McGRAW-HILL
ENCYCLOPEDIA OF
SCIENCE &
TECHNOLOGY

15 R - RYE

Rabies

An acute, encephalitic viral infection. In humans it is almost invariably fatal. Human beings are infected from the bite of a rabid animal, usually a dog. *See* ANIMAL VIRUS.

Infectious agent. The virus is bullet-shaped with a cylinder diameter of about 10 nanometers and a length of about 210 nm. The nucleic acid is ribonucleic acid (RNA). The virion is enveloped, covered with projections, and ether-sensitive. A grouping of rabies virus, together with vesicular stomatitis virus of cattle and certain viruses of fish, insects, and plants, has been included under the name rhabdovirus group. Canine rabies can infect all warm-blooded animals, and death usually results. Some animals show chiefly paralytic signs, whereas others manifest encephalitic hyperexcitability and viciousness. However, the vampire bat may transmit virus for months while apparently not infected, and in other wild animals infection may take other courses than the fatal encephalitis. The virus will grow in chick embryos or in chick or mouse embryo tissue cultures. Strains freshly isolated from dogs and humans are called street virus; strains of altered pathogenicity and a stable, shortened incubation period, produced by serial passage in rabbit brains, are called fixed. Except in bats, street virus invariably produces cytoplasmic inclusion bodies (Negri bodies) in infected nerve cells. *See* EMBRYONATED EGG CULTURE; TISSUE CULTURE; VIRAL INCLUSION BODIES.

Pathogenesis. The virus is believed to move from the saliva-infected wound through sensory nerves to the central nervous system, multiply there with destruction of brain cells, and thus produce encephalitis, with severe excitement, throat spasm upon swallowing (hence hydrophobia, or fear of water), convulsions, and death—with paralysis sometimes intervening before death.

Diagnosis. Diagnosis in the human is made by observation of Negri bodies in brains of animals inoculated with the person's saliva, or in the person's brain after death. Diagnosis in dogs is essential for guidance concerning human vaccination. A dog which has bitten a person is isolated and watched for 10 days for signs of rabies; if none occur, rabies was absent. If signs do appear, the animal is killed and the brain examined for Negri bodies, or for rabies antigen by testing with fluorescent antibodies. *See* FLUORESCENCE MICROSCOPE.

Epidemiology. Rabies occurs throughout the world. About 1000 cases of human rabies are reported annually to the World Health Organization, almost all from developing countries. Only rarely are human cases reported in the United States, but a reservoir is constantly present, as shown by outbreaks in domestic and wild animals. In South and Central America, vampire bats and fruit- and insect-eating bats transmit the disease to cattle, and occasionally to humans. Since bats may carry the virus without becoming ill, they are important as a constant source of infection for wildlife predators, domestic animals, and human beings. In addition, aerosol transmission of rabies by bat has been proved, adding to the problem of rabies control, especially among persons who explore caves which are inhabited by bats.

Immunity and prevention. Individuals at high risk must receive preventive immunization. If exposure is believed to have been dangerous, postexposure prophylaxis should be undertaken. The rabies virus may remain latent in tissues for some time after being introduced from a bite, but if antibody or immunogenic vaccine is administered promptly, the virus can be prevented from invading the central nervous system. Passively administered antibody provides additional time for a vaccine to stimulate active antibody production before the central nervous system is invaded.

An inactivated rabies virus vaccine is available in the United States. It is made from virus grown in human or monkey cell cultures and is free from brain proteins that were present in earlier Pasteur-type

vaccines. This material is sufficiently antigenic that only four to six doses of virus need be given to obtain a substantial antibody response. An experimental recombinant virus oral vaccine, which consists of vaccinia virus carrying the rabies surface glycoprotein gene, has successfully immunized animals, and it should prove valuable in the immunization of both wildlife reservoir species and domestic animals. *See* BIOLOGICALS; VACCINATION.

Since 1960, one to five cases of human rabies have occurred annually in the United States, but 20,000 to 30,000 persons require some treatment for possible bite-wound exposure each year. All bites should immediately be cleaned thoroughly with soap and water, and a tetanus shot should be considered. The decision to administer rabies antibody, rabies vaccine, or both depends on four factors: the nature of the biting animal; the existence of rabies in the area; the manner of attack (provoked or unprovoked) and the severity of the bite and contamination by saliva of the animal; and recommendations by local public health officials. *See* PUBLIC HEALTH.

Control. Isolated countries, such as Great Britain, that have no indigenous rabies in wild animals can establish quarantine procedures; imported domesticated animals are quarantined for 6–12 months. In countries where dog rabies exists, stray animals should be destroyed, and vaccination of pet dogs and cats should be mandatory. Where wildlife rabies exists—and where contact between domestic animals, pets, and wildlife is inevitable—all domestic animals and pets should be vaccinated, and the incidence of rabies in wild animals should be continually monitored.

Preexposure vaccination is desirable for all persons who are at high risk of contact with rabid animals, particularly veterinarians, animal care personnel, certain laboratory workers, and spelunkers. Persons traveling to countries where rabies control programs for domestic animals are not optimal should receive preexposure prophylaxis if they plan to stay for more than 30 days. Persons who stay in rabies-endemic areas for extended periods and who are at risk of inapparent exposure to rabies or a delay in postexposure prophylaxis should be advised to have a booster vaccination every 2 years or to undergo serum testing for rabies-neutralizing antibody every 2 years. If the antibody titer is inadequate, booster vaccination is advisable. Preexposure prophylaxis does not, however, eliminate the need for postexposure prophylaxis if exposure to rabies occurs.   Joseph L. Melnick

Bibliography. D. M. Knipe et al. (eds.), *Field's Virology*, 4th ed., 2001; S. A. Plotkinet al. (eds.), *Vaccines*, 3d ed., 1999.

Raccoon

A member of the mammalian family Procyonidae, along with ringtails, coatis, kinkajous, olingos, and lesser pandas. These carnivores inhabit areas from southern Canada to northern Argentina and Uruguay

Raccoon, a member of the family Procyonidae.

and have been introduced into Russia and Germany. Only one species, *Procyon lotor*, inhabits North America.

Morphology. The raccoon is a medium-sized, heavily built, partially arboreal mammal with long, loose, gray to almost black fur (see **illustration**). A white-bordered black mask extends across the eyes and down onto the cheeks. The foxlike head is broad with a pointed muzzle, and the ears are prominent, erect, and somewhat pointed. The moderately long, well-haired tail is banded with five to seven blackish rings alternating with wider, lighter bands. The plantigrade feet have naked smooth soles and five very long digits, each of which possesses a nonretractile claw. No webbing is present between the toes. The "hands" are well adapted for grasping and manipulating objects. The dental formula is I 3/3, C 1/1, Pm 4/4, M 2/2 for a total of 40 teeth. Raccoons have excellent senses of hearing, vision, smell, and touch. *See* DENTITION.

Adult raccoons are 79.5–135 cm (31–53 in.) long, including a 19–40 cm (7–16 in.) long tail. They stand 22.8–30.4 cm (9–12 in.) at the shoulder and generally weigh 3–10 kg (6–21 lb). Males tend to be 20 to 30% larger than females. Body size is correlated with severity of the climate; raccoons with the smallest overall body size are found in the subtropical Florida Keys, and those with the largest body size are found in areas with 100 or more days of snow cover.

Ecology. Raccoons prefer moist areas and are found mainly in timbered swamps, on river bottoms, along the banks of streams and lakes, and in coastal salt marshes. These primarily nocturnal mammals are excellent climbers, and they swim well, but not far or

often. Dens may be in hollow trees, hollow logs, old stumps, burrows, muskrat houses, or in outbuildings such as barns. Seasonal activity cycles vary with latitude. Raccoons do not hibernate, although in northern areas they may remain in a den for much of the winter. Although normally a solitary species, communal denning is common with up to 23 raccoons having been reported in a single den.

Raccoons are opportunistic omnivores and feed on a wide variety of plant and animal foods including wild fruits, berries, acorns, shellfish, crustaceans, frogs, fish, small mammals, reptiles, and birds. Most male raccoons range over about 600 hectares (1500 acres) annually. Female home ranges are smaller and often overlap.

Mating and development. Mating in the United States extends from December to August, with peaks in February and March. Following an average gestation of 63 days, the annual litter is born in April or May. Litter sizes average 3 or 4, but range 1–7. Newborn animals have blackish skin and are well covered with yellowish-gray fur. Indistinct darker areas show where the facial mask and tail rings will appear prominently later. Weaning is completed at 3–4 months of age. Mother and young remain together into the late fall and often well into the winter. Sexual maturation is attained at about 12–15 months, but most raccoons do not mate until the following year. The average life span is 2–3 years in the wild, although some wild individuals have lived more than 12 years.

Threats. The activities of humans (hunting, trapping, and automobiles) are the principal causes of raccoon mortality. Though the demand for its pelt has fluctuated, the raccoon has long been an important fur species in North America. Foxes, bobcats, hawks, owls, and snakes may prey on young raccoons. Distemper and rabies can temporarily reduce populations.

Other species. Central and South American raccoons include the Tres Marias raccoon (*P. insularis*) which inhabits Maria Madre Island, Mexico; the Barbados raccoon (*P. gloveralleni*) of Barbados; the crab-eating raccoon (*P. cancrivorus*) that ranges from Costa Rica south to northern Argentina; the Cozumel Island raccoon (*P. pygmaeus*) of Cozumel Island, Yucatan, Mexico; the Guadeloupe raccoon (*P. minor*) of Guadeloupe; and *P. maynardi* found on Nassau Island in the Bahamas. Recent researchers feel that the insular forms (*P. insularis*, *P. gloveralleni*, *P. pygmaeus*, *P. minor*, and *P. maynardi*) are conspecific with *P. lotor*. All of the insular forms are classified as Endangered. The Barbados raccoon may be extinct. *See* CARNIVORA; MAMMALIA.   Donald W. Linzey

Bibliography. *Grzimek's Encyclopedia of Mammals*, McGraw-Hill, 1990; D. W. Linzey, *The Mammals of Virginia*, McDonald & Woodward, 1998; D. Mcdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford Limited, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999; J. O. Whitaker, Jr. and W. J. Hamilton, Jr., *Mammals of the Eastern United States*, 3d ed., Comstock Publishing Associates, 1998; D. E. Wilson and S. Ruff (eds.), *The Smithsonian Book of North American Mammals*, Smithsonian Institution Press, 1999.

Racemization

The formation of a racemate from a pure enantiomer. Alternatively stated, racemization is the conversion of one enantiomer into a 50:50 mixture of the two enantiomers (+ and −, or R and S) of a substance. Racemization is normally associated with the loss of optical activity over a period of time since 50:50 mixtures of enantiomers are optically inactive. *See* OPTICAL ACTIVITY.

Racemization is an energetically favored process since it reflects a change from a more ordered to a more random state. But the rate at which enantiomers racemize is typically quite slow unless a suitable mechanistic pathway is available, since racemization usually, but not always, requires that a chemical bond at the chiral center of an enantiomer be broken. Racemization of enantiomers possessing more than one chiral center requires that all chiral centers of half of the molecules invert their configurations. *See* ENTROPY.

