficient high-speed counterrotating propellers and a unique scheme that does away with the gearbox (Fig. 1*e* and Fig. 2*b*). *See* AIRCRAFT PROPULSION; PROPELLER (AIRCRAFT); TURBINE PROPULSION.

Fredric F. Ehrich

Bibliography. M. Kroes et al., *Aircraft Powerplants,* 7th ed., 1995; P. Jackson, (ed.), *Jane's All the World's Aircraft,* annually; Rolls-Royce plc, *The Jet Engine,* 5th ed., 2005; I. E. Treager, *Aircraft Gas Turbine Engine Technology,* 3d ed., 1995.

Turbulent flow

A fluid motion in which velocity, pressure, and other flow quantities fluctuate irregularly in time and space. **Figure 1** shows a slice of a water jet emerging from a circular orifice into a tank of still water. A small amount of fluorescent dye mixed in the jet makes it visible when suitably illuminated by laser light, and tags the water entering the tank. In this and similar realizations of the flow, there is a small region close to the orifice where the dye concentration does not vary with position, or with time at a given position. This represents a steady laminar state. Generally in laminar motion, all variations (if

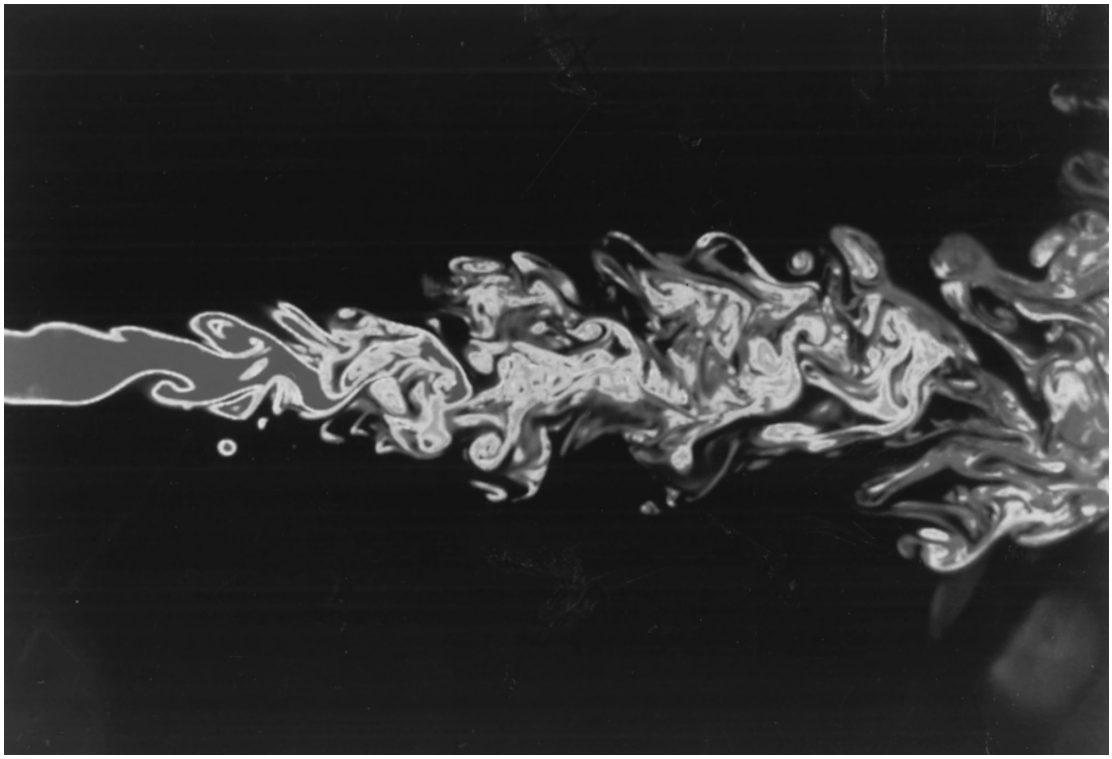


Fig. 1. Two-dimensional image of an axisymmetric water jet, obtained by the laser-induced fluorescence technique. A neodymium:yttrium-aluminum-garnet laser beam, shaped into a sheet of 250-micrometer thickness by using suitable lenses, was directed into a water tank into which the jet fluid, containing small amounts of uniformly dispersed fluorescing dye (sodium fluorescein), was emerging. The laser had a power density of 2×10^7 J/s per pulse and a pulse duration of about 10 nanoseconds. The flow is thus frozen to a good approximation. The region imaged extends from 2 diameters downstream of the orifice to about 18 diameters. The Reynolds number based on the orifice diameter, the velocity at the orifice, and the viscosity of water is about 2000. (*From R. R. Prasad and K. R. Sreenivasan, Measurement and interpretation of fractal dimension of the scalar interface in turbulent flows, Phys. Fluids A, 2:792–807, 1990*)

they occur at all) of flow quantities, such as dye concentration, fluid velocity, and pressure, are smooth and gradual in time and space. Farther downstream, the jet undergoes a transition to a new state in which the eddy patterns are complex, and flow quantities (including vorticity) fluctuate randomly in time and three-dimensional space. This is the turbulent state. *See* JET FLOW; LAMINAR FLOW.

Turbulence occurs nearly everywhere in nature [in the Earth's boundary layer (extending to more than a few hundred meters upward from the ground), the jet stream, cumulus clouds, rivers and oceans, the stellar atmosphere, interstellar gas clouds, and so forth] and in technology (in flow over airplanes, flow over turbine blades, flow of natural gas and oil in pipelines, combustion systems, and so forth). Two important characteristics of turbulence are the efficient dispersion and mixing of vorticity, heat, and contaminants. In flows over solid bodies such as airplane wings or turbine blades, or in confined flows through ducts and pipelines, turbulence is responsible for increased drag and heat transfer. Turbulence is therefore a subject of great engineering interest. On the other hand, as an example of collective interaction of many coupled degrees of freedom, it is also a subject at the forefront of classical physics. *See* DEGREE OF FREEDOM (MECHANICS); DIFFUSION; HEAT TRANSFER; PIPE FLOW; PIPELINE.

Figure 1 demonstrates the principal issues associated with turbulent flows. The first is the mechanism (or mechanisms) responsible for transition from the steady laminar state to the turbulent state even though, for both states, the governing equations (the Navier-Stokes equations) are the same, and the same smooth and symmetric boundary conditions are imposed on the flow everywhere. A second issue concerns the description of fully developed turbulence typified by the complex state far downstream of the orifice. To understand and describe the essential features of these spatial patterns, their interactions and temporal evolution, and to develop on this basis a rational theory capable of predicting flow features, is at the heart of turbulence theories. Finally, it is of technological importance to be able to alter the flow behavior to suit particular needs: Delaying transition to turbulence, or promoting it, or affecting the spread rate of the jet, or decreasing the drag of an airplane wing, or relaminarizing a turbulent flow some distance downstream of where it has become turbulent are some examples. Together, these three aspects—the origin of turbulence, the dynamics of fully developed turbulence, and the control of turbulent flows—constitute the turbulence problem. The problem assumes richer complexion when effects such as buoyancy, compressibility, electromagnetic conductivity, and chemical reactions are included. In spite of sustained efforts, turbulence

has remained unsolved. Less is known about eddy motions on the scale of centimeters and millimeters than about atomic structure on the subnanometer scale, reflecting the complexity of the turbulence problem. *See* NAVIER-STOKES EQUATION.

Origin of turbulence. A central role in determining the state of fluid motion is played by the Reynolds number. In general, a given flow undergoes a succession of instabilities with increasing Reynolds number and, at some point, turbulence appears more or less abruptly. It has long been thought that the origin of turbulence can be understood by sequentially examining the instabilities. This sequence depends on the particular flow and, in many circumstances, is sensitive to a number of details even if the gross features in a given flow are nominally fixed. The program of precisely identifying the various instabilities culminating in fully developed turbulence has not been carried out for any flow, but a careful analysis of the perturbed equations of motion has resulted in a good understanding of the first two instabilities (primary and secondary) in a variety of circumstances. *See* REYNOLDS NUMBER.

Since the onset of turbulence resembles the onset of complexity in nonlinear systems in general, the universality theories describing the onset of chaos have been thought to bear on the transition to turbulence in fluid flows. The spirit of universality is that, no matter what equations govern a low-dimensional system, its behavior in the vicinity of bifurcations depends on certain generic features in phase space. This issue is an active area of research, and the experience so far has been that the onset of chaos in special types of flows under special circumstances follows these theories, at least to a very good approximation, but the relation between chaos (or temporal stochasticity) and fluid turbulence (which possesses temporal as well as spatial randomness, and large-scale order underlying the latter) remains unclear. *See* CHAOS.

Fully developed turbulence. Some of the principal difficulties in fully developed turbulence are the following: (1) The equations of motion are nonlinear, possess no general solutions, and permit few statements of general validity to be made; there is no small parameter in the problem on the basis of which approximate solutions can be deduced rationally. (2) There is no well-understood working model of turbulence that replicates its essential properties. (3) Turbulent velocity fluctuations at small scales are strongly nongaussian, this being an essential feature. (4) The number of degrees of freedom is very large. *See* DISTRIBUTION (PROBABILITY).

An estimate of the number of degrees of freedom is given by the quantity $(L/\eta)^3$, where L is the characteristic size of the large eddy in the flow (or an upper bound for the eddy scale either excited by inherent instability or forced by an outside agency), and η is the smallest scale below which all eddy motions are damped by viscosity. This number increases with the flow Reynolds number according to its 9/4 power. Three-quarters of the way downstream from the orifice in Fig. 1, the ratio L/η is of the order of 100. Although such flows can now be computed directly, the prospect at high Reynolds numbers remains discouraging—for the atmosphere, L is of the order of a few kilometers whereas η is of the order of a millimeter—even though computational capabilities have continued to increase rapidly and parallel processing has been much considered as a tool for expanding the scope of computation. *See* CONCURRENT PROCESSING; SUPERCOMPUTER.

Quite often in engineering, the detailed motion is not of interest, but only the long-time averages or means, such as the mean velocity in a boundary layer, the mean drag of an airplane or pressure loss in a pipeline, or the mean spread rate of a jet. It is therefore desirable to rewrite the Navier-Stokes equations for the mean motion. The basis for doing this is the Reynolds decomposition, which splits the overall motion into the time mean and fluctuations about the mean. These macroscopic fluctuations transport mass, momentum, and matter (in fact, by orders of magnitude more efficiently than molecular motion), and their overall effect is thus perceived to be in the form of additional transport or stress. This physical effect manifests itself as an additional stress (called the Reynolds stress) when the Navier-Stokes equations are rewritten for the mean motion (the Reynolds equations). The problem then is one of prescribing the Reynolds stress, which contains the unknown fluctuations in quadratic form. A property of turbulence is that the Reynolds stress terms are comparable to the other terms in the Reynolds equation, even when fluctuations are a small part of the overall motion. An equation for the Reynolds stress itself can be obtained by suitably manipulating the Navier-Stokes equations, but this contains third-order terms involving fluctuations, and an equation for third-order terms involves fourth-order quantities, and so forth. Thus, at any stage of the process, which can be continued indefinitely, there are more unknowns than equations; that is, the system of equations is not closed. This is the closure problem in turbulence. The Navier-Stokes equations are themselves closed, but the presence of nonlinearity and the process of averaging result in nonclosure.

Given this situation, much of the progress in the field has been due to (1) exploratory experiments and numerical simulations of the Navier-Stokes equations at low Reynolds numbers; and (2) plausible hypotheses in conjunction with dimensional reasoning, scaling arguments, and their experimental verification.

Experiments, for long the central tool of research in turbulence, are limited to measuring a small number of parameters at a few positions in high-Reynolds-number flows. Low-Reynolds-number flows (at least some of their features) can be quantitatively mapped in three dimensions by using lasers and advanced optical techniques; this Reynolds number range is also the one for which numerical simulations are currently possible. From a combination of such studies, it has been learned, among other things, that the magnitude of the dissipation rate of turbulent kinetic

energy is independent of viscosity (even though viscosity is essential for dissipation); that the boundary between the turbulent and nonturbulent regions in high-Reynolds-number free shear flows such as jets is sharp and fractallike; that the dissipation of energy is highly intermittent in space; that some events that appear to be dynamically significant are also intermittent and perhaps quasicyclic; and that, when the flow scales are suitably coarse-grained, some degree of spatial order on scales of order L is visible even at very high Reynolds numbers, especially if the flow development in time is observed. The true significance of each of these features in accomplishing transport is still under active research. *See* FRACTALS; VISCOSITY.

The intermittency in space of the turbulent energy dissipation is shown at moderate (**Fig. 2***a*) and high (Fig. 2*b*) Reynolds numbers. The signal becomes less space-filling or more intermittent as the Reynolds number increases. In particular, the big spikes in Fig. 2*b* are many times larger than the corresponding ones in Fig. 2*a*. This intermittency, representing the fact that there is a limit to the mixing at small scales, is believed to be an important feature of turbulence. It is not entirely clear how this feature arises dynamically, but it can be modeled well by a simple multiplicative process.