Mechanisms. Of the several known racemization mechanisms, those involving the temporary formation of intermediates which possess reflection symmetry are the most common. Depending upon the substance and the conditions employed, the intermediate may be a free radical, carbocation, or carbanion. The racemization of an alkyl halide in the presence of a Lewis acid catalyst [reaction (1)]
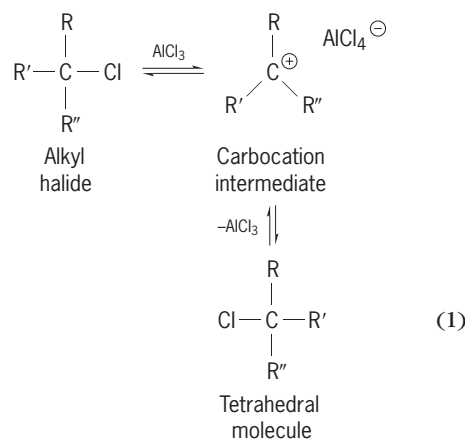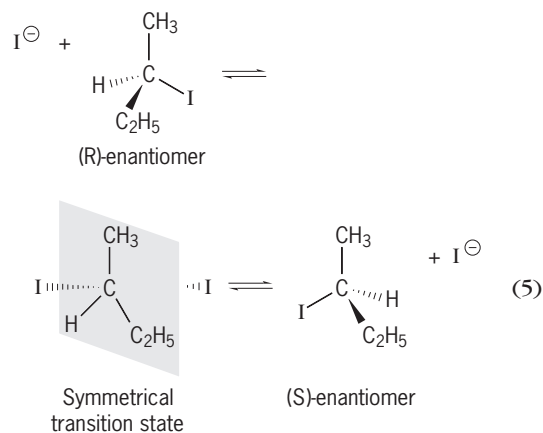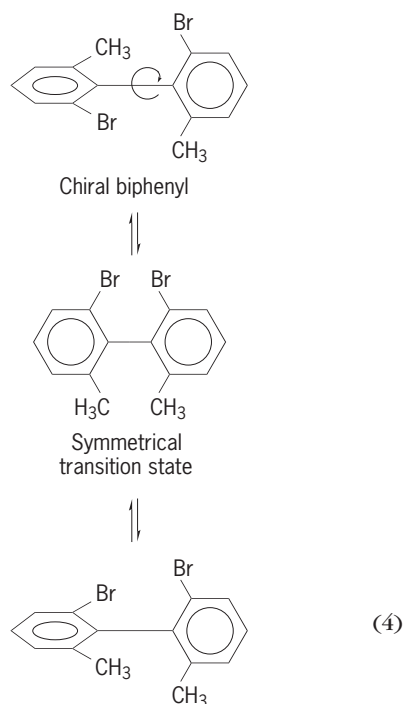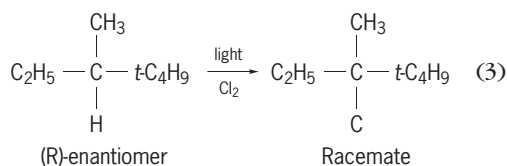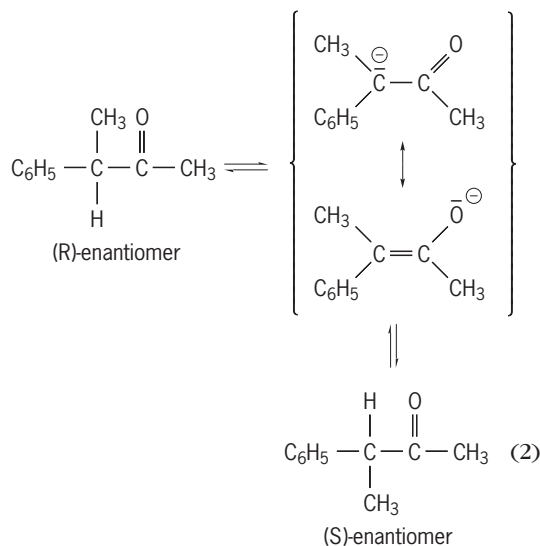

illustrates the intervention of a planar, hence achiral, carbocation intermediate. Reversal of the ionization or capture of the cation by a solvent molecule can occur with equal probability to either side of the planar intermediate which then reforms a molecule of tetrahedral geometry. When equal numbers of (R)- and (S)-enantiomers are present, racemization is complete.

Chiral compounds with acidic hydrogens at the chiral center are racemized by base with the

intervention of carbanions [reaction (2)]. Delocalization of the unshared electrons makes the enolate intermediate planar. For carbanions for which delocalization is not possible, racemization takes place because a nonplanar carbanion usually inverts its configuration readily.

Racemization via free radicals is exemplified by the chlorination of hydrocarbon which involves an intermediate free radical mechanism [reaction (3)].

(R)-enantiomer

(2)

(S)-enantiomer

(3)

(R)-enantiomer   Racemate

Chiral biphenyl

Symmetrical
transition state

(4)

(R)-enantiomer

Symmetrical
transition state   (S)-enantiomer

(5)

The product is racemic. The *t* in the structures stands for tertiary (as intertertiary butyl group) and is a descriptor defining one of the isomers of the butyl group (C_4H_9). Heat can also cause racemization without breaking bonds if it is possible for a molecule to pass through a state which possesses reflection symmetry. For example, the chiral biphenyl may be racemized thermally by rotation about the central carbon-carbon bond. In this case the benzene rings become coplanar in the transition state [reaction (4)]. Nucleophilic displacement on chiral alkylhalides may also lead to racemization via symmetrical transition states if the displacing ion is identical to the leaving ion [reaction (5)]. This type of change in configuration is known as Walden inversion. *See* FREE RADICAL; ORGANIC REACTION MECHANISM; REACTIVE INTERMEDIATES.

Significance. The observation and study of racemization have important implications for the understanding of the mechanisms of chemical reactions and for the synthesis and analysis of chiral natural products such as peptides. Moreover, racemization is of economic importance since it provides a way of converting an unwanted enantiomer into a useful one. Synthetic medicinal agents are often produced industrially as racemates. After resolution and isolation of the desired enantiomer, half of the product would have to be discarded were it not for the possibility of racemizing the unwanted isomer and of recycling the resultant racemate.

Inversion of configuration at some but not all chiral centers in diastereomeric peptides and sugars does not lead to total loss of configuration. This process, while analogous to racemization, does not lead to complete loss of optical activity. It is called epimerization. Epimers are diastereomers which differ at only one of several chiral centers. Epimerization of amino acids in proteins takes place very slowly on standing, and is used as a means of dating very old protein samples (aminostratigraphy). *See* STEREOCHEMISTRY.   Samuel H. Wilen

Bibliography. E. L. Eliel and S. H. Wilen, *Stereochemistry of Organic Compounds*, 1994; D. Naspuri, *Stereochemistry of Organic Compounds: Principles and Applications*, 1991; M. Nogradi, *Stereoselective Synthesis*, 2d ed., 1994; F. Vogtle and E. Weber (eds.), *Stereochemistry*, 1984.

Radar

An acronym for radio detection and ranging, the original and basic function of radar. The name is applied to both the technique and the equipment used. Radar is a sensor; its purpose is to provide estimates of certain characteristics of its surroundings of interest to a user, commonly the presence, position, and motion of aircraft, ships, or other vehicles in its vicinity. In other uses, radars provide information about the Earth's surface (or that of other astronomical bodies) or about meteorological conditions. To provide the user with a full range of sensor capability, radars are often used in combinations or with other elements of the complete system.

Radar operates by transmitting electromagnetic energy into the surroundings and detecting energy reflected by objects. If a narrow beam of this energy is transmitted by the directive antenna, the direction from which reflections come and hence the bearing of the object may be estimated. The distance to the reflecting object is estimated by measuring the period between the transmission of the radar pulse and reception of the echo. In most radar applications, this period will be very short since electromagnetic energy travels with the velocity of light.

Kinds of radar. Radar has so many valuable applications that the physical nature of radars varies greatly. Familiar radars that are very small and compact include those used on small boats or in small aircraft for navigation and safety and, even smaller, the handheld radars used for measurement of vehicle speed or even that of a baseball pitch. See MARINE NAVIGATION.

The largest radars cover acres of land, long arrays of antennas all operating together to monitor the flight of space vehicles and orbiting debris, an increasing space-flight problem, and also to contribute to radar astronomy. Other very large radars provide coverage at substantial distances well beyond the normal radar horizon; they monitor flight activity and sea-surface conditions, contributing to both air-traffic and meteorological interests. These radars are very large because they must use longer-than-usual radio wavelengths [those in the high-frequency (HF) part of the radio spectrum, and not the much higher microwave frequencies more common in radar] that result in the ionospheric containment of the signal for over-the-horizon operations; these longer wavelengths require very large antennas to form the narrow beams useful in the measurement of target positions or the surface spot being observed. See RADAR ASTRONOMY; SPACECRAFT GROUND INSTRUMENTATION.

More common in size are those radars at airports, with rotating antennas 20–40 ft (6–12 m) wide, a room or two of electronic equipment, and operator consoles the size of a desk. Radars intended for mobile use, particularly those in aircraft and in trucks, buses, and automobiles, are quite compact and very modular in construction, well adapted to fit into the vehicle. See AIRBORNE RADAR.

Radars intended principally to determine the presence and position of reflecting targets in a region around the radar are called search radars. Other radars examine further the targets detected: examples are height finders with antennas that scan vertically in the direction of an assigned target, and tracking radars that are aimed continuously at an assigned target to obtain great accuracy in estimating target motion. In some modern radars, these search and track functions are combined, usually with some computer control. Surveillance radar connotes operation of this sort, somewhat more than just search alone. See SURVEILLANCE RADAR.

The development of quite complex and versatile radars, computer-controlled, generally using

(a)

(b)

Fig. 1. Multifunction phased-array radars. (*a*) One of the four planar faces of the radar of the AEGIS combat system, which is in use on many U.S. cruisers and destroyers (*Lockheed Martin*). (*b*) Pave Paws radar installation. (*Raytheon Company*)

phased-array antennas and called multifunction radars, has been quite successful worldwide. Such antennas (described below) require no physical motion to position the radar beam, so various coverage routines and modes of operation can be scheduled with an efficiency and expediency unachievable in lesser radars. The U.S. Navy's AEGIS combat system features such a radar (**Fig. 1***a*), as does the Army's Patriot missile system. Large land-based multifunction phased-array radars include the Cobra Dane (and variants) system and the Pave Paws radars (Fig. 1*b*), which have contributed to United States defense surveillance for several decades. Pave Paws, as an example, used two array faces, each 72.5 ft (21.8 m) in diameter composed of approximately 14,000 solid-state radiating elements, operating in the UHF frequency band. Many other nations have now deployed such radars of various scale, and developments continue in such systems for land-based, shipboard, airborne, and space-based radars. No longer are such multifunction radars a developmental curiosity.

An important airborne multifunction radar that uses mechanical rotation of its antenna in azimuth and electronic beam positioning in elevation is the Air Force's S-band AWACS (Airborne Warning and Control System) radar, the antenna of which is mounted in a rotodome atop the aircraft (**Fig. 2***a*). The precision-built waveguide array (Fig. 2*b*) features extremely low sidelobes, essential in almost any airborne radar because of the strong ground echo at all angles. New multimission military aircraft will include multifunction radars as an integral part of their design. *See* MILITARY AIRCRAFT.