A classic and celebrated hypothesis is the concept of local isotropy, which assumes that small scales of motion are isotropic irrespective of the gross orientation of the mean flow, and thus possess some universality. A second notion is the matchability between behaviors of highly disparate scale ranges so that a functional form for average quantities of interest can be determined for the intermediate scale range. For example, in the turbulent boundary layer over a flat wall, this type of argument leads to a logarithmic variation of mean velocity with height for heights large compared to the viscous scale and small compared to the overall thickness of the boundary layer. Similarly, intermediate scales that are large compared to η but small compared to L (the so-called inertial range) are expected to possess self-similarity, leading to power-law variations for the spectral densities of energy, dissipation, variance of concentration fluctuations, and so forth. These predictions have received experimental support, and, in fact, they seem to be realized under conditions where they are not necessarily expected to be valid, based on first principles. This raises the possibility that the basic theoretical arguments of this type (which, incidentally, do not make much use of the Navier-Stokes equations) have a much wider range of validity. It should be emphasized that these arguments predict an absolute number in the inertial range, but not elsewhere. *See* BOUNDARY-LAYER FLOW.

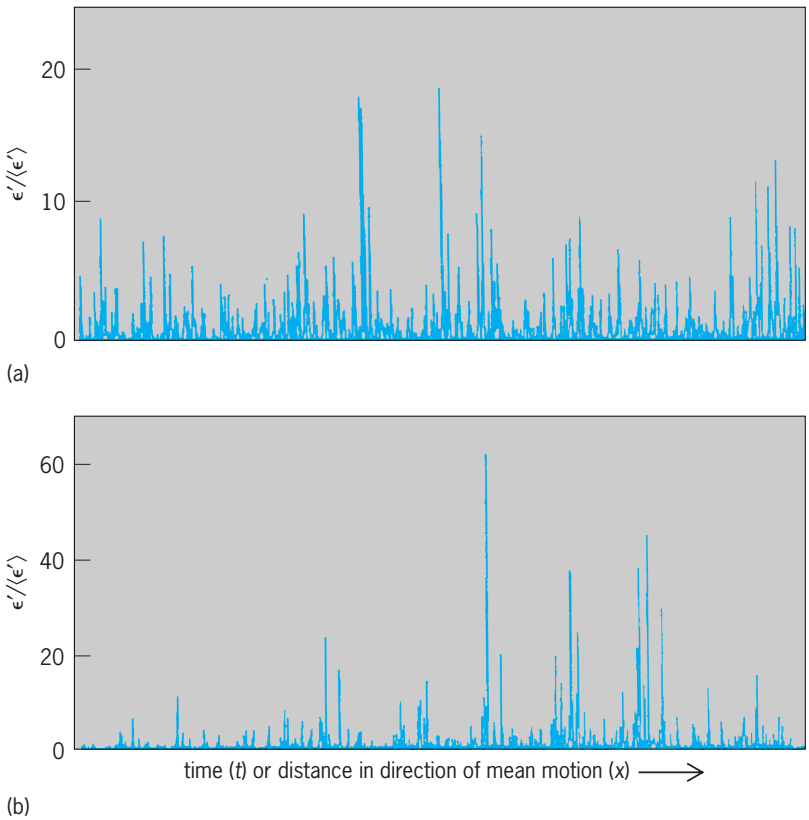


Fig. 2. Typical signals of ϵ', a component of the turbulent energy dissipation, normalized by its mean value $\langle\epsilon'\rangle$. (*a*) Signal obtained in a laboratory turbulent boundary layer at a moderate Reynolds number (defined suitably). (*b*) Signal obtained in the atmospheric surface layer at a high Reynolds number. (*After C. Meneveau and K. R. Sreenivasan, Simple multifractal cascade model for fully developed turbulence, Phys. Rev. Lett., 59:1424–1427, 1987*)

Certain specific questions of engineering interest can be answered quickly by modeling the Reynolds stress in a variety of ways and closing the Reynolds equations. The earliest model, based on analogies with molecular motion, postulated that the physical motion of eddies over well-defined distances accomplishes transport. The resulting eddy viscosity—essentially the product of a characteristic velocity scale and a length scale of turbulent motion—is the analog of molecular viscosity. Even though there are circumstances where the eddy-viscosity approach works roughly, the concept has many drawbacks and is not very useful as a general idea; in any case, the eddy viscosity varies from flow to flow and from point to point in a given flow. In the next level of models, separate equations are written for the length and velocity scales making up eddy viscosity, but there are several unknown coefficients that have to be determined empirically. There are more complex models, all of which resort to empiricism of dubious quality at some level. In spite of this drawback, they are quite useful once the limits of their validity have been established. At present, they represent a practical way of computing high-Reynolds-number flows of technological interest. Another fruitful approach is the large-eddy simulation, which models the small-scale motion but simulates large eddies on the computer.

Control of turbulent flows. Unlike several other issues in turbulence, questions concerning flow control can be posed in specific terms. However, because of this specificity, a broad-brush approach to the control problem encompassing all circumstances is unlikely to succeed. Some typical objectives are the

reduction of drag of an object such as an airplane wing, the suppression of combustion instabilities, and the suppression of vortex shedding behind bluff bodies. A surge of interest in flow control is due in part to the discovery that some turbulent flows possess a certain degree of spatial coherence at the large scale. An example of successful control, based on an unrelated idea, is the reduction of the skin friction on a flat plate by making small longitudinal grooves, the so-called riblets, on the plate surface, imitating shark skin.

Prospects. Progress in the turbulence problem depends on the capability to make accurate measurements in high-Reynolds-number flows, the increase in computer power, the invention of new tools for handling large streams of stochastic data, and a judicious combination of all of them. Unfortunately, simply computing or making measurements in a highly nonlinear system such as turbulent flow does not always add to understanding. Although several new analytical tools, graphical display capabilities, and data-compression and data-handling techniques are being explored, it is difficult to predict what true progress is likely to occur through the 1990s. It is clear, however, that turbulence will spur important activity in a number of disciplines at the forefront of science and technology; conversely, it will benefit from them. In the long run, perhaps, all these tools can enhance the qualitative understanding of turbulence; to obtain quantitative data in a specific context, reliance may always have to be placed on experiment, as well as modeling and computation that use this qualitative knowledge in a sensible way. This would resemble to some extent the situation in quantum chemistry. *See* FLUID-FLOW PRINCIPLES.

K. R. Sreenivasan

Bibliography. H. Bai-Lin (ed.), *Chaos II*, 1990; D. M. Bushnell and C. B. McGinley, Turbulence control in wall flows, *Annu. Rev. Fluid Mech.*, 21:1–20, 1989; R. J. Garde, *Turbulent Flow*, 1994; J. L. Lumley, *Whither Turbulence? Or, Turbulence at Crossroads*, 1990; H. L. Swinney and J. P. Gollub (eds.), *Hydrodynamic Instabilities and the Transition to Turbulence*, 2d ed., 1985; H. Tennekes and J. L. Lumley, *A First Course in Turbulence*, 1972; A. A. Townsend, *The Structure of Turbulent Shear Flows*, 2d ed., 1976.

Turmeric

A dye or a spice obtained from the plant *Curcuma longa*, which belongs to the ginger family (Zingiberaceae). It is a stout perennial with short stem, tufted leaves, and short, thick rhizomes which contain the colorful condiment. As a natural dye, turmeric is orange-red or reddish brown, but it changes color in the presence of acids or bases. As a spice, turmeric has a decidedly musky odor and a pungent, bitter taste. It is an important item in curry and is used to flavor and color butter, cheese, pickles, and other food. *See* SPICE AND FLAVORING; ZINGIBERALES.

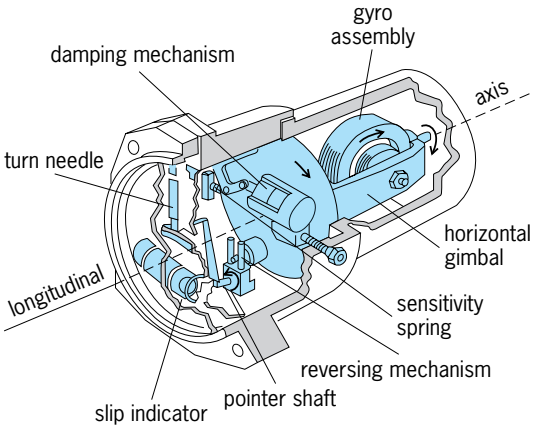
Perry D. Strausbaugh; Earl L. Core

Turn and bank indicator

A combination instrument which provides an aircraft pilot with two distinct pieces of information: the aircraft's rate of turn about the vertical axis, and the relationship between this rate and the aircraft's angle of bank. It is also known as the needle and ball indicator or the turn and slip indicator.

The turn needle is operated by a gyroscope and indicates the rate at which the aircraft is turning about the vertical axis in degrees per second. Semirigid mounting of the gyro permits it to rotate freely about the lateral and longitudinal axes while restricting motion about the vertical axis (see **illus.**). In a turn, gyroscopic precession causes the rotor to tilt in the direction opposite the turn with a magnitude proportional to the turn rate. A mechanical linkage converts this precession to reversed movement of a turn needle, thus indicating proper turn direction. A spring attached between the gyro assembly and the instrument case holds the gyro upright when precession force is not present and allows calibration for a given turn rate, while a damping mechanism is included to prevent excessive oscillation of the turn needle. Because of the criticality of this instrument under instrument flight conditions, it is normally powered by a separate energy source from the other flight instruments. *See* GYROSCOPE.

The bank or slip indicator is a simple inclinometer consisting of a curved glass tube containing fluid and a black ball bearing which is free to move in the fluid. The fluid provides dampening for the ball movements, and a small projection at one end of the tube contains an air bubble to compensate for fluid expansion during temperature changes. The tube is curved upward at the ends so the ball seeks the lowest point in wings-level flight. The indicator is actually a balance indication, showing the relationship between the rate of turn and the angle of bank of the aircraft. During a turn the ball experiences two forces in the horizontal plane: the component of the lift force in the horizontal plane pulling to the inside of the turn, which is created by the bank angle or the tilt of the lift force; and the centrifugal force



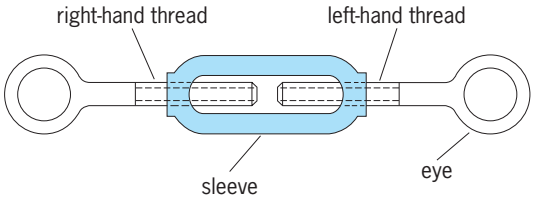
Typical turn and bank indicator. (*U.S. Air Force*)

created by the turn rate producing a force to the outside of the turn. It is the sum of these two forces that the ball displays. *See* AIRCRAFT INSTRUMENTATION; ROTATIONAL MOTION.

Grady W. Wilson

Turnbuckle

A device for tightening a rod or wire rope. Its parts are a sleeve with a screwed connection at one end and a swivel at the other or, more commonly, a sleeve with screwed connections of opposite hands (left and right) at each end so that by turning the sleeve, the connected parts will be drawn together, taking up slack and producing necessary tension (see **illus.**). Types of ends available are hook, eye, and



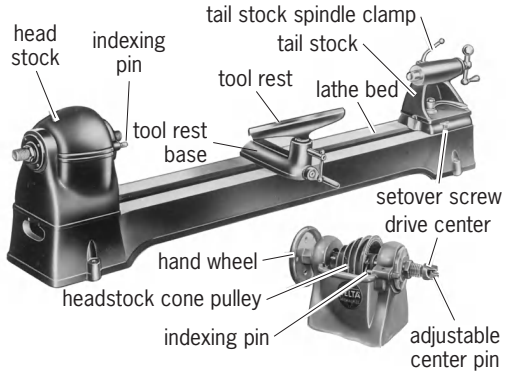
Turnbuckle with eyes.

clevis. The turnbuckle can be connected at any convenient place in the rod or rope, and several may be used in series if required. *See* ROPE; SCREW THREADS.

Paul H. Black

Turning (woodworking)

The shaping of wood by rotating it in a lathe and cutting it with a chisel. The lathe consists essentially of a bed on which are mounted a headstock, a tailstock, and a tool rest (see **illus.**). The headstock is rotated by a motor and holds one end of the wood to be turned. The tailstock holds the other end of the wood, allowing it to rotate freely. The tool rest provides a fixed guide along which the operator can handle the chisels if the turning is by hand, or along which the tool is driven if the turning is mechanized.



Wood-turning lathe and detail of headstock. (*Delta*)

Variously shaped chisels are used, all with longer handles than the woodworking chisels intended to be driven by mallet, and thus providing a firmer grip. A gouge is used for roughing cuts, for example, in turning the work as it comes from the sawmill to a nearly cylindrical shape. A skew chisel with a straight cutting edge is used in finishing. A parting chisel with a tapered shank is used in separating the finished work from the stock. *See* WOODWORKING.

Alan H. Tuttle

Bibliography. S. Hogbin, *Wood Turning*, 1980.