Very accurate tracking radars intended for use at missile test sites or similar test ranges are called instrumentation radars. Radars designed to detect clouds and precipitation are called meteorological or weather radars. Improved techniques in Doppler-sensitive signal processing have enabled radars to examine more precisely the motion of reflecting objects (windborne dust, insects, and precipitants, for example) indicative of weather disturbances. Such radars are used in critical places such as airports where sudden wind shears (as from downdraft microbursts in the area) have threatened air safety.

Many weather radars also use microwave polarimetry to examine the precipitants and characterize the disturbance more accurately. *See* DOPPLER EFFECT; DOPPLER RADAR; METEOROLOGICAL RADAR; STORM DETECTION.

Airborne and spaceborne radars have been developed to perform ground mapping with extraordinary resolution by special Doppler-sensitive processing while the radar is moved over a substantial distance. Such radars are called synthetic-aperture radars (SARs) because of the very large virtual antenna formed by the path covered while the processing is performed. Interferometry can provide topological information (3D SAR), and polarimetry and other signal analysis can provide more information on the nature of the surface (type of vegetation, for example). *See* REMOTE SENSING; SYNTHETIC APERTURE RADAR (SAR).

Some radars have separate transmit and receive antennas sometimes located miles apart. These are called bistatic radars, the more conventional single-antenna radar being monostatic. Some useful systems have no transmitter at all and are equipped to measure, for radarlike purposes, signals from the targets themselves. Such systems are often called passive radars, but the terms "radiometers" or "signal intercept systems" are generally more appropriate. Regular radars may, of course, occasionally be operated in passive modes. Doing so helps locate sources of interference, either deliberate (jamming) or unintentional. Estimating the angle-of-arrival of signal sources can be used to draw other sensors (infrared sensors, for example; other active radars) to those angles for further examination. Angle data also contribute to triangulation processes (with similar data from other sensors) to locate sources more completely. The spectral data, showing perhaps the less occupied frequencies, can be used in governing regular radar operations somewhat adaptively. Passive radars (or modes) can be equipped to analyze the signals being received to help identify targets by association, a common data fusion process in military systems particularly. *See* PASSIVE RADAR.

The terms "primary" and "secondary" are used to describe, respectively, radars in which the signal received is reflected by the target and radars in which

(a)

(b)

Fig. 2. Air Force AWACS (Airborne Warning and Control System) in use. (*a*) AN/APY-1 rotodome installed. (*b*) Precision-array antenna of the AN/APY-1. (*Northrop/Grumman Corp.*)

the transmission causes a transponder (transmitter-responder) carried aboard the target to transmit a signal back to the radar. The Identification Friend or Foe (IFF) system in both military and civil use is a secondary radar. *See* AIR-TRAFFIC CONTROL; ELECTRONIC NAVIGATION SYSTEMS.

Fundamentals of operation. It is convenient to consider radars as having four principal subsystems: the transmitter, the antenna, the receiver (and signal processor), and the control and interface apparatus.

The transmitter provides the radio-frequency (RF) signal in sufficient strength (power) for the radar sensitivity desired and sends it to the antenna, which causes the signal to be radiated into space in a desired direction. The signal propagates (radiates) in space, and some of it is intercepted by reflecting (or "scattering") bodies. The part of the incident signal that is backscattered returns to the radar antenna, which collects it and routes it to the receiver (along with all incoming signals) for appropriate signal processing and detection. The presence of an echo of the transmitted signal in the received signal reveals the presence of a target. The echo is indicated by a sudden rise in the output of the detector, which produces a voltage (video) proportional to the sum of the RF signals being received and the RF noise inherent in the receiver itself. The range to the target is half the product of the speed of electromagnetic propagation (3×10^8 m/s), and the time between the transmission and the receipt of the echo. The direction or bearing of the target is disclosed by the direction to which the beam is pointed when an echo is received.

In some radars, pulsing may be too rapid to permit range determination by direct echo timing in this unambiguous way. Such radars are said to employ higher pulse-repetition-frequency (prf) modes of operation; the higher sampling rate is necessary for clearly distinguishing moving targets from highly reflective but nearly stationary backgrounds (clutter) through the Doppler effect involved. Range determination then requires resolution of the ambiguities in echo timing by using more than one such pulse repetition frequency. Airborne radars are particularly dependent upon such Doppler-sensitive modes of operation.

Most radars indeed use pulses in their operation. Those that do not are called continuous-wave (cw) radars. Some missile seekers, mostly interested only in the direction to a target, are of this type; police radars interested only in the Doppler shift of the echo for speed measurement are another example. In pulsed radars (**Fig. 3**) with a transmitter of the master-oscillator power-amplifier (MOPA) type, the conversion of electrical power to transmitted signal power takes place in a chain of amplifiers of increasing power level, each controlled by the modulator forming the radar pulse. In some simpler radars, a single high-power oscillator may be used with no further amplification. In newer high-performance radars, the final stage of amplification may be in transmit-receive modules in the antenna structure itself, one such module at each element of a large active array. *See* AMPLIFIER; CONTINUOUS-WAVE RADAR; MODULATOR; OSCILLATOR; POWER AMPLIFIER; RADIO TRANSMITTER.

A duplexer permits the same antenna to be used on both transmit and receive, and is equipped with protective devices to block the very strong transmit signal from going to the sensitive receiver and damaging it. The antenna forms a beam, usually quite directive, and, in the search example, rotates throughout the region to be searched, or otherwise causes the beam to be positioned as desired, perhaps without mechanical motion, as in radars using a computer-controlled phased-array antenna (Fig. 3*b,c*). In such phased arrays, the duplexing function is generally built into the feed system and phasing logic. *See* ANTENNA (ELECTROMAGNETISM).

Radars searching with mechanical antennas generally transmit several pulses (perhaps 20–40) as the antenna scans past a target, whereas phased-array radars may search with various "dwells," sometimes with just one transmission per beam position. The use of several pulses allows a buildup of the echo being received, a process called noncoherent or video integration. Most radars are equipped with low-noise RF preamplifiers to improve sensitivity. The signal is then "mixed" with (multiplied by) a local oscillator signal to produce a convenient intermediate-frequency (IF) signal, commonly at 30 or 60 MHz; the same principle is used in all heterodyne radio receivers. The local oscillator signal, kept offset from the transmit frequency by precisely this intermediate frequency, is supplied by the transmitter oscillators during reception. After other significant signal processing in the IF circuitry (of a digital nature in many newer radars), a detector produces a video signal, a voltage proportional to the strength of the processed IF signal. This video can be applied to a cathode-ray-tube (CRT) display so as to form a proportionately bright spot (a blip), which could be judged as resulting from a target echo. However, radars increasingly use artificial computerlike displays based on computer analysis of the video. Automatic detection and automatic tracking (based on a sequence of dwells) are typical of such data processing, reports being displayed for radar operator management and also made instantly available to the user system. *See* CATHODE-RAY TUBE; ELECTRONIC DISPLAY; HETERODYNE PRINCIPLE; MIXER; PREAMPLIFIER; RADIO RECEIVER.

The waveform of an elementary search radar consists of a train of pulses. Duty factors (the ratio of pulse width or transmit "on" time, to total time) in such radars are typically 0.1 to 1.0%. With modern signal processing, much more complicated waveforms are often used, and with the advent of transistors and other solid-state amplifiers, radar systems have been developed with duty factors of 10 to 50%.

Radar carrier-frequency bands. Radar carrier frequencies are broadly identified by a widely accepted nomenclature. The spectrum is divided into bands (see **table**). In addition, some special-purpose
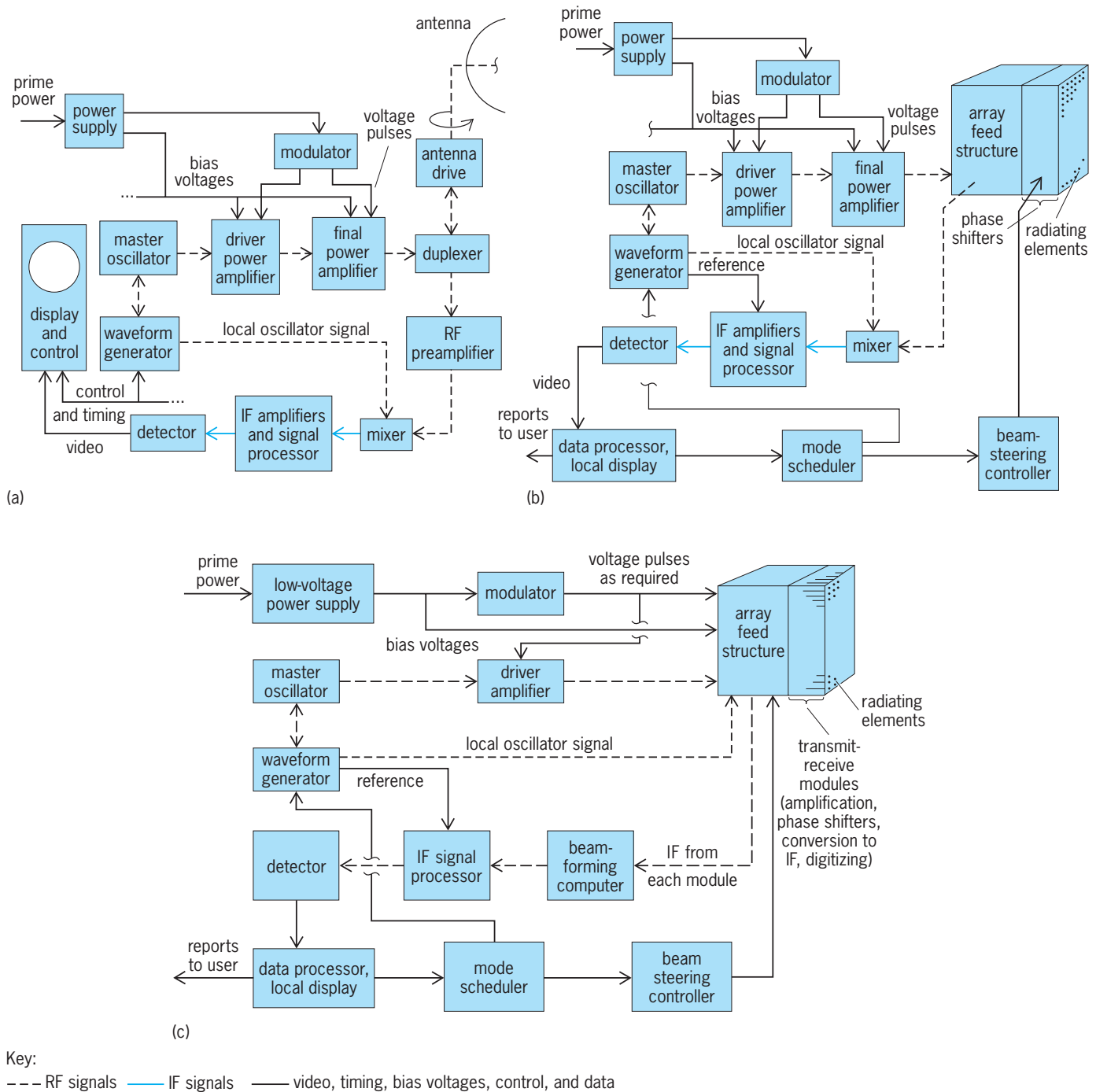
Fig. 3. Block diagram of pulse radar. (*a*) Radar with a mechanically scanned antenna. (*b*) Radar with a phased-array antenna for electronic scanning. Array feed structure includes transmit distribution, receive beam forming, duplexing, and preamplification. (*c*) Radar with active phased array using transmit-receive modules at every element and digital beam forming on receive.

radars are operated with laser transmitters at infrared and near-optical frequencies. *See* LIDAR; RADIO SPECTRUM ALLOCATIONS.

Propagation. There are several noteworthy characteristics of the radar's surroundings in terrestrial operation. The charged layers of the ionosphere present a highly refractive shell at radio frequencies well below the microwave frequencies of most radars. Consequently, over-the-horizon radars operate in the 10-MHz region to exploit this skip path. Further,

the Earth's surface makes such an abrupt change of medium that electromagnetic reflections take place. In microwave radars these forward-scattered surface reflections cause a "multipath" condition resulting in the lobed interference structure seen in field strength graphs of radar coverage (**Fig. 4**). These lobes in the plane normal to the reflecting surface result from the alternately constructive and destructive phase interference of the direct and the reflected signals, varying with the elevation angle. When

Radar carrier-frequency bands		
Band designation	Nominal frequency range	Representative wavelength
HF	3–20 MHz	30 m at 10 MHz
VHF	30–300 MHz	3 m at 100 MHz
UHF	300–1000 MHz	1 m at 300 MHz
L	1.0–2.0 GHz	30 cm at 1 GHz
S	2.0–4.0 GHz	10 cm at 3 GHz
C	4.0–8.0 GHz	5 cm at 6 GHz
X	8.0–12.0 GHz	3 cm at 10 GHz
K_u	12.0–18 GHz	2 cm at 15 GHz
K	18–27 GHz	1.5 cm at 20 GHz
K_a	27–40 GHz	1 cm at 30 GHz
mm	40–300 GHz	0.3 cm at 100 GHz

terrain is not so reflective, the lobe structure is less pronounced. Because there is always destructive interference (giving minimum signal strength) at 0° elevation, there exists an inherent performance limitation for very low flying targets, often said to fly under the radar beam. Similarly, narrow-beam tracking radars have difficulty measuring the true elevation angle of low-angle targets when the beam grazes the Earth's surface and incurs multipath interference. *See* INTERFERENCE OF WAVES; IONOSPHERE; RADIO-WAVE PROPAGATION; REFLECTION OF ELECTROMAGNETIC RADIATION.

The atmosphere also acts as a lens since its dielectric constant decreases with increasing altitude. Consequently, microwave rays are bent downward slightly, and the radar horizon is somewhat beyond the visual. Under certain conditions defined as standard, the effect is such that if one were to draw the Earth with a radius 4/3 its actual value, the refracted rays would be bent back to straight lines.

Frequently, an atmospheric condition involving a pronounced departure from the smoothly varying dielectric constant with altitude occurs, and a superrefractive or ducting condition results with low-altitude targets detected well beyond 4/3-Earth predictions. The variation in such conditions over time of day, season, geographic location, and local weather makes reliable accounting for the phenomenon difficult. Attenuation of radar signals in the

antenna's "free-space" vertical pattern

interference lobes in resulting pattern

height

range

radar at origin

(a)   (b)

Fig. 4. Effect of a reflecting Earth surface on the radiation pattern of an antenna. (*a*) Reflection nearly perfect. (*b*) Reflection much reduced.

Earth's atmosphere is due to both molecular absorption by resonant excitation of uncondensed gases and scattering by particles such as dust or water-droplets in fog and clouds. The attenuation due to water vapor and to oxygen molecules is negligible below the L-band and tolerable through the X-band. The resonance peaks for these molecules cause much more concern in choosing frequencies in the millimeter region for special-purpose radars. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION.

Target characteristics. This discussion assumes that targets of interest are vehicles such as aircraft, trucks, ships, or spacecraft. Terrain features, buildings, and vegetation are certainly of interest to other (imaging) radars in which signal processing addresses the reflectivity, polarimetric, and other characteristics of such distributed targets. Targets such as vehicles are composed of many scatterers or reflecting surfaces, all of which contribute to the whole reflection. The individual reflections add coherently (that is, with both constructive and destructive phase interference, depending on the many path lengths involved) in composing the total reflection. Consequently, over successive observations, the reflections will differ in magnitude if even the slightest repositioning of the scatterers occurs, or if a different carrier frequency is employed. *See* COHERENCE.

Such targets have been categorized by the statistical nature of these fluctuations in echo strength; the Swerling models (after Peter Swerling) serve as a basis. Advances in more precise modeling of the composition of targets, aided by computer analysis, provide assistance to the designers of both radars and aircraft (or other potential targets). A single average value of a target's radar cross section often suffices for basic radar calculations, provided a statistical allowance is made for the fluctuations. Some targets are designed to have very little reflection back to the radar; stealth aircraft and missiles are examples. Others, however, may be designed to concentrate the reflections back to the radar. For safety reasons, small boats and aircraft are often equipped with small corner reflectors of right-intersecting planes, so that the radar signal is reflected back to the radar regardless of the direction from which it came, making the radar return far larger and more visible to a searching radar than would be the random reflections from the target itself. These and other retrodirective reflectors are available in various physical sizes and aid in instrumentation, navigation, and search-and-rescue situations.

The strength of the echo from complex targets, then, is properly treated as a random variable. An expected value (a mean value) might be used in elementary performance calculations for radar. Radar cross section is defined as a fictitious area which, if intercepting the incident signal and reradiating isotropically the signal intercepted, would return to the radar the same strength echo that the real target returns. Radar cross sections of ships can be in the thousands of square meters, of commercial aircraft several tens of square meters, of smaller
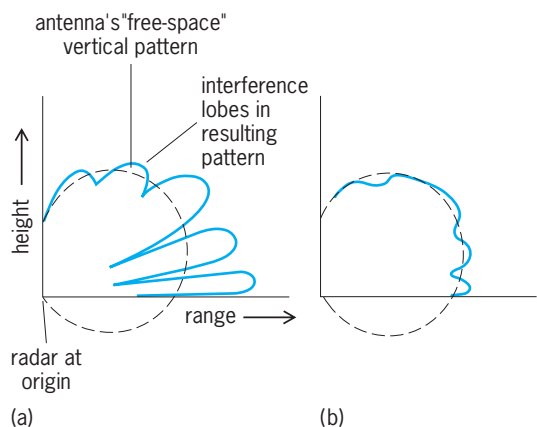
aircraft and missiles a few square meters, and of deliberately stealthy aircraft a small fraction of a square meter (perhaps about the same as a bird or large insect). Retro-reflective aids to navigation can have radar cross sections of tens or hundreds of square meters, even though they are small in physical size.

Noise. The detectability of a radar echo is determined by its strength relative to various competing signals also causing a detector output. Among the sources of such signals is the receiver itself. The receiver produces RF and IF signals as a result of thermal conditions (molecular agitation in the components) and of the mixing process (unintended constituents of the local oscillator signal, for example).

A measure of the receiver's noise production is called its noise factor. In general terms, the noise factor is the ratio of noise power at the output of a network to the noise power that would have resulted from the network simply acting on the noise power at the input. The added noise of the network itself causes a degradation in the signal-to-noise ratio from input to output. It can be seen, then, that a sensitive radar receiver might to good advantage use a first stage of low noise factor but of considerable gain, so that subsequent stages are provided with a relatively strong signal (and noise) at their inputs, their own noise production further degrading the now-amplified signal-to-noise ratio very little. Noise factor when expressed in decibels is more commonly called noise figure. *See* ELECTRICAL NOISE.

Other noise sources can be external to the radar. They include the Sun, other cosmic sources, the Earth's own surface, electromagnetic interference from urban areas, and deliberately energized sources intended to reduce the radar's sensitivity (jamming). *See* ELECTRICAL INTERFERENCE; ELECTRONIC WARFARE; JAMMING; RADIO ASTRONOMY.

Clutter. While some radars are designed to examine land features and others, weather phenomena, generally such radar returns are considered a nuisance, hindering the detection of targets of greater interest (such as aircraft) and producing false alarms. It is possible to state elementary formulas for the power of such returns, but in general the simple expressions fail to account sufficiently for the statistical distribution of clutter backscatter encountered in practice.

Several types of clutter have been carefully investigated. Surface clutter includes scattering from the sea surface (for various sea-state conditions) and from land surface (for various terrain types). Volume clutter includes fog, clouds, rain, snow, and chaff (reflective threads deliberately dispersed in the air as a radar countermeasure), and even birds and insects have been taken into account. Clear-air turbulence and thermal inversions can be reflective; these are often called angels by radar users. *See* CLEAR-AIR TURBULENCE.

In the radar bands, clutter backscatter generally increases with increasing frequencies—in a given amount of rain, the clutter power will be worse for an X-band radar than for an L-band one. Radars using frequencies below L-band are often called all-weather radars. Radars at S-band and higher may use high resolution (narrow beams, very short pulses) in part to reduce the amount of clutter competing with a target at any one time.

In addition to the strength of the backscatter, the spread in frequency of clutter return is of concern in radar design. Such spectral spread in the Doppler-shifted return comes from turbulence among the scatterers as experienced by a surface-based radar; it is particularly severe in airborne radar because the return from the Earth's surface comes from all angles (even through the sidelobes of the directive antenna) while the aircraft moves. In signal processing intended to suppress returns from presumed stationary clutter, this clutter spectrum is a major consideration. Airborne radars generally feature modes of pulse Doppler processing that are unnecessary in most surface-based radars.

Detection process. Since a radar receiver is always exposed to a background of noise (and possibly other disturbance such as clutter, interference), the declaring at some instant that a desired signal (a target echo) is also present is clearly a probabilistic matter. The detection process is, in fact, a direct application of hypothesis testing in which one must decide whether the receiver's output (the video) was produced by the background only or by a target echo in addition to the background (the two hypotheses).

There are, of course, two ways to err in such a process: first, to declare a target to exist when none does (a false alarm), and second, to fail to declare a target when in fact one exists (a miss). Statistical methods are used to quantify the probabilities of false alarm and detection (the complement of miss). A useful relationship, often used in very general radar analysis and with certain simplifying assumptions about the background and the target, is that a ratio of target signal to average noise of about 13 decibels (13 dB, a power ratio of 20:1) permits detection with 0.5 probability while maintaining a false alarm probability of just 10^{-6}. Since a typical radar might be making about one million decisions per second (resolution-dependent—here as if using a pulse width of 1 microsecond), a threshold setting producing 10^{-6} false alarm probability would yield about one false alarm each second, a rate perhaps quite tolerable as an input to successive steps (multipulse integration, confirmation) of the entire detection process. An often misunderstood point in radar design is that this initial threshold should be as low as possible—for as high a rate of false alarms as can be tolerated—so that maximum possible sensitivity to the desired signal is achieved.

Today, the detection process often uses computer analysis of the receiver's output in the radar's control and interface subsystem, described more fully later. In older radars with a human operator viewing a cathode-ray-tube (CRT) display of that output (the video), the process is quite the same, even if unwittingly so. When the operator makes receiver

and display adjustments so that the background is just barely visible (minimally exciting the CRT), he is actually adjusting the "threshold" relative to the experienced background so that he would not likely mistake a noise excursion for a target and yet be reasonably likely to recognize a true echo causing a bright spot, a "blip," on the display. *See* CATHODE-RAY TUBE.