Turnip

The plant *Brassica rapa*, or *B. campestris* var. *rapa*, a cool-season, hardy crucifer of Asiatic origin belonging to the order Capparales and grown for its enlarged root and its foliage, which are eaten cooked as a vegetable (see **illus.**). The plant is an annual when planted early, a biennial if seeded late in the summer. Propagation is by seed. Popular white-fleshed varieties (cultivars) grown for their roots are Purple Top Globe and White Milan; Yellow Globe and Golden Ball are common yellow-fleshed

Turnip, of the Brassicaceae family.

varieties. Shogoin is a popular variety grown principally in the southern United States for turnip greens. Turnip harvesting begins when the roots are 2–3 in. (5–7.5 cm) in diameter, usually 40–70 days after planting. Principal areas of production in the United States are in the South. *See* CAPPARALES; RUTABAGA.

H. John Carew

Turnip and rutabaga can be damaged by several diseases throughout their growth period. Seedling diseases caused by soil-inhabiting fungi may kill or debilitate young plants soon after emergence. Proper planting-site selection and preparation, along with seed spacing to avoid crowding, can minimize losses. Leaf-spotting diseases are caused by fungi that produce wind- or rain-disseminated spores. Alternaria leaf spot, anthracnose, and blackleg are important foliar diseases and may be controlled by protective fungicides. Turnip leaves grown for food are

particularly susceptible to leafspot damage. Blackleg may also damage roots before harvest and during storage.

Bacterial blackrot is often a destructive disease. Symptoms, damage, and control measures are similar to those for blackrot of cabbage.

Clubroot is a serious disease of worldwide distribution. The causal organism, which persists for several years in the soil, infects young roots; severely deformed, unattractive turnips and rutabagas are produced as a result. Infested soil should be avoided, or disease-resistant cultivars should be planted. Deformed, mottled leaves or stunted plants may be caused by several insect-transmitted viruses.

Turnips and rutabagas require adequate boron to prevent unsightly, internal breakdown of root and stem tissues. *See* CABBAGE; PLANT PATHOLOGY.

J. O. Strandberg

Bibliography. G. R. Dixon, *Vegetable Crop Diseases: American Edition*, 1981; A. F. Sherf and A. A. McNab, *Vegetable Diseases and Their Control*, 2d ed., 1986; J. C. Walker, *Diseases of Vegetable Crops*, 1952; J. C. Walker, R. H. Larson, and A. L. Taylor, *Diseases of Cabbage and Related Plants*, USDA Agr. Handb. 144, 1958.

Turquoise

A mineral of composition $CuAl_6(PO_4)_4(OH)_8 \cdot 5H_2O$ in which considerable ferrous ion (Fe^{2+}) may substitute for copper. Ferric ion (Fe^{3+}) may also substitute for part or all of the aluminum (Al), forming a complete chemical series from turquoise to chalcosiderite [$CuFe_6(PO_4)_4(OH)_8 \cdot 5H_2O$]. Turquoise with a strong sky-blue or bluish-green to apple green color is easily recognized, and such material is commonly used as a gem. Some variscite, of composition $AlPO_4 \cdot 2H_2O$ with minor chemical substitutions of Fe^{3+} and or chromium ion (Cr^{3+}) for aluminum and with a soft, clear green color, may be marketed as green turquoise.

Crystals of turquoise are rare; they are triclinic in symmetry, with space group $P\bar{1}$. Most turquoise is massive, dense, and cryptocrystalline to fine-granular. It commonly occurs as veinlets or crusts and in stalactitic or concretionary shapes. It has a hardness on the Mohs scale of about 5 to 6 and a vitreous to waxy luster. The distinctive light blue coloration of much turquoise is the result of the presence of cuprous ion (Cu^{2+}); limited substitution of the copper by Fe^{2+} produces greenish colors. *See* CRYSTAL STRUCTURE.

Occurrence. Turquoise is a secondary mineral, generally formed in arid regions by the interaction of surface waters with high-alumina igneous or sedimentary rocks. It occurs most commonly as small veins and stringers traversing more or less decomposed volcanic rocks. Since the times of antiquity, turquoise of very fine quality has been produced from a deposit in Persia (now Iran) northwest of the village of Madeň, near Nishapur. It occurs also in Siberia, Turkistan, China, the Sinai Peninsula, Germany, and France.

The southwestern United States has been a major source of turquoise, especially the states of Nevada, Arizona, New Mexico, and Colorado. Extensive deposits in the Los Cerillos Mountains, near Santa Fe, New Mexico, were mined very early by Native Americans and were a major early source of gem turquoise. However, much of the gem-quality turquoise has been depleted in the Southwest. Some mining districts that still supply gem material are the Sleeping Beauty district between Globe and Miami, Arizona; Morenci, Arizona; Crow Springs, Stormey Mountain, and Blue Moon district, Nevada; Menassa, Colorado; and Hachita and Oro Grande, New Mexico.

Imitations. Because of much demand for turquoise at relatively low prices and because of the scarcity of gem-quality turquoise, which is generally high priced, many materials that are not completely natural or are turquoise-colored imitations have appeared on the market. These varieties are known as stabilized, oil-treated, treated, reconstituted, and imitation turquoise.

Stabilized turquoise is a poor-quality natural turquoise (often referred to as chalk in the trade) that has been chemically impregnated and hardened with organic resins to improve the color and the hardness of the final product. Such material becomes very workable, permanently hardened, and stable, and can be very attractive. Major sources of such low-grade turquoise are found in Mexico and China.

Oil-treated turquoise is a material in which the color of the natural turquoise has been enhanced by impregnation with oil, paraffin, or oil-based polishes. Such treatment generally results in only temporary improvement of color, with subsequent fading.

Treated turquoise is a natural or stabilized turquoise that has been altered (dyed) to produce a change in the coloration of the natural material.

Reconstituted turquoise is fabricated by bonding dust, turquoise particles, or nuggets together with plastic resins.

Imitation turquoise results from a process in which a natural compound is treated or a synthetic compound is manufactured to closely approximate turquoise's appearance. Included are turquoise-colored plastics, glass enamel, and dyed chalcedony.

Stabilized, oil-treated, and reconstituted turquoise can be distinguished by quantitative chemical microanalysis for carbon (C), hydrogen (H), and nitrogen (N), the elements that were added to the original material in the various modification processes. Treated turquoise can generally be recognized by a well-trained eye. Physical testing usually suffices to distinguish the materials comprising imitation turquoise. *See* GEM; PHOSPHATE MINERALS.

Cornelis Klein

Bibliography. C. S. Hurlbut, Jr., and R. C. Kammerling, *Gemology*, 1991; C. Klein, and C. S. Hurlbut, Jr., *Manual of Mineralogy*, 21st ed., 1993; C. Palache, H. Berman, and C. Frondel (eds.), *System of Mineralogy*, vol. 2, 7th ed., 1951.

Twilight

The period between sunset and darkness in the evening and darkness and sunrise in the morning. The following statements apply to evening twilight; the reverse would apply to morning twilight.

The characteristic light is caused by atmospheric scattering, which transmits sunlight to the observer for some time after the Sun has set. It depends geometrically on latitude, longitude, and elevation of the observer, and on the time of the year. Physically it depends also on local conditions, particularly the weather. *See* METEOROLOGICAL OPTICS.

Three degrees of twilight are conventionally distinguished. Civil twilight ends when the center of the Sun is 6° below the horizon; if the sky is clear, it is usually practicable to carry on ordinary outdoor occupations without artificial light during civil twilight. Nautical twilight ends when the depression of the Sun is 12°; at this time both the horizon and the brighter stars are visible. Astronomical twilight ends when the depression of the Sun is 18°; at this time no trace of illumination by the Sun appears in the sky. As thus defined, the times of ending of the three sorts of twilight can be precisely calculated.

Since the angle at which the Sun sets depends on an observer's latitude, twilights are relatively short near the Equator, where the Sun sets perpendicularly, and can last for days near or above the Arctic and Antarctic circles. *See* ANTARCTIC CIRCLE; ARCTIC CIRCLE.

Gerald M. Clemence; Jay M. Pasachoff

Bibliography. A. B. Meinel and M. P. Meinel, *Sunsets, Twilights, and Evening Skies*, 1991; G. V. Rozenberg, *Twilight*, 1966; V. J. Schaefer and J. A. Day, *A Field Guide to the Atmosphere*, 1999.

Twinkling stars

A phenomenon by which light from the stars, as it passes through fluctuations in the Earth's atmosphere, is rapidly modulated and redirected to make the starlight appear to flicker. Although it is familiar to those who have looked with the unaided eye at the night sky, the twinkling phenomenon affects all wavelengths that manage to penetrate the Earth's atmosphere, from the visible to the radio wavelengths. At visible wavelengths, atmospheric fluctuations are caused predominantly by temperature irregularities along the line of sight. Minor contributions are made by irregularities in atmospheric density and in water vapor content. All such irregularities introduce slight changes in the index of refraction of air, and these changes affect light waves in two ways: they modulate the intensity of the light, and they deflect the light waves in one direction and then another. An analogous phenomenon is often observed when light grazes across the surface of a hot highway: light is bent and distorted by pockets of hot air rising over the pavement's surface. At radio wavelengths, electron density irregularities in the ionosphere modulate and redirect radio waves. *See* REFRACTION OF WAVES.

Unaided-eye observations. High-altitude winds and local air currents carry atmospheric irregularities across the line of sight. Because the entrance pupil of the human eye is much smaller than the characteristic size of the atmospheric irregularities, any twinkling that is visible to the unaided eye is caused primarily by the modulation of the light intensity and not by the deflection of the light waves. Modulation is best understood as an interference phenomenon acting on adjacent light waves. Like random waves in water, sometimes the waves add constructively, causing the star to brighten, and other times the waves add destructively, causing the star to dim. *See* INTERFERENCE OF WAVES.

While stars twinkle, planets generally do not. This difference occurs because stars are essentially infinitesimally small points of light. All light from a star travels along the same path through the atmosphere, and all the light from a star is modulated simultaneously. Planets have a definite angular extension. Light from one side of the planet traverses one path through the atmosphere, while light from the other side of the planet traverses another path through the atmosphere. Along each path the modulation of light is different, so that when all the light is added together by the human eye the total modulation or twinkling is averaged out. *See* PLANET; STAR.

Telescope observations. The twinkling phenomenon is of utmost interest to astronomers who view the skies from ground-based telescopes. While modulation variations are present, it is the deflection of light that causes the most serious problems. The entrance pupil of an optical telescope is often much larger than the characteristic size of atmospheric irregularities. The composite star image produced by a large telescope is a blurry circle that results when the randomly deflected light waves are added together in an extended time exposure. To diminish atmospheric effects, telescopes are built on high mountains, and are placed at least 30–45 m (100–150 ft) above the ground. The best observatories are those where the twinkling phenomenon is minimized. Examples include the Mauna Kea Observatory in Hawaii and observatories in the foothills of the Andes Mountains in northern Chile. *See* OBSERVATORY, ASTRONOMICAL.

To completely remove the twinkling effects of the atmosphere, there are two alternatives. The first is to place a telescope in orbit above the atmosphere, as with the Hubble Space Telescope. The second alternative is to monitor the random deflections of the atmosphere and, within the telescope, to bend the deflected light back onto its original path. This optical technique is given the name adaptive optics. *See* ADAPTIVE OPTICS.

Radio sources. Small radio sources have been found to vary rapidly in brightness because of the flow of density fluctuations in the solar wind across the path between the radio source and the Earth. The method of interplanetary scintillations has proved a useful means of probing the interplanetary medium near the Sun and also of establishing the very small size of some of the radio sources. Radio sources

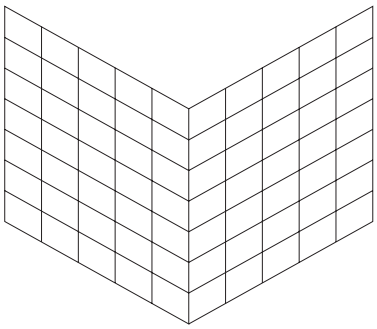
whose angular diameters are large do not show interplanetary scintillation at all, just as planets do not share in the stellar twinkling. Pulsars show erratic changes of amplitude in their radio emission, caused by radio scintillation both near the neutron star and in interstellar space. *See* PULSAR; RADIO ASTRONOMY; SOLAR WIND. Laird A. Thompson

Bibliography. J. M. Beckers, Adaptive optics for astronomy: principles, performance, and applications, *Annu. Rev. Astron. Astrophys.*, 31:13-62, 1993; H. G. Booker and W. E. Gordon, A theory of radio scattering in the troposphere, *Proc. Inst. Radio Eng.*, 38:401–412, 1950; B. J. Rickett, Interstellar scattering and scintillation of radio waves, *Annu. Rev. Astron. Astrophys.*, 15:479–504, 1977; N. J. Woolf, High resolution imaging from the ground, *Annu. Rev. Astron. Astrophys.*, 19:367–398, 1981; A. T. Young, Seeing: Its cause and cure, *Astrophys. J.*, 189:587–604, 1974.

Twinning (crystallography)

A process in which two or more crystals, or parts of crystals, assume orientations such that one may be brought to coincidence with the other by reflection across a plane or by rotation about an axis. Crystal twins represent a particularly symmetric kind of grain boundary; however, the energy of the twin boundary is much lower than that of the general grain boundary because some of the atoms in the twin interface are in the correct positions relative to each other.