In modern automatic detection, a threshold voltage is established based on a relatively large amount of data by which the background—no matter how composed of clutter, jamming, and noise—is estimated. Detections are declared whenever the video exceeds this threshold and some further criteria are met (as with dwells of multiple transmissions). Such automatic detection may promise no particular improvement in basic sensitivity, but offers very consistent performance and a greater capacity in dense target situations than could be expected of human decision making. Abrupt changes in the background must be recognized in such automation and the threshold adjusted to maintain a constant false alarm rate (CFAR); this is perfectly analogous to human operators tempering their declaring of targets when they sense changes in the background shown on the display.

In automated radar, then, an operator's display will likely show symbols and other computer-generated indications of targets and local conditions useful for his further action and overall management of the system, but generally will not show the receiver's video output directly.

The advent of stealth targets (ones with very low radar cross section) has prompted radar detection methods that involve more than thresholding on signal strength. Processes involving polarimetric signature, greater Doppler dependence, and track behavior estimation have been refined to accommodate detection of target signals no stronger than the background present but differing from it in these other ways.

Tracking. Many processes are used in radar systems to associate successive declared detections (now called contacts or plots) to form a history and a prediction of target motion, that is, to form its track. Historically, such processing was done rather manually, on or near the radar display, showing the tracks on a plotting board. Modern radars use computers to form such tracks by associating the contacts into probable sequences, that is, the tracks. In dense target environments the association process itself is a challenge. Methods of accounting for measurement accuracy, the quality (strength) of each contact, the localized history of successful and failed associations (indicating local confusing clutter conditions or the like), and very probabilistic analysis involving multiple associations as multiple hypotheses (until untrue pairings naturally fail) have all been explored to help in this important step. After association, the track is formed using mathematical algorithms for estimating course and speed (the velocity vector) which are called filters; the process is called smoothing. In automated radars that use regularly

scanning antennas, this processing is called track-while-scan; in multifunction phased-array radars, it is called sampled-data-tracking since track updating can be scheduled quite independently of the search mode. Adaptive tracking refers usually to using track filters in which some of the terms can be adjusted to the behavior of either the track or the radar itself being experienced. Sometimes multiple radars (or other sensors) share a common field of view and separately generated contacts are used to form a single set of tracks; this is called integrated automatic tracking, a form of data fusion. *See* ESTIMATION THEORY.

Radars have been built that are dedicated to tracking just one target but to do so very accurately (**Fig. 5**). In these radars a dish antenna is steered to the direction of a known target. Such a designation may have come from a search radar, and the tracking radar may have to perform a small local search in order to acquire the target. After acquisition, a continuous stream of contacts results and is used to keep the antenna steered to the target. The dish antenna produces a beam narrow in both horizontal and vertical angles (a pencil beam). If this beam is scanned in a very small cone around the target's presumed direction, the amplitude and phasing of the resulting modulations of the returns can be detected and used

Fig. 5. AN/FPS-16 tracking radar, a very accurate monopulse radar, operating in the C-band, installed at more than 50 sites around the world. (*RCA Corp.*)

to drive the motors that control the position of the antenna.

In these con-scan trackers, usually only the feedhorn in front of the dish is caused to nutate; the whole dish and feed assembly is repositioned by the director. The final estimate of the target's angle is made by sensing the director's position; the inertia due to its mass acts as a smoothing filter in the track estimate.

Another technique used in such tracking radars is monopulse angle measurement, in which multiple beams are generated by a cluster of feedhorns in a conventional antenna or by special beam-forming networks in phased arrays. A sequence of transmissions is not required, as it is in conical scanning, and fine angle measurement can be made on each transmission. While a tracking radar with a mechanical antenna (such as used in instrumentation radars and in missile guidance and gunfire control systems) dwells continuously on just one target, phased arrays permit scheduling of track dwells only as required to maintain a specified accuracy, and do so on dozens of targets at a time. *See* MONOPULSE RADAR.

Composition of radar systems. It is convenient to discuss radar systems in terms of four principal subsystems, namely the transmitter, the antenna, the receiver and signal processor, and the control and interface apparatus (likely including a display of some sort).

Transmitter. The transmitter converts electrical power (as from a motor-generator set or a utility system) to electromagnetic power at the carrier frequency and in the waveform desired. The three principal components of a transmitter are the power supply, the modulator, and the RF devices. The power supply changes the utility electrical power to the levels and type of electrical power needed by the particular transmitter, the modulator applies that power to the RF devices, and the RF devices convert that applied electrical power to electromagnetic power, the RF signal to be sent to the antenna and radiated.

The main functions of power supplies are transformation, rectification, and filtering. Step-up transformers change the voltage levels of the input (or prime) utility power to the high voltage required, rectifiers change that power from alternating current (ac) to direct current (dc) required by the RF devices, and the filters smooth out the dc high voltage, removing the ripples inherent in rectification. Radar power supplies must also be well regulated, which means that the voltage levels applied to the RF devices must remain constant regardless of the power being drawn by the various waveforms and modes of operation involved. Transmitters using solid-state amplification only (described below) will generally not require the high voltages (often tens of kilovolts) required for high-power vacuum tubes, but good regulation and the handling of higher total current (for similar power levels) must be accommodated. *See* ELECTRIC FILTER; ELECTRONIC POWER SUPPLY; RECTIFIER; RIPPLE VOLTAGE; TRANSFORMER.

Modulators act as switches to apply the power-supply voltage to the RF devices during transmission. Generally, vacuum tubes require bias-voltage modulation whereas solid-state amplifiers do not. The timing (when to switch and for how long) is governed by the waveform generator, often a part of the receiver equipment since the same waveform details are required there during reception. Gas-filled switch tubes (thyratrons, for example) were key in older designs, whereas solid-state circuitry is used increasingly in both modulators and power supplies today. *See* GAS TUBE.

The RF devices in many radar transmitters are vacuum tubes that amplify the signal in multiple stages (predriver, driver, and final stages, for example). Typical amplifiers are traveling-wave tubes, klystrons, and crossed-field amplifiers. The signal itself may be formed by a master oscillator or a set of oscillators constituting a frequency synthesizer, and then fed to the amplifiers. The modulator applies to these tubes the full cathode-anode voltage, or sometimes a lesser voltage to a modulating grid, just as the RF signal passes through, effecting the conversion of electrical power to RF power. *See* KLYSTRON; MICROWAVE TUBE; TRAVELING-WAVE TUBE; VACUUM TUBE.

The development of solid-state devices, generally transistors, as amplifiers at microwave frequencies has provided a design alternative to the use of large high-power vacuum tubes, appealing in several ways. Specifically, the power supply voltages required are far lower (tens of volts) than the biases required for the vacuum tubes, modulators of even that voltage are generally not required (transistors are designed to amplify only when fed an RF signal), and the probability of complete failure of the transmitter in the field is much reduced.

At frequencies from UHF through L-band, the relatively economic doped silicon compounds are commonly used as the semiconductor material, whereas the more costly gallium-arsenide devices provide higher power at frequencies above S-band. In either case, power capability lessens with increasing frequency, from perhaps 100 watts output per transistor to only 10 watts or less over the frequency ranges cited for each above. Therefore, thousands must often be used to achieve useful radar powers. Also, such amplifiers can handle an average power that is a considerable fraction of their peak power capability; hence, radar designers are led to consider modes and waveforms with duty factors of 10% and more (rather than the 0.1 to 1% duty usually associated with vacuum tubes).

In using such a large number of parallel amplifiers in high-power radar (small radars, of course, need far less power transmitted, far fewer transistor modules), two configurations prevail. In one, the power from large assemblies of such amplifiers is combined and the total power routed to the antenna in a rather conventional waveguide (typically) transmission line; in the other, such amplifiers are part of transmit-receive modules, one at each radiating element of an array-type antenna. This "active array" approach is mentioned again in the antenna discussion

below. *See* MICROWAVE SOLID-STATE DEVICES; TRAN-SISTOR.

Much simpler radar transmitters are also in wide use, in which a single oscillator (such as a magnetron) is used rather than an amplifier chain or large assembly of transistors. The modulator supplies a cathode-to-anode voltage pulse which causes an RF signal to be generated. *See* MAGNETRON.

The transmitter subsystem, particularly a subsystem that generates lethal high voltages, must include many safety features. Transmitters also require very involved cooling systems, since the overall power conversion efficiency is generally quite low and much damaging heat is generated.

Antenna. Of the four radar subsystems, the antenna is the most visible and often the most indicative of the nature of the radar. Antennas form the directive beam being used and position it in space by either physical movement of the antenna or by purely electronic means in an otherwise fixed structure. In the latter, an array of many discrete radiating elements with individual phase control is commonly used and is called a phased array. Some antennas use mechanical positioning in one dimension and electronic steering in another (for example, rotation in azimuth and electronic steering in elevation).

Antennas for mechanical scan are often of the feedhorn and reflector type; transmitter power is carried to the antenna by waveguide (usually rectangular metal piping) through a duplexer (if the antenna is used on both transmit and receive) and a rotary joint to a radiating horn at the focus of a reflector. The reflector (generally a parabolic section) forms the directive beam for both transmit and receive. The nature of the antenna's pattern, containing both the desired main beam and inescapable sidelobes in every direction, is determined by the way the feedhorn distributes power across the reflector's aperture; on receive, the same structure forms a pattern of directional sensitivity, the receive beam (and its sidelobes). Antennas designed for very low sidelobe levels (even 40 dB or more below the main beam) are a valuable part of modern radar. The design and manufacture of such low sidelobe antennas, requiring such care in the illuminating function across the aperture, is greatly assisted by computer-aided methods. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING; MICROWAVE OPTICS; WAVEGUIDE.

Optical properties are such that the beam formed has a width inversely proportional to the dimension of the illuminated aperture—the wider the antenna (in units of wavelengths), the narrower the beam in that dimension. A vertical fan beam is suitable for two-dimensional (2D) search radar making estimates of just the range and bearing of a target; such radars use, then, oblong reflector shapes (**Fig. 6**). A horizontally oriented fan beam might be used in a height-finding radar; its antenna, pointed to the azimuth of a designated target, nods vertically, scanning the narrow beam over the target to estimate its elevation angle.

Three-dimensional (3D) radars measure azimuth, elevation, and range as they rotate in azimuth, using

beams that are narrow in both angles (pencil beams). Often, a stack of pencil beams (requiring separate receivers) is used to measure elevation over a small angle for each transmission, then the stack is repositioned vertically for the next transmission. Such multibeam operation permits coverage and timing that is much more limited in single-beam 3D radars. Beams being narrow in both angles, these radars are characterized by rather square or round antenna shapes, equal in size in both dimensions.

Pencil beams are also used in precision tracking radars, those with dish antennas. Both the conical scanning technique and the multibeam monopulse angle measurement technique were described in the "Tracking" discussion earlier in this article.

Many radars use phased-array antennas, in which the transmitter power is divided among many radiating elements (possibly thousands) and in which the phase of each element is controlled by a computer (**Fig. 7**). In such a fixed-aperture antenna, the beam can be steered in only microseconds from one position to another many beam widths away. Consequently a wide variety of dwell routines can be implemented that are not at all constrained by mechanical rotation. Phased-array antennas are quite complex and expensive; they are used, therefore, only where great demand for very versatile operation exists. Phased arrays usually are designed to form pencil beams since 3D operation and even precision tracking are likely among the modes of operation intended.

Some modern radars use active arrays, in which the final stage of amplification of the transmitted signal takes place in transmit-receive modules located at every radiating element. The transmit-receive

Fig. 6. AN/SPS-49 search radar antenna, which forms a vertical fan beam for the two-dimensional L-band radar widely used in the U.S. Navy. (*Raytheon Co.*)
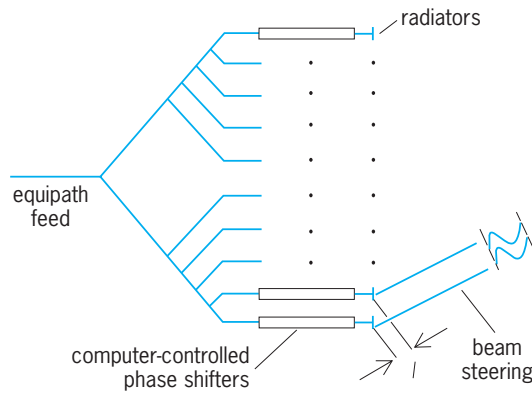
Fig. 7. Basic principle of electronic beam steering with a phased array. Equipath feed is for transmit or receive. All radiators radiate in essentially all forward directions. In beam steering, signals add in phase in the direction shown only if phase shifters are set to establish a phase increment element-to-element to compensate exactly for the path-length difference, l, shown.

modules may also contain duplexing circuitry and perform receive preamplification and conversion to intermediate frequency. If receive beam forming is to be done digitally in a computer (Fig. 3*c*) rather than in microwave circuits (Fig. 3*b*), the modules may contain the analog-to-digital converters as well. Still further development of self-adaptive processes, whereby the antenna's pattern can be made inversely proportional to directional sources of interference in its field of view, is well under way. *See* ANALOG-TO-DIGITAL CONVERTER.

Using rather narrow beams provides radars with useful angular (and lateral or cross-range) resolution. The ability to separate closely spaced targets is important in many radar applications, so the maximum practical antenna size (producing the narrowest practical beam width) often limits design. In airborne and spaceborne radars, fine cross-range resolution can be achieved by processing the echoes from the ground in a careful Doppler-sensitive way while the aircraft or satellite moves over a considerable path. The effect is somewhat like using an antenna as long as that segment of the flight path, perhaps a few hundred meters. This synthetic aperture radar (SAR) technique is particularly valuable in imaging the Earth's surface for both military and civil purposes. *See* SYNTHETIC APERTURE RADAR (SAR).

Receiver and signal processor. The signals collected by the antenna are amplified in some radars (those with RF preamplifiers), then converted to the IF for further amplification and processing. Several parallel channels of receivers may be necessary (for the several beams in a stacked-beam radar or for the angle error channels in a monopulse tracker, for example), each requiring all the processing described here.

Many radars require a long pulse for the requisite energy on each transmission, yet require the range resolution of a short pulse. The effect in range resolution of a very short pulse can be achieved by modulating the long pulse with either frequency or phase modulation so that it has the same bandwidth as the

very short pulse desired. Then, on receive, the waveform generator supplies a replica of the modulation on the transmitted pulse to the receiver so that a matched-filter action takes place in circuitry called the pulse compressor to produce a narrow output pulse.

In many radar applications it is desired to determine if returns are Doppler-shifted, that is, if certain returns are coming from radially moving targets rather than from stationary clutter. At the microwave frequencies, the Doppler shift is so slight, even for rather high-speed targets, that it cannot be detected on just one short pulse transmission. Instead, multiple pulses must be used over an interval of time that compares to a cycle (at least) of the Doppler shift. A radar equipped to compare the phase of each echo with the phase of the transmission, and then to see if a consistent phase progression exists from pulse to pulse (indicating that the range to the target is changing) is called a "coherent" radar. A simple way of processing such echoes is to compare the phase of each pulse to that of the previous one; this is what a two-pulse canceller circuit does in a moving target indication (MTI) radar. More elaborate cancellers may compare pulses in sets of three or more, with a better shaping of the Doppler response function of the process. More valuable (often) Doppler filtering is achieved by processing the phase data from all the pulses of a single coherent dwell, perhaps scores of pulses. The process involves spectral analysis of that set, wherein as many filters can be formed as there are samples (pulses) and the resolution of the filters (their width and separation in Doppler frequency, roughly speaking) is equal to the reciprocal of the time spanned by the coherent dwell (perhaps several milliseconds). *See* MOVING-TARGET INDICATION.

The fine quantization and processing speeds of modern computers permit these coherent processes to be performed in computerlike circuitry after an appropriate digitizing of the IF signal (as in paired synchronous detection giving two scalar values of the quadrature components of the signal). Most valuable is the process of digital Doppler filtering using discrete and fast Fourier transform methods. Such computer use has also prompted development of adaptive processing wherein heavy clutter that is not uniformly distributed in Doppler (as experienced in airborne radar flying over certain terrain features) can be more effectively countered in the Doppler filtering. *See* FOURIER SERIES AND TRANSFORMS.

Pulse compression and Doppler processing are both coherent processes since complex (amplitude and phase) information is processed. Further steps in the signal processing might now be noncoherent; for example, angle estimation in monopulse radar often uses just the amplitudes of signals in the parallel receiver channels involved. Following Doppler filtering, the amplitude of the signal accumulated in each filter becomes the observable upon which a detection decision is to be made. In older radars, this output might be the video voltage from a radio detector circuit had the filtering been in analog circuitry at the intermediate frequency. In modern
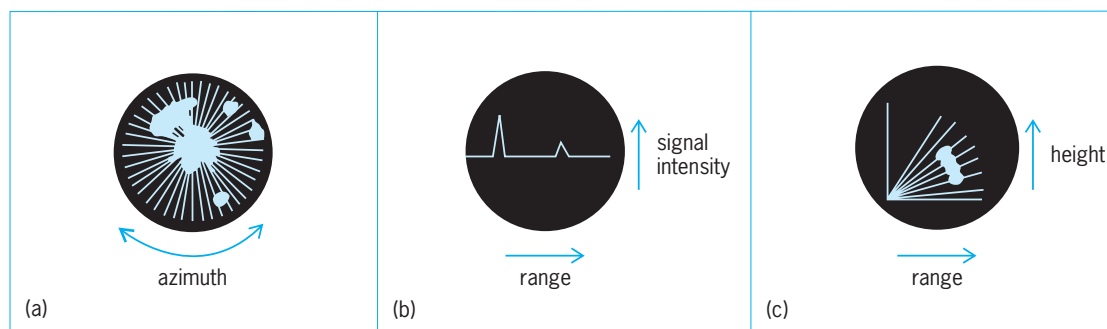
Fig. 8. Cathode-ray-tube radar display formats. (*a*) Type P (plan position indication, or PPI). Range is measured radially from center. (*b*) Type A. (*c*) Range-height indication (RHI).

radar, the digital processing produces numeric values for that output; it is upon these values that the interpretation of the output (to generate contacts, in the search-radar context) is effected in the control and interface apparatus. *See* SIGNAL PROCESSING.

Control and interface apparatus. In simple radars, the operator may observe the output of an elementary video-producing detector on one of several formats on a cathode-ray-tube display. The most common format is the PPI (plan position indication), in which a radial trace (the path of the cathode-ray-tube electron beam exciting the phosphor) is synchronized to the pulse repetition frequency, resulting in radar range being displayed radially from the center (**Fig. 8***a*), and in which this trace is moved around the display, linked electronically to the antenna rotation. The detector output is applied to the intensity grid of the cathode-ray tube so that a bright spot results when a strong video voltage is produced by the detector. The display circuitry may also superimpose upon the radar video certain calibration pulses (range marks and bearing lines, for example), maps of the local terrain, or other symbols that are of importance in the entire system of which the radar is a part.

Other analog displays are illustrated (Fig. 8*b,c*) and, together with the PPI format, remain useful display formats for more automated radars. However, such display is generally of computer-generated symbols and data for operator awareness to manage the system, rather than of receiver video for operator performance of the detection process itself.

In the earlier discussion of the "Detection process," the fundamentals of constant false alarm rate (CFAR) automatic detection were described. In the "Tracking" discussion, the association and track filtering processes were described. The computers and attendant displays of this subsystem perform those functions.

Another important part of the control and interface apparatus, particularly in modern multifunction radars, is the mode scheduler. The outputs of the detection and tracking functions combined with external data provided to the radar (mission priorities, dimensions of function-critical regions in the radar surveillance space, information from other sensors of the network, and so on) must feed into a control logic to produce a list of what next should be done. That list must be prioritized and fed to a con-

trol apparatus to operate the radar itself. Further, this scheduling must be done within the radar's time and energy constraints. Often such control is planned in short time "modules," planning periods of perhaps a few tens of milliseconds, enough time for a dozen or so radar dwells. The response time from a radar result being interpreted and a new action taken might need to be a small fraction (a tenth) of a second—the most critical loop-closure time in the various modes of the radar will bear heavily upon the dimensioning of the time modules and the architecture of the scheduler.

Finally, the monitoring of the state of the radar itself is an important function of this subsystem. Modern radars have, of necessity, much built-in test equipment (BITE). There may be a resource console where data are collected from various monitoring points throughout the radar equipment. In military radars, this function is further burdened with battle damage assessment and the direction of compensating measures if further operation is possible. This function may include conducting readiness tests from time to time. Internal test target stimuli are affected and results analyzed for assuring acceptable performance. Physical maintenance procedures can be based on these results.

Less sophisticated testing of radars (and certainly of those with less BITE than described above) is still carried out by trained technicians in the field. A rather fundamental measure of a radar's state is called the performance factor, the ratio of the power output of a radar's transmitter and the minimum discernable signal (MDS) of its receiver. For example, a radar with transmitter power (pulse power) of 1 megawatt and an MDS of 10^{-14} watt (an MDS of -140 dBw or -110 dBm -110 decibels below a milliwatt) has a performance factor of 200 dB (that is, of 10^{20}). A convenient way to test for this value is built into many radars; separate test equipment can be used in others.

The "echo box" is familiar to all technicians who have worked in the field with radar. It consists of a resonant cavity into which can be fed a set fraction of the transmitter's power (by a directional coupler built into the radar for this purpose). The well-calibrated and decaying ringing of the resonating pulse is fed into the receiver, where the receiver's response to it (at an appropriate test point) can be

viewed on an oscilloscope. The length of time after the pulse that the ringing remains visible above the receiver noise background is a direct measure of the radar's performance factor. Some echo boxes are equipped with precise tuning and a power meter inside the cavity so that a more careful spectral analysis of the transmitter's performance can be made across the radar's operating band. Knowing the power level also permits the receiver's MDS to be determined. These performance values differ day-to-day in actual radar operation, and this relatively simple diagnostic tool is helpful in keeping the radar working well.   Robert T. Hill

Bibliography. D. K., Barton, *Modern Radar System Analysis*, Artech House, Norwood, MA, 1988; L. V. Blake, *Radar Range Performance Analysis*, Artech House, Norwood, MA, 1986, reprint 1991; R. Buderi, *The Invention that Changed the World*, Simon and Schuster, New York, NY, 1996; F. E. Nathanson, *Radar Design Principles*, SciTech Publishing, Mendham, NJ, 1999; S. Sabatini and M. Tarantino, *Multifunction Array Radar*, Artech House, Norwood, MA, 1994; M. I. Skolnik, *Introduction to Radar Systems*, 3d ed., McGraw-Hill, New York, NY, 2001; M. I. Skolnik (ed.), *Radar Handbook*, 2d ed., McGraw-Hill, New York, NY, 1990; W. D. Wirth, *Radar Techniques Using Array Antennas*, Institution of Electrical Engineers, London, 2001.