In the general grain boundary, all the neighbors of the atoms of the interface are in distorted positions. The usual definition of a twin relationship between two crystals states that there exists a set of parallel equivalent crystal planes of atoms which is common to both twins, but that rows of atoms are discontinuous across the interface. Quite commonly, twins are mirror images of each other, as in the **illustration**. Also, it is common for a twin in a crystal to leave the nearest neighbors of the atoms in the interface unchanged in orientation, but to place the atoms in the second neighbor shell in altered positions. This feature is also true of the twin in the illustration. It is sometimes possible to create a twin in a crystal by putting an external stress on the crystal; in other cases, twins are found "grown in." *See* CRYSTAL GROWTH. Robb M. Thomson



Example of a twinned crystal. One atom plane is common to each half of the crystal, but the other lines of atoms suffer a discontinuity at the twin boundary.

Bibliography. A. Holden and P. S. Morrison, *Crystals and Crystal Growing*, 1982; R. Raj and S. L. Sass (eds.), Interface Science and Engineering. '87, *Journale de Physique*, vol. 49, colloque C5, 1989; J. P. Van der Eerden, *Fundamentals of Crystal Growth*, 1993.

Twins (human)

Two babies born to a mother at one birth. Knowledge about the biological bases of twinning, as well as sophistication in techniques for data collection, research design, and analysis, have increased substantially in recent years. Consequently, twin research has been incorporated into a growing number of behavioral science and medical science research programs.

Biology of twinning. There are two types of twins, monozygotic and dizygotic. Members of a twin pair are called co-twins. Controversy surrounding the definition of a twin arose with the advent of reproductive technologies enabling the simultaneous fertilization of eggs, with separate implantation. The 1996 cloning of Dolly the lamb in Scotland directed attention toward the promises and pitfalls of possible human cloning. The unique "twinlike" relationships that would result between parents and cloned children (who would be genetically identical to their parents) also challenge current conceptions of twinship. Monozygotic twins are clones (genetically identical individuals derived from a single fertilized egg), but parents and cloned children would not be twins for several reasons, such as their differing prenatal and postnatal environments. *See* REPRODUCTIVE TECHNOLOGY.

Monozygotic twins. The division of a single fertilized egg (or zygote) between 1 and 14 days postconception results in monozygotic twins. They share virtually all their genes and, with very rare exception due to unusual embryological events, are of the same sex.

A common assumption is that because monozygotic co-twins have a shared heredity, their behavioral or physical differences are fully explained by environmental factors. These environmental influences may occur during the prenatal, perinatal, or postnatal periods. However, monozygotic twins are never exactly alike in any measured trait, and may even differ for genetic reasons. For example, the random "shutting down" of one X chromosome in every female cell soon after conception (a process called lyonization) can cause monozygotic female twins to differ in X-linked traits, such as color blindness. *See* COLOR VISION; SEX-LINKED INHERITANCE.

Sometimes chromosomes fail to separate after fertilization, causing some cells to contain the normal chromosome number (46) and others to contain an abnormal number. This process, called mosaicism, results in monozygotic co-twins who differ in chromosomal constitution. These unusual cellular processes explain the presence of monozygotic pairs in

which one co-twin is normal while the other shows a genetic anomaly reflecting a mixture of normal and abnormal cells. A rare case of monozygotic triplets including two healthy males and a co-triplet with Turner's syndrome (loss of one X chromosome) has been reported. *See* MOSAICISM.

There are several other intriguing variations of monozygotic twinning. Splitting of the zygote after day 7 or 8 may lead to mirror-image reversal in certain traits, such as handedness or direction of hair whorl. The timing of zygotic division has also been associated with placentation. Monozygotic twins resulting from earlier zygotic division have separate placentae and fetal membranes (chorion and amnion), while monozygotic twins resulting from later zygotic division share some or all of these structures. Compared with two-chorion monozygotic twins, monozygotic twins sharing a chorion are more likely to be born prematurely, to differ in birth weight, and to die early (in extreme cases). Associations between mirror-image reversal and placentation are expected, but relationships among these events do not appear to be straightforward. Should the zygote divide after 14 days, the twins may fail to separate completely. This process, known as conjoined twinning, occurs in approximately 1 monozygotic twin birth out of 200. The many varieties of conjoined twins differ as to the nature and extent of their shared anatomy. Approximately 70% of such twins are female. There do not appear to be any predisposing factors to conjoined twinning.

Dizygotic twins. Dizygotic twins result when two different eggs undergo fertilization by two different spermatozoa, not necessarily at the same time. Dizygotic twins share, on average, 50% of their genes, by descent, so that the genetic relationship between dizygotic co-twins is exactly the same as that of ordinary brothers or sisters. Theoretically, dizygotic twins may share between 0 and 100% of their genetic makeup, but most are close to the 50% average. Some dizygotic co-twins share higher or lower proportions of genes for certain traits, so they may be more or less similar in those traits. Dizygotic twins may be of the same or opposite sex, outcomes that occur with approximately equal frequency.

There are some unusual variations of dizygotic twinning. There is the possibility of polar body twinning, whereby divisions of the ovum prior to fertilization by separate spermatozoa could result in twins whose genetic relatedness falls between that of monozygotic and dizygotic twins, or between dizygotic twins and unrelated individuals. Blood chimerism, another variation, refers to the presence of more than one distinct red blood cell population, derived from two zygotes, and has been explained by connections between two placentae. In humans, chimerism can occur in dizygotic twins. New techniques estimate that chimerism occurs in 8% of dizygotic twins and 21% of dizygotic triplets. Superfecundation is the conception of dizygotic twins following separate fertilizations, usually within several days, in which case each co-twin could have a different father. Superfecundation may cause significant developmental discrepancies between co-twins due to their differing paternal heredity. Superfetation, which refers to multiple conceptions occurring several weeks or even one month apart, may be evidenced by delivery of full-term infants separated by weeks or months and by the birth or abortion of twin infants displaying differential developmental status. *See* FERTILIZATION; OOGENESIS.

Epidemiology. According to conventional twinning rates, monozygotic twins represent approximately one-third of twins born in Caucasian populations and occur at a rate of 3–4 per 1000 births. The biological events responsible for monozygotic twinning are not well understood. It is generally agreed that monozygotic twinning occurs randomly and not as a genetically transmitted tendency. Some recent evidence from Sweden suggests an increased tendency for mothers who are monozygotic twins to bear same-sex twins themselves; further work will be needed to resolve this question.

Dizygotic twinning represents approximately two-thirds of twins born in Caucasian populations. The dizygotic twinning rate is lowest among Asian populations (2 per 1000 births), intermediate among Caucasian populations (8 per 1000 births), and highest among African populations (50 per 1000 births in parts of Nigeria). The natural twinning rate increases with maternal age, up to between 35 and 39 years, and then declines. A possible causal mechanism is the increased ovarian activity that continues until a woman reaches her late thirties. Elevated levels of follicle-stimulating hormone (FSH) may increase the probability of dizygotic twinning among some women. Dizygotic twinning has also been linked to increased parity, or the number of children to which a woman has previously given birth. However, parity is associated with older maternal age, which is more closely linked to dizygotic twinning. Mothers of dizygotic twins are also significantly taller and heavier, on average, than mothers of monozygotic twins and singletons. *See* OVARY.

Dizygotic twinning appears to be genetically influenced, although the pattern of transmission within families is unclear. One study found that parents of both monozygotic and dizygotic twins are twice as often left-handed as their own same-sex singleton siblings, but the specific mechanism linking twinning and left-handedness is unknown. Recent work in New Zealand uncovered a gene in sheep associated with dizygotic twins and triplets that could assist the understanding of the genetics of human multiple birth.

A higher frequency of dizygotic twinning among lower social classes than among higher social classes has been reported. This relationship may simply reflect the larger size of the lower-class families. A relationship may exist between dizygotic twinning and dietary factors, but it could be quite complex: A reduction in the dizygotic twinning rate has been associated with reduced nutritional supply in some European countries during World War II, although there have been mixed trends regarding nutrition and overall twinning rates. Some researchers have

implicated the yam (a food with estrogenlike substances) in the extremely high dizygotic twinning rate among the Yoruba of western Nigeria, although this is not definitive. Higher and more constant food supplies may explain the increased dizygotic twinning rate in southwest Finland's archipelago of Aland and Aboland, relative to the mainland.

Increased coital rate, following extended periods of sexual abstinence, may also increase the dizygotic twinning rate. The rate of dizygotic twinning is elevated in extramarital conceptions, during the first year of marriage, and in couples immediately after soldiers return from war. It may be that, in some women, coitus stimulates ovulation, and frequent ovulation leads to superfecundation. Increased dizygotic twinning has also been associated with an increased number of marriages.

The United States' twinning rate rose 65% between 1980 and 2002 and 38% between 1990 and 2002. It went up 3% between 2001 and 2002, to 31 twin births per 1,000 births. These increases were mainly due to advances in fertility treatments (for example, in vitro fertilization and ovulation induction), but were also due to delays in child-bearing. The increase mostly involved dizygotic twinning, in which multiple ovulation and maternal age are key factors. In vitro fertilization has, to a smaller extent, increased monozygotic twinning rates, possibly due to laboratory manipulation of the early embryo. However, in 2002 the birth rate of triplets and other higher order multiples dropped slightly to 184 per 100,000 births. This represented the third such decrease in the previous four years, following a 400% increase between 1980 and 1998.

The birth rate of preterm babies, defined as infants born after less than 37 weeks' gestation, increased 12.1% in 2002, and increased by 15% since 1990. This change partly reflects the increase in the rate of multiple births, which usually occur early.

Determination of zygosity. The accurate assignment of twin pairs as monozygotic or dizygotic is crucial for avoiding misleading estimates of genetic influence upon traits of interest. Co-twin comparison of selected DNA markers (six to eight) is the most accurate (greater than 99%) scientific method for classifying twins as monozygotic or dizygotic. Except for monozygotic twins, it would be extremely rare for family members to show matching DNA patterns. Cells for analysis can be obtained easily by gently scratching the inner cheek to obtain a buccal smear. Several laboratories are equipped to receive samples by mail and to provide twin-typing results in approximately 2 weeks. Blood-typing is another objective method for classifying twin pairs. Blood-group differences identify dizygotic co-twins with complete certainty. However, an absence of blood group differences can indicate only the probability that the twins are monozygotic, because in rare cases dizygotic co-twins may share all measured blood groups because of shared parentage. Analyses of anthropometric characteristics, such as fingerprint ridgecount (dermatoglyphic analysis), ponderal index, and cephalic index, improve the accuracy of twin classification when used in conjunction with blood-typing. *See* BLOOD GROUPS; DEOXYRIBONUCLEIC ACID (DNA); FINGERPRINT.

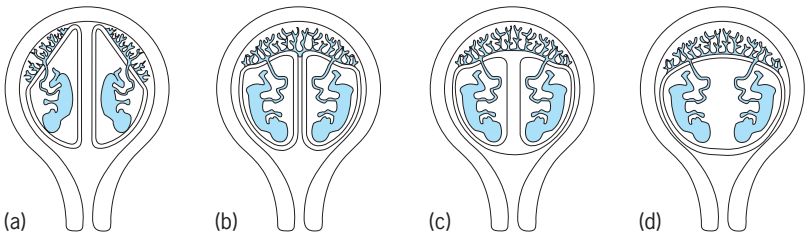


Fig. 1. Arrangements of placentae and placental membranes for monozygotic and dizygotic twin pairs in the uterus. (*a*) Monozygotic or dizygotic twins with separate amnions, chorions, and placentae. (*b*) Monozygotic or dizygotic twins with separate amnions and chorions and fused placentae. (*c*) Monozygotic twins with separate amnions and single chorion and placenta. (*d*) Monozygotic twins with single amnion, chorion, and placenta. Analysis of these arrangements can aid in diagnosing zygosity. (*After C. E. L. Potter, Fundamentals of Human Reproduction, McGraw-Hill, 1948*)

Other methods for diagnosing zygosity include physical-resemblance questionnaires and examination of placentae and placental membranes (**Fig. 1**). The more accurate of these methods is the physical resemblance questionnaire, which shows 93–99% agreement with the results from blood typing. Twins' and parents' judgments of twin type are often questionable.

Twin studies. Francis Galton reasoned late in the nineteenth century that comparative analyses of the behavioral and physical resemblance of the two types of twins would provide useful insights into the relative contributions of heredity and environment. (Although the biology of twinning was not established during Galton's time, he correctly recognized that there were two types of twins.) In particular, he suggested that greater resemblance within twin pairs derived from a single ovum (monozygotic twins), relative to twin pairs derived from separate ova (dizygotic twins), would demonstrate a contribution from genetic factors. This monozygotic–dizygotic twin comparison, called the classic twin method, assumes that differences between one-egg twins are generally due to environmental effects, given their shared heredity, and that differences between two-egg twins are due to genetic and environmental differences between them. The twin method additionally assumes that trait-relevant environmental influences are equal for both types of twins (equal environments assumption). Galton obtained detailed analyses of the life histories of twins, some of whom displayed early physical and behavioral similarities, and some an absence of early resemblance. He concluded that nature made a major contribution to many domains of human development.