Radar-absorbing materials

Materials that are designed to reduce the reflection of electromagnetic radiation from a conducting surface in the frequency range from approximately 100 MHz to 100 GHz. The level of reduction that is achieved varies from a few decibels to greater than 50 dB, reducing the reflected energy by as much as 99.999%.

Theory. The reflection of electromagnetic radiation from the surface of any material can be calculated from Maxwell's equations if the electromagnetic properties are known. The relevant properties are the magnetic permeability μ_m and the dielectric permitivitty ε_m. Both μ_m and ε_m may be expressed as complex numbers, in which case the imaginary part represents the magnetic loss or dielectric loss, respectively. These parameters in turn define the characteristic impedance Z_m, given by Eq. (1), and the propagation coefficient γ_m, given by Eq. (2) [where

$$Z_m = \sqrt{\frac{\mu_m}{\varepsilon_m}} \qquad (1)$$

$$\gamma_m = \frac{2\pi_c}{\lambda}\sqrt{\mu_m\,\varepsilon_m} \qquad (2)$$

λ is the wavelength of the radiation and c is the speed of light], which determine the reflection at the material interface and the attenuation of the radiation in the material, respectively. In the case of radar-absorbing materials, the objective is to minimize the reflection by matching the input impedance Z_i of the material layer to the characteristic impedance Z_0 of free space. *See* ELECTRICAL IMPEDANCE; ELEC-TROMAGNETIC RADIATION; MAGNETISM; MAXWELL'S EQUATIONS; PERMITTIVITY.

Equation (3) gives the input impedance of a layer

$$Z_i = Z_m \tanh\left(\gamma_m d\right) \qquad (3)$$

of material of thickness d directly adjacent to a conducting surface. The energy reflection coefficient is given by expression (4). Zero reflection will occur

$$\left(\frac{1 - Z_i/Z_0}{1 + Z_i/Z_0}\right)^2 \qquad (4)$$

in the case where Z_i is equal to Z_0.

In the simple case where the attenuation in the material is such that there is no reflected energy from the conducting surface, the hyperbolic tangent (tanh) function in Eq. (3) tends to unity, Z_i equals Z_m, and the only reflection is at the free-space/material interface. Zero reflection is obtained when the characteristic impedance of the material and free space are identical, which also implies that Eq. (5) is sat-

$$\frac{\mu_m}{\varepsilon_m} = \frac{\mu_0}{\varepsilon_0} \qquad (5)$$

isfied. This condition must be met for μ_m and ε_m as complex numbers. Since either μ_m or ε_m must contain an imaginary component in order to attenuate the radiation, it necessarily follows that in this case both μ_m and ε_m must be complex numbers with both the real parts and imaginary parts in the same ratio as the ratio of μ_0 and ε_0. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION.

Design methods. In theory, this simple solution offers the possibility of a perfect absorber. In practice, materials have not been found which will give such a good impedance match over an appreciable frequency range. It is therefore necessary to adopt specific design methods to make practical absorbing materials.

Two methods have been widely adopted in order to produce such absorbers. The first is to avoid a discrete change of impedance at the material surface by gradually varying the impedance. The removal of the discontinuity at the surface allows the microwave energy to be transmitted into the absorbing medium without reflection. This transition from the impedance of free space to that of the bulk material is commonly achieved by a geometric profile. The carbon-loaded foam pyramids used as the lining of anechoic chambers are typical of this type of absorber. To produce such absorbers, it is necessary in practice to taper the material over distances which are large compared with the wavelength of the frequencies to be absorbed. Therefore, practical absorbers of this type giving greater than 20 dB absorption vary in thickness from about 0.8 in. (2 cm) at 10 GHz and above to 6 ft (2 m) at 100 MHz and above. **Figure 1** shows the approximate level of absorption at normal incidence as a function of the thickness of the material in wavelengths of the incident radiation. The absorber performance improves with increasing thickness until the point is reached
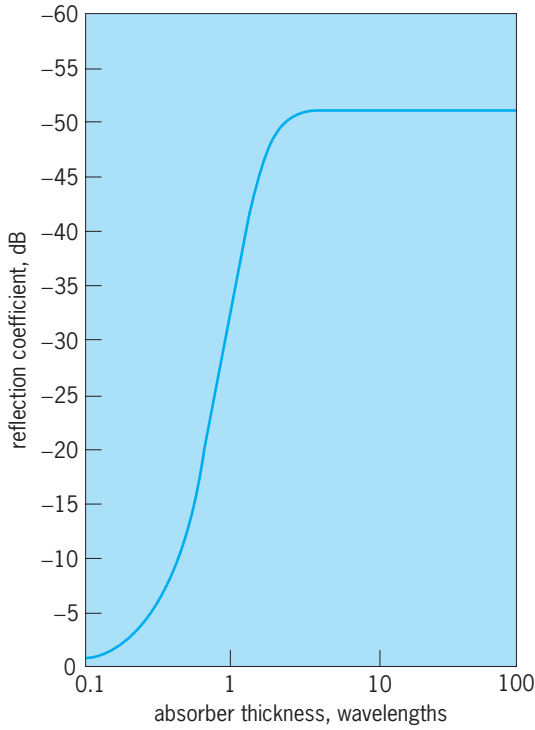
Fig. 1. Absorption of a thick, profiled lossy layer at normal incidence as a function of the thickness of the material in wavelengths of the incident radiation.

frequency is shown when the electric field (E) vector is parallel and perpendicular to the plane of incidence for a range of angles. *See* POLARIZED LIGHT.

It is possible to tune these materials to any desired center frequency over most of the range from 100 MHz to 100 GHz. When the absorber is optimized for an angle other than normal, it will generally be for only one polarization. While this type of material intrinsically provides a relatively narrow bandwidth frequency performance, it is also possible to broaden the effective bandwidth by various techniques such as multiple-layer absorbers. With two layers of material it is possible to tune one absorber to two different frequencies. By placing these two frequencies appropriately, for example, within one
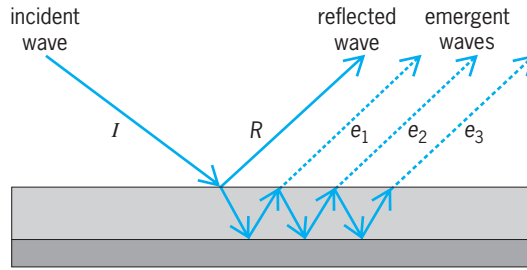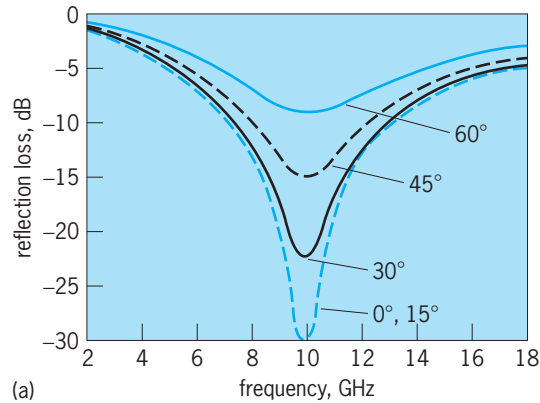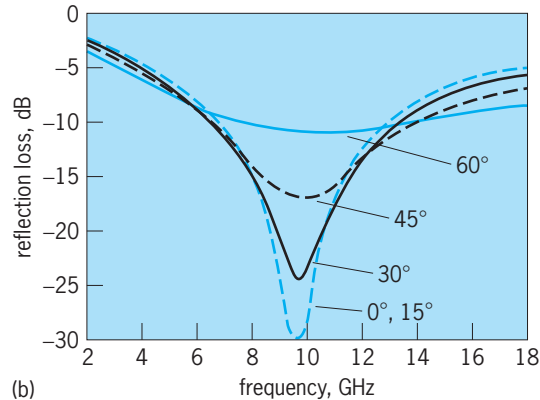

Fig. 2. Mode of operation of a lossy quarter-wave layer. There is a mutual cancellation of the reflected wave and the emergent waves e_1, e_2, e_3, and so forth.

where all of the energy that enters the material is absorbed and only the front-face reflection is left. In the example given, the front face reflection is -50 dB relative to a perfect reflector, which is a typical figure for anechoic chamber material. While this type of absorber is capable of producing a very high degree of absorption over a broad bandwidth, it is at the same time a relatively thick material. *See* ANECHOIC CHAMBER.

The second method of absorber design produces much thinner absorbing layers which are capable of producing good absorption (≥ 25 dB) with restricted bandwidths. These materials achieve the absorption by a combination of attenuation within the material and destructive interference at the interface. The electromagnetic properties and the thickness of the layer are such that the initial reflected wave and the sum of the emergent rays (e_1, e_2, e_3, and so forth) resulting from the multiple reflections within the material are equal in magnitude and opposite in phase. The thickness of the layer is close to a quarter-wavelength at the frequency of operation, giving a 180° phase difference between the interface reflection and the emergent waves (**Fig. 2**). *See* INTERFERENCE OF WAVES.

The characteristic impedance and the propagation coefficient of the materials change with the angle of incidence and polarization of the radiation on the material surface. As a result, the input impedance of the layer and hence the level of absorption are angle- and polarization-dependent. The performance of a typical quarter-wave absorber designed for normal incidence is shown in **Fig. 3**. The absorption against

(a)

(b)

Fig. 3. Performance of a lossy quarter-wave layer designed for optimum performance when the incident radiation is normal to the surface as a function of frequency, polarization, and angle of incidence. (*a*) Perpendicular polarization. (*b*) Parallel polarization. Angles of incidence are indicated.

octave of each other, a broadband absorber is obtained in practice.

Absorbers based on destructive interference can also be made by placing a thin resistive layer one-quarter of a wavelength from a conducting surface. In this case, we consider the system in terms of admittance, which is the reciprocal of impedance. The admittance of free space is 1/377 siemens, and that of a quarter-wavelength space in front of a reflector is zero. Therefore, a thin layer of resistance 377 ohms spaced one-quarter wavelength from a reflector will match the admittance of free space, and there will be no reflection. The absorption characteristics with frequency are similar to those shown in Fig. 3.

Uses. Microwave-absorbing materials are widely used both within the electronics industry and for defense purposes. Their uses can be classified into three major areas: (1) for test purposes so that accurate measurements can be made on microwave equipment unaffected by spurious reflected signals (the most common example is the anechoic chamber); (2) to improve the performance of any practical microwave system by removing unwanted reflections which can occur if there is any conducting material in the radiation path; and (3) to camouflage a military target by reducing the reflected radar signal. *See* ELECTRONIC WARFARE; MICROWAVE; MICROWAVE MEASUREMENTS; REFLECTION OF ELECTROMAGNETIC RADIATION.   S. B. Morris

Bibliography. E. F. Knott, J. F. Shaeffer, and M. T. Tuley, *Radar Cross Section*, 2d ed., 1993; A. R. Von Hippel, *Dielectrics and Waves*, 2 vols., 1955, reprint 1995.