Challenges to the classic twin method have been raised, despite its successful implementation in numerous behavioral science and medical science fields. Social biases refer to biases in treatment of monozygotic and dizygotic twins; for example, some critics of twin studies assert that there might be more similar treatment of monozygotic than dizygotic twins that may amplify behavioral similarity between the former and reinforce behavioral

dissimilarity between the latter. Empirical testing has, however, demonstrated that similarity of treatment is unrelated to similarity of behavior among monozygotic twins. Primary biases refer to unique prenatal influences, such as shared intrauterine circulation, which tend to reduce resemblance between monozygotic co-twins. These effects do not, however, appear to seriously affect results from twin data based upon large samples. Recruitment biases refer to differences in representation among the types of twins who volunteer for research: approximately two-thirds of adult same-sex twin volunteers are female, and two-thirds are monozygotic. Twin studies carried out in Norway and the United States (Minnesota), however, were unable to demonstrate differences in resemblance for intelligence quotient (IQ), personality traits and interests, or demographic variables between solicited pairs in which both co-twins completed questionnaires by mail, as compared with pairs in which one twin participated. Twins who volunteer for research without solicitation, however, may not be representative of the general twin population. Another concern is that some twin studies are based on small samples, limiting the power of their conclusions.

Intraclass correlation. The intraclass correlation (r_i) is the preferred measure of resemblance between co-twins. It is based on a ratio of the variation between groups to the total variation (group refers to a twin pair) and is given by the equation below, where

$$r_i = \frac{S^2b - S^2w}{S^2b + (n-1)S^2w}$$

S^2b = variance between families, S^2w = variance within families, and n = average number of individuals per family.

Heritability expresses the proportion of population variation associated with genetic differences in a given characteristic. It is a function of the particular population under study, the time of measurement, and the measuring instrument, and is therefore not a constant figure but one that reflects changes in these factors. Subtracting the dizygotic intraclass correlation from the monozygotic intraclass correlation estimates half the genetic effect because dizygotic twins share half as many genes, on average, as do monozygotic twins. Doubling the resulting value yields an estimate of the genetic effect. This procedure works only if the genetic effects are additive; that is, the effects of the genes are constant, regardless of the genes with which they are paired. However, sometimes genes behave interactively or nonadditively—that is, the effects of the genes change depending upon the genes with which they are paired. This may increase the monozygotic twin intraclass correlation and heritability estimates because monozygotic twins always share gene combinations while dizygotic twins do not.

Methods for analyzing twin data have moved beyond analyses of correlations to more complex model-fitting approaches. Researchers develop a model specifying genetic and environmental relatedness and apply it to data from twins and other relatives. They can then estimate genetic and environmental contributions to the trait, based on how well the model fits the data.

Twin and adoption studies. Twins reared apart offer informative tests of genetic and environmental influences on behavior. If monozygotic twins are separated early in infancy and raised in separate homes selected at random, their degree of similarity (intraclass correlation) provides a direct estimate of genetic effects. The study of dizygotic twins reared apart enables additional tests of the possibility of genetic interactions. Another approach is to study the adopted siblings of twins that are reared apart, a procedure that enables comparison between individuals who share environments but not genes; this is the reverse of studying separated monozygotic twins except that each twin and his/her adopted sibling are not the same age. This can be overcome by studying "virtual twins," genetically unrelated children of the same age raised together from infancy. They may consist of two adopted children or one adopted child and one near-in-age biological child in the family. They share their environments but not their genes so they truly are the reverse arrangement of monozygotic twins reared apart. One laboratory in the United States (at California State University, Fullerton) is studying these "twinlike" sibships to determine the extent to which common environments affect behavioral similarities.

Comparison of resemblance between monozygotic and dizygotic twins reared together and those reared apart also provides information on the influence of common rearing on behavioral and physical similarity. The level of personality resemblance is similar for monozygotic twins reared apart and reared together (approximately 50% of the variance is associated with genetic factors), demonstrating that common environmental influences do not contribute importantly to personality similarity. Resemblance in the body mass index (which is the weight in kilograms divided by the height, in squared meters) is only slightly lower for monozygotic twins reared apart than for monozygotic twins reared together. These analyses suggest that both genetic factors and environmental factors unique to the individual are important influences on these characteristics.

Twin gender studies. There has been recent interest in gender-related behaviors of opposite-sex twins. This is due to findings in the nonhuman literature showing that female mice and gerbils demonstrate masculine behaviors if situated next to a male fetus in utero. It is suggested that prenatal exposure to male hormones may be responsible for these results. Human twin research has produced mixed findings in this regard. It is also important to consider that female co-twins' behaviors could reflect the psychological and social effects of being raised with a same-age male sibling. However, the finding that females with male co-twins show the same frequency of otoacoustic emissions ("hums" discharged in the inner ear to raise the volume of weak sounds) as their brothers is difficult to explain without reference to opposite-sex

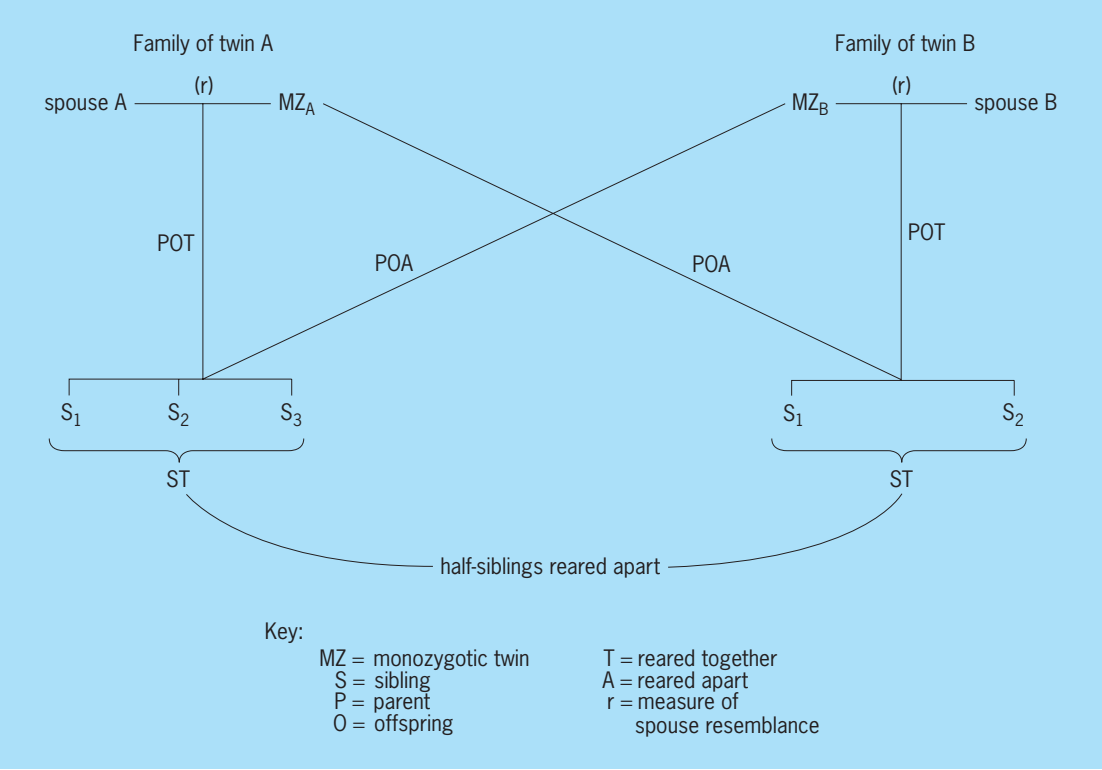


Fig. 2. Twin-family design. (*Adapted from S. Scarr-Salapatek, in K. W. Schaie, ed., Developmental Human Behavior Genetics*)

twins' prenatal biological circumstances.

Twin-family studies. The offspring of monozygotic twins share unique relationships with one another. The children of monozygotic twins are genetically equivalent to half-siblings, because they each have a parent that is genetically identical (either a twin mother or a twin father). Similarly, monozygotic twins share the same genetic relationship with the co-twin's children (nieces and nephews) as they do with their own children. This family constellation thus enables comparisons between co-twins, marital partners, twins and children, twins with nieces and nephews, full siblings, and half-siblings. Studies of mental abilities and physical characteristics that use the twin-family method are available. Studying social relationships among members of these unique families is also potentially informative with respect to genetic and social influences (**Fig. 2**).

Twins as couples. This approach to twin research compares monozygotic and dizygotic twins' social relationships with reference to their joint behaviors, such as cooperation and competition. The focus of interest is the behavior of the pair (for example, their level of cooperation or competition), not their separate behaviors as in classic twin studies. Research using this design has generally revealed greater cooperation between monozygotic twins than dizygotic twins in experimental games and tasks. These tendencies mirror the increased social closeness that monozygotic twins typically express toward one another and the increased grief they experience following the loss of their co-twin, relative to dizygotic twins. There is, however, overlap among twin types—some dizygotic twins are especially close to one another, while some monozygotic twins are less so.

QTL studies. Quantitative traits, such as height, intelligence, and certain aspects of personality, are influenced by many genes working together. People vary continuously on such traits, ranging from short to tall; average to highly intelligent; and outgoing to shy. Quantitative trait loci (QTLs) are the many genes that affect these traits, although they each differ in their effect. Identifying genes with the greatest influence on traits of interest can tell us about individual differences in the development of that trait. Now that the Human Genome Project has mapped the full complement of 20,000 to 25,000 genes, behavioral scientists and molecular geneticists have increased their efforts at finding QTLs for mental retardation, aggression, and psychopathology. It has how been shown that individuals with the apolipoprotein e4 gene are at increased risk for late-onset Alzheimer's disease. Twin studies will play important roles in the quest to find QTLs. The most valuable information may come from studying monozygotic twins who differ in their expression of disease and behavior. In the case of twins differing in schizophrenia, it may be that one twin was exposed to specific environmental factors that activate relevant genes or QTLs, while the other twin was not. Furthermore, even monozygotic twins are not genetically identical, as indicated above. Unraveling the link between these genetic differences and behavior will be an informative, albeit challenging task. *See* HUMAN GENOME PROJECT.

Twin registries. Twin studies depend on large samples in order to draw valid conclusions about behavior and development. Statistically significant findings are those occurring by chance fewer than one time in twenty. Statistical significance is associated with sample size and the size of the correlation. A .25 correlation would be statistically significant with 45 twin pairs, while a .50 correlation would be statistically significant with 12 twin pairs. Large twin samples are also required for estimating heritability, especially if the heritability of a given trait is low in the population. However, a large number of subjects does not ensure valid findings because the quality of the information may be difficult to determine. It has also been shown that heritability may be calculated more accurately and efficiently using twins reared apart rather than twins reared together. For example, 400–500 monozygotic and 400–500 dizygotic twin pairs reared together allow heritability to be estimated with the same degree of confidence as 50 monozygotic twin pairs reared apart. This is because heritability is estimated directly from the reared apart twins and indirectly from the reared together twins.

Modern twin research has come a long way in recruiting twins since Francis Galton did his famous study of nature and nurture in 1875. Galton gathered twins by sending notices to twins and relatives of twins. Respondents were asked to provide the names and addresses of other twins that they knew, thus enlarging his potential participant pool. Galton received 80 replies, although he did not indicate what percentage of his initial inquiries this number represented. Thirty-five twin pairs, presumably monozygotic, showed "close similarity" based on written responses to his questions. Galton clearly showed that twin studies could illuminate genetic and environmental influences on human behavior. However, findings from small samples gathered unsystematically urge cautious interpretation of the results. Fortunately, there is now a growing number of population-based twin registries.

The oldest twin registry in the world is the Danish Twin Registry, established in 1954. It now includes over 65,000 twin pairs, comprising 127 birth cohorts. A recent review summarized characteristics for 16 national twin registries, located in the United States, Australia, Belgium, Canada, China, Denmark, Finland, Germany, Italy, Japan, Korea, the Netherlands, Norway, Sri Lanka, Sweden, and the United Kingdom. There are also statewide twin rosters such as the Minnesota Twin Registry and the Wisconsin Twin Panel, and commercial registries such as the United Kingdom's *Gemini Genomics*, dedicated to finding disease-related genes. Investigators interested in specific conditions can try to identify specific subjects within these sources. There are, however, registries that target twins with particular conditions such as HIV exposure and chronic fatigue syndrome. Some registries include specific subgroups of twins, such as World War II veterans, Vietnam veterans, and elderly citizens. Aside from offering researchers access to large and diverse twin samples, the registries facilitate international collaboration. Studies of rare medical conditions can proceed by pooling cases across sites. Attempts at cross-cultural replication of findings can also be conducted.

Value of twin studies. Studies of monozygotic and dizygotic twins provide valuable information on genetic and environmental influences on human behavioral and physical development. Studies of twins, therefore, offer many opportunities for examining novel hypotheses and generating new explanations of observations. For example, twin studies have been used to assess predictions generated by human evolutionary biology and theories of economics. New model-fitting approaches to data analysis enable testing for the presence of specific genetic and environmental effects. Refinement of twin samples and the study of more well-defined characteristics is another important trend. In recent years, twin research has begun to address the origins of less frequently studied behaviors such as humor, love styles, athleticism, and religiosity. Continued extension and elaboration of the twin design, replication of research, and continued testing of assumptions will help researchers to improve twin methods and respond to criticism. *See* BEHAVIOR GENETICS; HUMAN GENETICS.

Nancy L. Segal

Bibliography. M. G. Bulmer, *The Biology of Twinning in Man*, Clarendon Press, Oxford, 1970; A. Busjahn (ed.), Special issue on Twin Registries across the globe: What's out there in 2002, *Twin Res.*, no. 5, 2002; F. Galton, The history of twins as a criterion of the relative powers of nature and nurture, *J. Anthropol. Inst.*, 6:391–406, 1875; International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, *Nature*, 43:931–945, 2004; K. K. Kirk and N. G. Martin (eds.), Special issue on religion, values and health, *Twin Res.*, no. 5, 1999; D. T. Lykken, S. Geisser, and A. Tellegen, Heritability estimates from twin studies: The efficiency of the MZA design (unpublished manuscript); G. A. Machin et al., Correlations of placental vascular anatomy and clinical outcomes in 69 monochorionic twin pregnancies, *Amer. J. Med. Gen.*, 61:229–236, 1996; P. McGuffin, B. Riley, and R. Plomin, Genomics and behavior: Toward behavioral genomics, *Science*, 291:1232–1249, 2001; National Center for Health Statistics, News Release, Dec. 17, 2003; N. L. Segal, *Entwined Lives: Twins and What They Tell Us About Human Behavior*, Plume, New York, 2000; N. L. Segal, Human cloning: A twin-research perspective. *Hastings Law J.*, 53:1073–1084, 2002; N. L. Segal, Virtual twins: New findings on within-family environmental influences on intelligence, *J. Educ. Psych.*, 92:442–448, 2000.

Two-phase flow

The simultaneous flow of two phases or two immiscible liquids within common boundaries. Two-phase flow has a wide range of engineering applications

such as in the power generation, chemical, and oil industries. Flows of this type are important for the design of steam generators (steam–water flow), internal combustion engines, jet engines, condensers, cooling towers, extraction and distillation processes, refrigeration systems, and pipelines for transport of gas and oil mixtures.

The most important characteristic of two-phase flow is the existence of interfaces, which separate the phases, and the associated discontinuities in the properties across the phase interfaces. Because of the deformable nature of gas-liquid and liquid-liquid interfaces, a considerable number of interface configurations are possible. Consequently, the various heat and mass transfers that occur between a two-phase mixture and a surrounding surface, as well as between the two phases, depend strongly on the two-phase flow regimes. Multiphase flow, when the flow under consideration contains more than two separate phases, is a natural extension of these principles. *See* FLUID MECHANICS; INTERFACE OF PHASES.

From a fundamental point of view, two-phase flow may be classified according to the phases involved as (1) gas-solid mixture, (2) gas-liquid mixture, (3) liquid-solid mixture, and (4) two-immiscible-liquids mixture. *See* GAS; LIQUID; PHASE EQUILIBRIUM.

Flow regimes in vertical pipes. For up-flow in vertical pipes, several different flow regimes can be observed for gas-liquid (or two-immiscible-liquids) mixtures.

Bubbly flow. Bubbly flow is the most widely known flow pattern and is recognizable from everyday experience. Here the gas phase is distributed in discrete bubbles within a liquid continuum.

Slug flow. As the bubble concentration becomes very high in a bubbly flow, coalescence may produce much larger bubbles. In this type of flow, some of the gas bubbles have nearly the same cross section as that of the channel, and their length may be many times the diameter. The gas bubbles or plugs have spherical caps and flat tails and move along like bullet-shaped bubbles (**Fig. 1**). The bubble is separated from the pipe wall by a descending liquid film.

Churn flow. If the velocity of a two-phase mixture flowing as a slug flow is increased, the flow structure will eventually become unstable and the gas plugs will begin to break up, leading to an unstable regime in which the gas flows in a chaotic manner through the liquid (Fig. 1). The liquid is mainly displaced to the channel walls and flows in oscillatory motion, upward and downward.

Annular flow. In this regime, the gas flows mainly in a central core, with the liquid flow almost completely confined to a liquid layer or film on the channel walls. This presents a more or less continuous interface to a stream consisting mainly of gas flowing in the center of the channel.

Drop flow. The drop flow regime (also called mist flow or liquid dispersed flow) is one in which there is complete dispersion of the liquid in a continuum gas flow. This is usually found within flow in heated channels which, incidentally, has been found to contain all the previous mentioned flow patterns developed in the order described here.

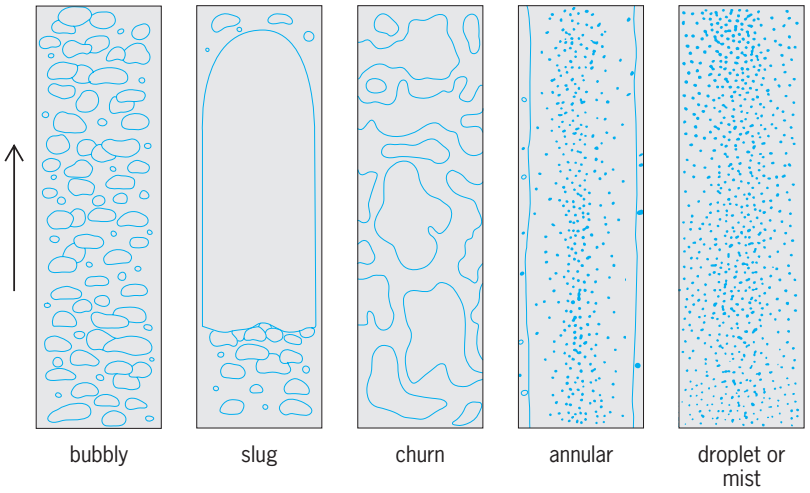


Fig. 1. Flow patterns of the gas-liquid flow in a vertical pipe.

Flow regimes in horizontal pipes. For flow in horizontal (and inclined) pipes, the flow patterns are more complex because gravity causes an asymmetric distribution of the phases (**Fig. 2**).

Bubbly flow. This is defined as in vertical flow, but there is a tendency for the bubbles to flow in the upper part of the channel (Fig. 2). At higher flow rates, the whole cross section may contain bubbles, but there will be a higher concentration toward the upper part of the duct.

Plug flow. The characteristic bullet-shaped bubbles occur as in vertical slug flow. However, in this case the plugs do not occupy as large a proportion of the cross section, and the liquid layer, separating the gas bubble from the wall, tends to be thicker at the bottom of the channel than at the top. In addition, the nose of the bubble is asymmetric (Fig. 2).

Stratified flow. At low velocities, the two fluids may flow in complete separation, with a flat interface. The liquid flows at the bottom of the channel and the gas at the top.

Wavy flow. As the gas velocity is increased in stratified flow, surface waves begin to build up on the liquid layer. This region is called wavy flow.

Slug flow. As the gas velocity is further increased in the wavy flow region, the waves grow in size, eventually reaching the top of the channel. These are propagated as slugs at high velocity and wet the whole of the channel surface, leaving a liquid film covering the surface in between the bridging waves.

Annular flow. At very high gas flow rates, the slugs become pierced, creating a gas core, and the flow becomes essentially annular. As in vertical flows, there may be some droplet entrainment and a dispersion of bubbles may exist in the liquid film. The film is very much thicker at the bottom of the channel than at the top.

Droplet flow. This is much the same as for vertical droplet flow, except the highest concentration of

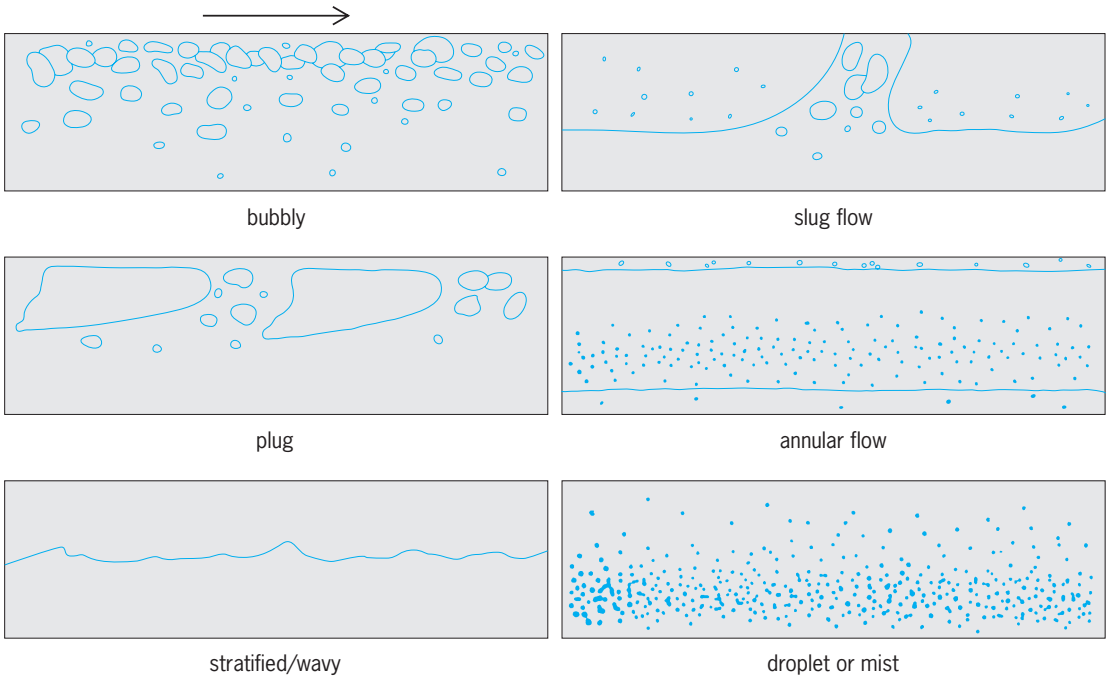


Fig. 2. Flow patterns of gas-liquid flow in a horizontal pipe.

droplets is below the pipe centerline.

Two-phase flow measurements. The primary condition for all two-phase flow is specified by the volumetric quality, which is defined as the ratio of the dispersed-phase flow rate to the total flow rate. Other required fundamental measurements include the average volumetric quality (obtained using methods such as quick-closing valves, and neutron, x-ray, and gamma-ray techniques) and local properties such as the void fraction or volume fraction, continuous- and dispersed-phase velocities, turbulent intensity, and bubble (drop) size. Dual optical and dual resistivity probes have been used for void fraction, bubble velocity, and bubble diameter measurements. The local void fraction (gas-liquid) or volume fraction (liquid-liquid) is defined (using a time-history record from a local probe) as the ratio of the time of the dispersed phase to the total measuring time. Continuous-phase velocity and turbulent intensity data have been obtained with hot-film anemometry. *See* FLOW MEASUREMENT.

Two-phase flow applications. Industrial applications of two-phase flow include systems that convert between phases, and systems that separate or mix phases without converting them (adiabatic systems).

Phase-change systems. Many of the practical cycles used to convert heat to work use a working fluid. In two or more of the components of these cycles, heat is either added to or removed from the working fluid, which may be accompanied by a phase-change process. Examples of these applications include steam generators, evaporators, and condensers, air conditioning and refrigeration systems, and steam power plants.

The majority of electricity-generating plants are variations of steam power plants. The function of the boiler in a fossil-fuel power plant is to supply the energy needed to vaporize the water, which passes through the tubes in the boiler. In nuclear power plants, the energy is released by a controlled nuclear reaction taking place in a separate reactor building. This energy is initially transferred to either pressurized water or to a liquid metal reactor coolant, with a subsequent energy transfer to the water circuit in secondary heat exchangers. Solar power plants have receivers for concentrating and collecting solar radiation in order to vaporize the working fluid. Regardless of the energy source, the vapor produced in the boiler expands, thus causing the rotation of the turbine shaft, which is connected to an electrical generator. The lower-pressure vapor leaving the turbine passes through a heat exchanger called a condenser. There it condenses on the outside of tubes carrying cooling water. The cooling water is pumped to a cooling tower, where the energy taken up in the condenser is released into the atmosphere by evaporation. *See* SOLAR ENERGY; STEAM ELECTRIC GENERATOR.

Adiabatic systems. In these types of systems, the process of phase mixing or separation occurs without heat transfer or phase change. Examples of these systems are airlift pumps, pipeline transport of gas-oil mixtures, and gas-pulverization of solid particles. *See* ADIABATIC PROCESS.

There are many different designs for phase separation that use the principles of gravity and centrifugal force. **Figure 3** shows a separating hydrocyclone in which the effect of centrifugal forces is used for solid-fluid separation or fluid-fluid separation. Hydrocyclones do not have any rotating parts, and the necessary vortex action is produced by pumping the

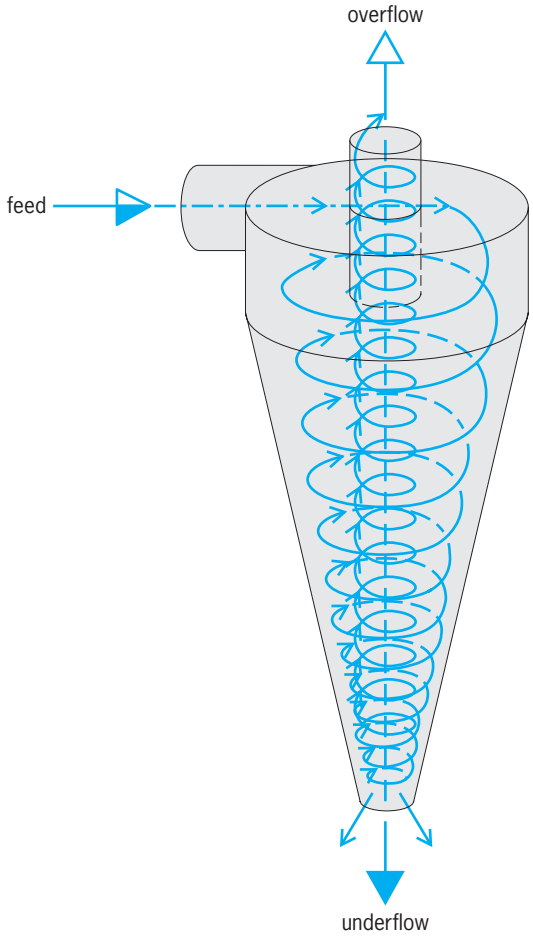


Fig. 3. Hydrocyclone phase-separation system.

fluid tangentially into a cylindrical/conical body. The end-plate, which covers the top of the cylindrical section, contains the liquid overflow pipe, which usually protrudes some distance into the cyclone body. The underflow, which carries the higher-density fluid or solids, leaves through the opening in the apex of the cone. H. H. Bruun; F. Hamad; B. K. Pierscionek

Bibliography. H. H. Bruun, *Hot-wire Anemometry: Principle and Signal Analysis*, Oxford University Press, 1995; M. J. Moran and H. N. Shapiro, *Fundamental of Engineering Thermodynamics*, John Wiley, New York, 1995; M. C. Roco, *Particulate Two-Phase Flow*, Butterworth-Heinemann, Boston, 1993; L. R. Weatherly, *Engineering Processes for Bioseparations*, Butterworth-Heinemann, Boston, 1994.

Tylenchida

An order of nematodes in which the labial region is variable and may be distinctly set off or smoothly rounded and well developed; the hexaradiate symmetry is most often retained or discernible. The amphids are porelike, oval, or slitlike, or clefts located on the lips. The internal circlet of six sensilla may be lacking. The external circle of 10 sensilla is often evident; however, these may be reduced to a visible 4 or some may be doubled. The hollow stylet is the product of the cheilostome (conus, "guiding apparatus," and framework) and the esophastome (shaft and knobs). Throughout the order and its suborders the stylet may be present or absent and may or may not be adorned with knobs. The variable esophagus is most often divisible into the corpus, isthmus, and glandular posterior bulb. The corpus is divisible into the procorpus and metacorpus. The metacorpus is generally valved but may not occur in some females and males, and the absence is characteristic of some taxa. The orifice of the dorsal esophageal gland opens either into the anterior procorpus or just anterior to the metacorporal valve. The excretory system is asymmetrical, and there is but one longitudinal collecting tubule. Females have one or two genital branches; when only one branch is present, it is anteriorly directed. Except for sex-reversed males there is only one genital branch. Males may have one (=phasmid) or more caudal papillae. The spicules are always paired and variable in shape; they may or may not be accompanied by a gubernaculum.

The order comprises five superfamilies: Tylenchoidea, Criconematoidea, Aphelenchoidea, Aphelenchoidoidea, and Sphaerularoidea.

Tylenchoidea. The superfamily Tylenchoidea comprises parasites of fungi, plants, and some insects. The most destructive and widespread endoparasitic plant nematodes are found in the families Pratylenchidae and Heteroderidae. The "meadow" or "lesion" nematode *Pratylenchus* is an obligate migratory endoparasite whose feeding results in cell death. In order to maintain themselves these nematodes must continuously seek out living cells in all directions, forming an ever-growing lesion. Nematode lesions can eventually girdle and kill roots.

Among heteroderids, species of interest include the cyst nematodes and root-knot nematodes. These two are very similar in biology but differ slightly in their life cycles and root response. They are extremely important to agriculture throughout the world and affect more than 2000 plant species.

Criconematoidea. This is a superfamily of ectoparasitic nematodes of plants ranging from migratory to highly specialized sessile forms that emulate endoparasites. Sedentary root parasitism is accomplished in several ways. In some species the spear or feeding stylet is nearly one-half the length of the body. This feature allows the nematode to stay in one place and feed on several cells by probing deeply and widely. Other species induce terminal root galls that are attractive to other individuals of the same species. The most sophisticated sessile forms embed the anterior portion of the body into the root; their feeding induces the production of giant nutritive plant cells. Two important species are *Criconemoides xenoplax* and *Tylenchulus semipenetrans*. The former, through excessive root damage, predisposes peach trees to bacterial canker, an aerial disease caused by *Pseudomonas syringae. Tylenchulus semipenetrans* is commonly known as citrus nematode and is the most important citrus parasite.

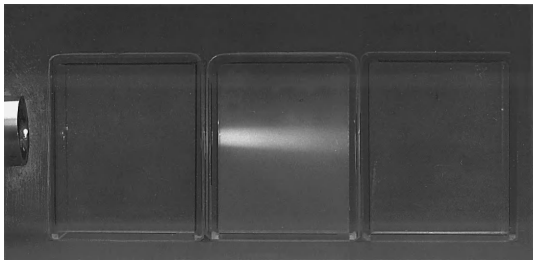
Aphelenchoidoidea and Aphelenchoidea. Aphelenchoidoidea includes parasitic forms, while Aphelenchoidea comprises mycetophagous soil inhabitants. As plant parasites, the Aphelenchoidoidea most often attack the aboveground parts; a few utilize insects as phoretic hosts to transport them from plant to plant. There are several important species of plant parasites that attack the aboveground parts, such as buds and leaves. Some important crops include strawberries, lilies, rice, chrysanthemums, and currants. An extremely important species is *Bursaphelenchus xylophilus*, on Japanese red and black pine. Seedlings can be killed by this nematode in 40 days. The nematode is transported from tree to tree by longhorn beetles in the family Cerambycidae. A single beetle, on the average, transports some 15,000 infective larvae. Another important disease is red ring of coconut, caused by *Rhadinaphelenchus cocophilus*. In this instance, the nematode is transported from tree to tree in the hemocoel of the weevil *Rhynchophorus palmarum*. Nematodes are deposited on palms when the weevil attempts to oviposit. *See* NEMATA; PLANT PATHOLOGY.

Sphaerularoidea. This is a superfamily of parasitic nematodes. Adult females are hemocoel parasites of insects and mites; a few taxa contain both plant and insect parasites. In general, nematodes belonging to this group have three distinct phases in their life cycles: two free-living and one parasitic. In the free-living phase the female gonad is single, anteriorly directed, and with few developing oocytes and a prominent uterus filled with sperm. In the parasitic phase either the body becomes grossly enlarged and degenerates to a reproductive sac, or the uterus prolapses and gonadal development takes place outside the body. The males are always free-living and not infective. The most interesting taxon is *Sphaerularia bombi*, which prolapses the uterus. When totally prolapsed, the gonad becomes the parasite and the original female a useless appendage. It is not unusual for the prolapsed gonad to attain a volume 15,000 times that of the original female. Armand R. Maggenti

Tyndall effect

Visible scattering of light along the path of a beam of light as it passes through a system containing discontinuities. The luminous path of the beam of light is called a Tyndall cone. An example is shown in the **illustration**. In colloidal systems the brilliance of the Tyndall cone is directly dependent on the magnitude of the difference in refractive index between the particle and the medium. In aqueous gold sols, where the difference in refractive index is high, strong Tyndall cones are observed.

For systems of particles with diameters less than one-twentieth the wavelength of light, the light scattered from a polychromatic beam is predominantly blue in color and is polarized to a degree which depends on the angle between the observer and the incident beam. The blue color of tobacco smoke is an example of Tyndall blue. As particles are increased in size, the blue color of scattered light disappears and the scattered radiation appears white. If this scattered light is received through a nicol prism which is oriented to extinguish the vertically polarized scattered light, the blue color appears again in increased brilliance. This is called residual blue, and its intensity varies as the inverse eighth power of the wavelength. *See* COLLOID; SCATTERING OF ELECTROMAGNETIC RADIATION. Quentin Van Winkle



Luminous light path known as the Tyndall cone or Tyndall effect. (*Courtesy of H. Steeves and R. G. Babcock*)

Type (printing)

The intelligible images organized into readable text of various styles and sizes. Type used in printing on paper or in video display is divided into four categories: foundry, machine-cast, photocomposed, and digitized type. In the first two the face of the letter is raised on one end of a piece of metal. It is from that surface, when inked, that the impression of type was made from its invention until the 1970s. In photocomposition, the type is reproduced photographically. Digitized methods assemble dots into typographic letters, lines, or pages.

Classification. Foundry type, also known as hand type, is cast as single characters in much the same way that Johannes Gutenberg, the inventor of movable type, produced them in Mainz, Germany, about 1440.

Machine-cast metal type is produced by Linotype, Intertype, Ludlow, and Monotype machines. Very few of these machines are still in operation. All but the last cast type in lines, or slugs. The Monotype—in reality two devices, a keyboard and a caster—produces individual types that are then manually set in lines of desired lengths. The Linotype, the earliest of the mechanical composing machines, was introduced in 1886; it was invented by Ottmar Mergenthaler.

The output of the photographic type machine is the image of type on film or on photosensitized paper in negative or positive form.

The term "cold type" is applied to text matter produced on a typewriter or laser printer and to words or lines made up of individual printed characters assembled or pasted together for photographic reproduction. Digital type uses a number of printout technologies, imaging small dots into lines (called rasters) and positioning them on pages from patterns of zeros and ones called bitmaps. Because this method is electronic, more sizes and variations (for

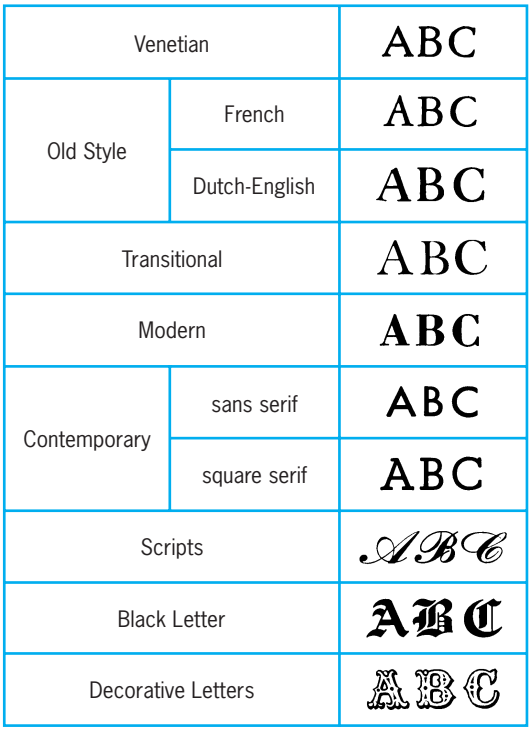


Fig. 1. Eight type classifications.

example, condensing or expanding or other distortion) are possible.

Of the more than 50,000 type styles created since Gutenberg's time, about 6000 are in everyday use throughout the world. The most widely used method for classifying them is the serif-evolution system, based on the different shapes of the terminals or endings of letters. This provides eight classifications: Venetian, Old Style (Dutch-English and French), Transitional, Modern, Contemporary (sans serif and square serif), Scripts, Black Letter, and Decorative Letters (**Fig. 1**).

Type measurements. Type is set in sizes from 4 point to 144 point. The sizes of type in most common use are 6, 7, 8, 9, 10, 11, 12, 14, 16, 18, 24, 30, 36, 42, 48, 60, and 72 point (**Fig. 2**). Modern sizes can range from 2 point to 500 point in increments up to three decimals (that is, 9.363 point), depending on the software.

The American point system, adopted in 1886, made the unit of type measurement a point, 0.01383 in. (0.35128 mm) or nearly 1/72 in., and all type sizes are multiples of this unit. This system replaced the sixteenth-century practice of giving all type sizes names such as pearl, agate, nonpareil, brevier, long primer, and pica. Some names remained and have been assigned other functions. Agate, $5^1/_2$-point type or 14 lines of type to an inch, has come to be used for measuring newspaper advertising space. Publications quote small space rates by the agate line. Nonpareil is used to designate 6 points of space in and around type. Pica is commonly used to denote a unit of space measuring 12 points. It is applied as a dimensional unit to the length of type lines and to the width and depth of page. With the advent of desktop publishing, many software programs have standardized on 72 points equaling exactly 1 inch. With these programs, the point now equals 0.01389 in., compared with the traditional point, 0.01383 in. Some of these programs allow the user to select which point system to use.

In continental Europe the unit of type measurement is the didot point, which equals 0.0148 in.; in the United States the point is 0.0138 in. In the didot system, the horizontal unit of measurement is the cicero, which equals 12 didot points, or 0.178 in. (4.51 mm). With the increasing internationalization of design measurement systems due to the use of computers, the didot point now is rarely used.

The standard height of metal type in the United States is 0.918 in. (2.3 cm), a dimension called type-high and one observed by photoengravers and electrotypers who make plates to be combined with cast type. Letterpress printing presses are adjusted to fit this standard height.

The hot composition type methods used metal composed of an alloy of lead, tin, and antimony. Large sizes, 72–144 point or more, were sometimes made of wood. Type size is the distance measured in points from above the ascending or capital letters to below the descending letters. In the days of metal setting, the point size of type was dependent on the point size of the body that the metal letter was cast onto. Thus, there is often a variance in the actual letter size of different fonts of the same point size, steming from the design and casting of letters in metal. Thus, the exact measurement from the top to the baseline of one 8-point A may not be the same as another 8-point A. The x-height is the height of the lowercase x, which is the typical height of the "readable" parts of the lowercase letters. It is the best determinant of type size.

Fonts. A complete complement of letters of one size and style from A to Z, together with the arabic numerals, punctuation, footnote reference marks, and symbols, is called a font (**Fig. 3**). Most fonts also include the ampersand and currency symbols and the ligatures, ff, fe, ffi, fl, ffl, æ, and œ. A letter from a different font that is included by accident in type composition is known as a wrong font. All typographic images, from letters to symbols to oriental ideographs, are called glyphs.

Type families. A family of type may be likened to the shades of a color in that it includes variations of a given type face or design. Weights are varied, from light to medium, bold, and extra bold; letters are condensed and expanded, as well as outlined, inlined, and shadowed (**Fig. 4**).

Space inserted between the lines of type to open them up vertically is called leading. Space inserted between letters of a word to open it up horizontally is known as letter spacing. Modifying space between two letters is called kerning. Modifying space between a series of letters is called tracking.

Special designs. Ornaments and rules (lines) can also be cast on machines or typeset digitally. These include individual designs, known as printer's flowers, fleurons, or florets; braces and brackets used to

Fig. 2. Sizes of type in most common use.

enclose or connect lines of type; chemical, mathematical, astronomical, and medical signs; and signs of the zodiac. Special symbols are called pi symbols or dingbats.

The arabic numerals introduced into Europe in the twelfth century are available in many fonts as old style and modern (or lining). They are part of the fonts of typefaces derived from designs of the period 1400–1800.

Each age since the Italian Renaissance has produced memorable type designers. Among them are Nicolas Jenson, Aldus Manutius (first to use italics and small caps), Claude Garamond, William Caslon, John Baskerville, Giambattista Bodoni, Frederic W. Goudy, W. A. Dwiggins, Rudolf Koch, Paul Renner, Jan van Krimpen, Stanley Morison, Eric Gill, Bruce Rogers, Matthew Carter, Hermann Zapf, and Erik Spiekermann.

Those who stress type in their design of books, publications, and advertising are known variously as typographers, typographic designers, or type directors. The size of type you are now reading is 9 point and is set $16\frac{1}{2}$ picas wide. *See* PRINTING.

Type for desktop publishing. Most typesetting today is performed using desktop publishing (DTP), which refers to the creation of documents through the use of a personal computer (small enough to fit on a desktop), off-the-shelf software that uses the PostScript® page description language, and laser-based output devices (consisting of a raster image processor and a marking engine) that convert the digital information into images on paper and other surfaces, such

Fig. 3. A font showing roman capitals, roman lowercase, small capitals, italic capitals, and italic style.

Fig. 4. A family of type.

as printing plates for offset lithography.

The type used for desktop publishing is also digital. Most of the fonts used today are Type 1, TrueType®, or OpenType® technology.

Type 1 fonts. The font technology known as Type 1 was developed by Adobe® as part of the PostScript language, and is the most widely used and best-supported method for imaging text on high-resolution output devices such as imagesetters and platesetters. Type 1 fonts consist of two components: the screen font and the printer font. Type 1 printer fonts (also called outline fonts) contain detailed information about the character's shape, expressed as mathematical formulas called vectors. This outline information consists of the location of the anchor points (which designate the corners and other important spots on the outline of the letters) and the shape of the lines drawn between the anchor points. These data are needed by the raster image processor (RIP), where it is used to create a high-quality representation of the typeface on the output device. The printer font's outline can be sent to the RIP as part of a PostScript file, or it can be downloaded (copied through the network) as needed from the desktop publishing workstation, or it can be "resident" in the RIP's font cache (a special partition of memory used to store font outlines).

The other component of a Type 1 font is called the screen font because it was developed to provide an on-screen representation of the font's shape at the desktop publishing workstation. Since all information shown on a computer monitor must somehow be converted to pixels (picture elements) in order to be displayed, the screen font consists of a pixel-by-pixel representation of each character at the low resolution used by the monitor (72 dots per inch).

In the early days of desktop publishing, screen fonts were supplied in only a limited number of common sizes (typically 6, 8, 10, 12, 14, 16, and 24 points). Any attempt to use the font at a size that did not have a matching screen font resulted in a poor representation of the font on the screen. Since each letter was captured as a digital image, any manipulation of text outlines (such as changes in scaling or kerning) resulted in jagged type. This common scenario came to be known as "bitmapped" text, since the poor on-screen representation made it obvious that the screen font was a bitmap made of pixels instead of a smooth vector outline. This problem was largely eliminated in 1989 by the introduction of a software program called Adobe Type Manager® (ATM) that performed on-the-fly rasterization of the printer font's vector equations into bitmaps for on-screen display.

TrueType fonts. TrueType is a more recent development in digital font technology. Unlike the Type 1 fonts, TrueType fonts have only one component. The same file provides information for the computer display and the output device. This is possible because capabilities very similar to ATM (the ability to rasterize outline fonts for the monitor on-the-fly) are built into both the Macintosh and the Microsoft Windows® operating system. TrueType also has improved facilities for controlling the "hinting" of its characters when displayed at low resolution. This is very important for improving the readability of small text on low-resolution devices such as computer monitors (72 dpi) and office laser printers (300–600 dpi).

Another technical distinction can be seen in the formulas used to describe font outlines: TrueType fonts tend to have more anchor points than similar Type 1 fonts due to the method used for describing curved lines within a PostScript font (known as Bézier). TrueType fonts use less complex (but easier to render) quadratic calculations; this reduced complexity requires more points to express elaborate curves. Conversion from quadratic to Bézier (or vice versa) is not lossless, so the transformation of fonts from one technology to another may lead to very slight changes in the shape of characters.

OpenType fonts. The OpenType format was jointly developed by Microsoft and Adobe Systems and is based on both TrueType and Type 1 font formats. It is a cross-platform font format that can be used on both Macintosh and Windows systems and has an expanded character set based on the international Unicode encoding standard. As a result, OpenType supports more languages, has greater typographic capabilities, and simplifies font management requirements. In addition, OpenType is compatible and can be used in the same document with Type 1 and TrueType fonts.

Eugene M. Ettenberg; Frank J. Romano; Thomas Destree; Roger Walton

Bibliography. M. Annenberg and S. O. Saxe, *Type Foundaries of America and Their Catalogs*, 1994; A. Faiola, *Typography Primer*, 2000; R. Goldberg and F. Romano, *Digital Typography*, 2000; A. Haley, *Hot Type Designers Make Cool Fonts*, 1998; R. Mclean, *Typographers on Type: An Illustrated Anthology from William Morris to the Present Day*, 1995; A. W. White, *Type in Use: Effective Typography for Electronic Publishing*, 1999.

Type method

The nomenclatural method for providing a fixed reference point for a taxon (plural, taxa), a group of organisms. When Linnaeus, the eighteenth-century

Swedish naturalist, ushered in the modern period of systematic biology through the publication of his classical work, *Systema Naturae*, he established the basis of binomial nomenclature. By means of this method the systematist who recognizes new, or supposedly new, taxa makes these known by means of a technical description and a scientific name. If the new taxon is a species, the material (specimens or parts of specimens) which the author had before him when he described it, together with such additional material as he may have gathered later, was considered typical, and such specimens were known as types. The types represented the author's notion of the new species and served as the basis of comparison. As knowledge increased, it was realized that each taxonomic name must be tied to a single specimen—the "type"—which is a nomenclatural, not a biological, type. Having a fixed reference for each name, later workers could correct taxonomic errors and modify the limits of species taxa without subsequent confusion in scientific names.

Type designations. For the first 50 or 100 years after Linnaeus, taxonomists did not make formal type designations. In fact, James E. Smith replaced some of Linnaeus's original specimens with better ones when they became available. Later taxonomists generally designated one specimen to serve as the type (or holotype) for newly described species. For existing species with competing syntypes or cotypes, one was selected as the type, usually called the lectotype. As the number of collections and type repositories increased and became less accessible to the taxonomist, as types became lost or destroyed, and as complications arose in trying to apply the single type concept retroactively, more refined type terminology and methods of type designation were developed. Although terms are available for more than 100 kinds of "type" specimens, these fall in two major classes: (1) nomenclatural types, the name-bearing specimens (onomatophores), which serve as reference points for nomenclatural purposes, and (2) taxonomic types which serve as standards of description and reference points for authors' concepts. The first category has been sanctioned by the International Commission on Zoological Nomenclature, the second category by taxonomic practice.

Among the more commonly used terms to designate these two kinds of types, nomenclatural and taxonomic, are the following:

1. *Nomenclatural types*

a. Holotype: the single specimen designated or indicated as "the type" by the original author at the time of publication of the original description or the only specimen known at the time of the original description.

b. Syntype (also called cotype): every specimen on which an author bases an original description when no single specimen has been designated as the holotype.

c. Lectotype: one of a series of syntypes which is selected subsequent to the original description and thenceforth serves as the definitive type of the species. To be effective, such selection must be made known through publication.

d. Neotype: a specimen selected as type subsequent to the original description when the primary types are definitely known to be destroyed. Here again selections must be made known through publication.

2. *Taxonomic types*

a. Paratype: a specimen other than the holotype which is before the author at the time of original description and which is designated as such or is clearly indicated as being one of the specimens upon which the original description was based.

b. Allotype: a paratype of the opposite sex to the holotype which is designated or indicated as such.

c. Topotype: a specimen not of the original type series collected at the type locality.

d. Plesiotype: a specimen or specimens on which subsequent descriptions or figures are based.

e. Metatype: a specimen compared with the type by the author of the species and determined by the author as conspecific with it.

f. Homotype: a specimen compared with the type by another than the author of a species and determined by the former to be conspecific with it.

Stabilization of names. Just as types are required as nomenclatural reference points for the application of scientific names to species, so they are also necessary for the stabilization of names for higher taxa. This is especially true at the level of the genus, not only because there are more generic names than those of any other higher category but also because generic names provide roots for the names of several taxa immediately above them, as tribe, family, and so forth. The type for a name of a higher taxon is a lower taxon; for example, the type of a genus is a species. As with species names, the necessity for refining early generic concepts and the retroactive application of the generic type concept to the works of previous authors have resulted in the development of a rather precise series of rules and practices governing their selection. These are specified in detail in the International Rules of Zoological Nomenclature and are concerned with two principal situations: (1) cases in which the generic type is to be accepted automatically based upon the method or form utilized by the author in the original publication in which the new name was proposed (such as type by original designation, type by monotypy, type by tautonomy), and (2) cases in which a subsequent author has a certain amount of freedom in making a selection from among two or more eligible names for type species (type by subsequent designation). As with other aspects of zoological nomenclature, priority of designation, if properly carried out, is binding upon later authors, except as the International Commission of Zoological Nomenclature, in the interest of conserving current usage, may set aside the rules in a given case. The Commission is also called upon to arbitrate difficult cases of generic type interpretation, as when the author misidentifies the species designated as the type of a genus and as a result, the generic

diagnosis applies to a species quite different from the one named in his publication. In such cases, the decision is usually influenced by whether the strict application of the author's concept or of a concept based upon the actual species name will have the most adverse nomenclatural effect. *See* ZOOLOGICAL NOMENCLATURE.

Walter J. Bock

Bibliography. R. E. Blackwelder, *Taxonomy: A Text and Reference Book*, 1967; D. Hegberg, *Systematics*, 1977; International Trust for Zoology Nomenclature, *International Code of Zoological Nomenclature*, 1964; E. Mayr, *Principles of Systematic Zoology*, 2d ed., 1991; G. G. Simpson, *Principles of Animal Taxonomy*, 1990.

Typewriter

A machine that produces printed copy, character by character, as it is operated. Its essential parts are a keyboard, a set of raised characters or a thermal print head, an inked ribbon, a platen for holding paper, and a mechanism for advancing the position at which successive characters are printed. The QWERTY keyboard (named for the sequence of letters of the top row of the alphabet worked by the left hand) was designed in the 1870s. It contains a complete alphabet, along with numbers and the symbols commonly used in various languages and technical disciplines. The manual typewriter was introduced in 1874, followed by the electrically powered typewriter in 1934. By the late 1970s, electronic typewriters offered memory capability, additional automatic functions, and greater convenience. Further advances in electronic technology led to additional capabilities, including plug-in memory and function diskettes and cartridges, visual displays, nonimpact printing, and communications adapters. Although many typewriters are still in use, computers and word processing software largely have supplanted them. *See* WORD PROCESSING.

Edward W. Gore, Jr.; Robert A. Rahenkamp