Radar astronomy

A powerful astronomical technique that uses radar echoes to furnish otherwise unavailable information about bodies in the solar system. By comparing a radar echo to the transmitted signal, information can be obtained about the target's size, shape, topography, surface bulk density, spin vector, and orbital elements, as well as the presence of satellites and in certain situations the target's mass and density. While other astronomical techniques rely on passive measurement of reflected sunlight or naturally emitted radiation, the illumination used in radar astronomy is a coherent signal whose polarization and time modulation or frequency modulation are tailored to meet specific scientific objectives. Through measurements of the distribution of echo power in time delay or Doppler frequency, radar achieves spatial resolution of a planetary target despite the fact that the radar beam is typically much larger than the angular extent of the target. This capability is particularly valuable for asteroids and planetary satellites, which appear as unresolved point sources through optical telescopes. Moreover, the centimeter-to-meter wavelengths used in radar astronomy readily penetrate cometary comas and the optically opaque clouds that conceal Venus and Titan, and also permit determination of near-surface roughness (abundance of wavelength-scale rocks), bulk density, and metal concentration in planetary regoliths. *See* ASTEROID; SATELLITE (ASTRONOMY); SATURN; VENUS.

Telescopes. A radar telescope is essentially a radio telescope equipped with a high-power transmitter (a klystron vacuum-tube amplifier) and specialized instrumentation that links the transmitter, low-noise receiver, high-speed data-acquisition computer, and antenna together in an integrated radar system. Planetary radars, which must detect echoes from targets at distances from about 10^6 km (10^6 mi) for closely approaching asteroids and comets to more than 10^9 km (nearly 10^9 mi) for Saturn's rings and satellites, are the largest and most sensitive radars on Earth. *See* KLYSTRON; RADIO TELESCOPE.

The two active planetary radar facilities are the Arecibo (radar wavelengths of 13 and 70 cm) and Goldstone (3.5 and 13 cm) instruments; for each, the shorter wavelength provides the greater sensitivity. The Arecibo radio-radar telescope in Puerto Rico consists of a 305-m-diameter (1000-ft) fixed reflector, the surface of which is a section of a 265-m-radius (870-ft) sphere. Movable feeds suspended from a triangular platform 130 m (427 ft) above the reflector can be aimed toward various positions on the reflector, enabling the telescope, located at 18°N latitude, to point up to 20° from the zenith. The Goldstone radar in California is part of the National Aeronautics and Space Administration's Deep Space Network. The Goldstone main antenna, DSS-14, is a steerable, 70-m (230-ft) parabolic reflector with horn feeds. Arecibo is the more sensitive of the two instruments, with twice the range of Goldstone, but Goldstone can track targets continuously for much longer periods and has access to the whole sky north of $-40°$ declination. Bistatic (two-station) experiments employing transmission from DSS-14 and reception of echoes at the 27-antenna Very Large Array (VLA) in New Mexico have synthesized a beamwidth as small as 0.24 second of arc, versus 2 minutes of arc for single-dish observations with Arecibo or Goldstone. Bistatic experiments using DSS-14 transmissions and reception of echoes at DSS-13, a 34-m (112-ft) antenna 22 km (14 mi) away, have been conducted on several very close targets. Bistatic observations between Arecibo and Goldstone, or using transmission from Arecibo or Goldstone and reception at the 100-m (328-ft) Greenbank Telescope (GBT) in West Virginia, have also proven very powerful.

Techniques. A typical transmit-receive cycle consists of signal transmission for a duration close to the round-trip light time between the radar and the target, that is, until the first echoes are about to return, followed by reception of echoes for a similar duration. The target's apparent radial motion introduces a continuously changing Doppler shift into the echoes. This shift is removed by continuously tuning the instrument according to an ephemeris based on an orbit calculated from optical (and perhaps radar) astrometric observations. In continuous-wave observations, a nearly monochromatic signal is transmitted and the distribution of echo power is measured as a
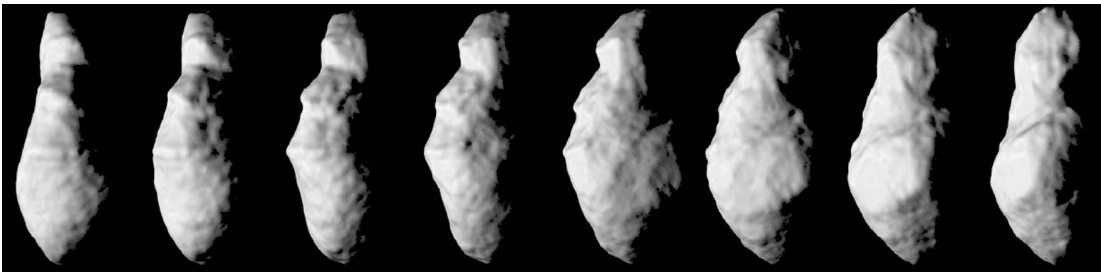
Fig. 1. A model of the shape of near-Earth asteroid 4179 Toutatis based on radar delay-Doppler images. The 4.5-km-long (2.8-mi) object has complex linear features as well as circular craterlike structures down to the ~34-m (112-ft) resolution limit. The noncraterlike features may arise from complex interior configurations involving monolithic fragments with various sizes and shapes, presumably due to collisions in various energy regimes. (*R. S. Hudson, S. J. Ostro, and D. J. Scheeres, High-resolution model of asteroid 4179 Toutatis, Icarus, 161:346–355, 2003*)

function of frequency. The resultant echo spectrum can be thought of as a one-dimensional image, or a brightness scan across the target through a slit parallel to the target's apparent spin vector. *See* DOPPLER EFFECT.

Range resolution can be obtained by using a coherent pulsed continuous waveform, but in practice, because of engineering considerations associated with klystrons, a pulsed waveform is simulated by encoding a continuous-wave signal with a preset sequence of $180°$ phase reversals. Fourier transformation of time samples taken at the same position within each of many successive range profiles yields the echo power spectrum for the corresponding range cell on the target. Spectral analysis of echoes from multiple range cells yields a delay-Doppler image. If the target's shape is known, it is a straightforward matter to convert that image into a radar brightness map. Heterogeneities in brightness can be introduced by variations in large-scale surface tilts, small-scale roughness, or surface bulk density. By measuring the echo's polarization state, the nature of the scattering process and the severity of wavelength-scale roughness near the surface can be deduced, thereby refining the interpretation of reflectivity measurements. If the shape is unknown, as with asteroids and comets, it can be estimated from delay-Doppler images of the target in a variety of orientations (**Fig. 1**). *See* FOURIER SERIES AND TRANSFORMS; POLARIMETRY.

Observational results. Radar-detected planetary targets include the Moon, Mercury, Mars, Venus, Phobos, the Galilean satellites, Saturn's rings, Titan and six other Saturnian satellites, nine comets, 95 main-belt asteroids, and 169 near-Earth asteroids (including 16 binaries and 95 potentially hazardous objects). The radar signatures of these bodies are extraordinarily diverse. For example, for each of the two most basic radar properties (albedo and circular polarization ratio), estimated values span two orders of magnitude. *See* ALBEDO; BINARY ASTEROID; COMET; JUPITER; MARS; MERCURY (PLANET).

Radar observations have revealed a metallic near-Earth asteroid, the dumbbell shape and metallic composition of a large main-belt asteroid, and the presence of large-particle swarms (with particles at least 1 cm or 0.4 in. in diameter) around many comets (including Halley's Comet). Radar has discovered the existence of contact binaries, binary systems, non-principal-axis spin states, and near-Earth-asteroid rotation periods ranging from minutes to weeks. Radar also has achieved the first detection of the nongravitational, thermal-recoil "Yarkovsky" acceleration of an asteroid, using the measurement to estimate the asteroid's mass. Radar's unique capabilities for trajectory refinement and physical characterization give it a natural role in predicting and preventing collisions with small bodies.

Other radar experiments have disclosed the extraordinary diversity of Mars's surface at decimeter-to-meter scales and have produced startling evidence for radar-bright, probably icy, polar caps on Mercury. Thirty years after the radar discovery of Venus's slow retrograde rotation, a radar instrument on the Venus-orbiting *Magellan* spacecraft mapped almost all of that planet's surface with striking geologic clarity. The first four-station radar interferometry assisted selection of the landing sites for the *Mars Exploration Rovers*, and a novel two-station radar speckle displacement technique has produced ultraprecise measurements of Mercury's spin state that indicate a liquid core.

Titan is the primary focus of the *Cassini* spacecraft, which went into orbit about Saturn in 2004. *Cassini*'s radar instrument, a 13.8-GHz (2.2-cm) synthetic aperture radar (SAR) imager, altimeter, scatterometer, and radiometer, is to operate during around half of the 44 Titan flybys, covering about a fifth of the surface with imaging resolution of 2 km (1.2 mi) or finer. Scatterometry observations
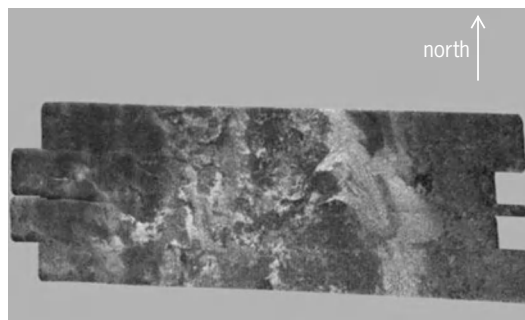

Fig. 2. A 150 × 250-km (90 × 150-mi) portion of *Cassini*'s first synthetic aperture radar (SAR) image strip on Titan, centered at 50°N, 82°W. The smallest details seen on the image are about 300 m (1000 ft) across. (*NASA/JPL*)

are to cover most of the surface, albeit at resolution no finer than tens of kilometers, and a limited number of short altimetry tracks are to give regional topographic information. *See* SYNTHETIC APERTURE RADAR (SAR).

The first Titan radar images (**Fig. 2**) reveal a complex surface with low relief and a wide variety of surface features and terrains, including flows that might be cryovolcanic and parallel linear features that might be eolian or tectonic. Radar-bright sinuous channels and associated fanlike structures suggest the presence of transported material with different radar properties. Dark patches are tentatively identified as low-dielectric materials such as hydrocarbon lakes, organic sludge, water ices, or ammonia hydrates. The small number of features unambiguously identified as craters strongly suggests a young, geologically active surface, possibly modified as well by deposition of atmospheric chemical products. *See* PLANET; RADAR; SPACE PROBE.　　　Steven J. Ostro

Bibliography. G. J. Black, D. B. Campbell, and P. D. Nicholson, Icy Galilean satellites: Modeling radar reflectivities as a coherent backscatter effect, *Icarus*, 151:167–180, 2001; J. K. Harmon, P. J. Perillat, and M. A. Slade, High-resolution radar imaging of Mercury's north pole, *Icarus*, 149:1–15, 2001; J. L. Margot et al., Topography of the lunar poles from radar interferometry: A survey of cold trap locations, *Science*, 284:1658–1660, 1999; S. J. Ostro, Planetary radar astronomy, *Rev. Mod. Phys.*, 65:1235–1279, 1993; S. J. Ostro et al., Asteroid radar astronomy, in *Asteroids III* (W. Bottke et al., eds.), pp. 151–168, University of Arizona Press, Tucson, 2002; G. H. Pettengill et al., *Magellan*: Radar performance and data products, *Science*, 252:260–265, 1991.

Radar meteorology

The application of radar to the study of the atmosphere and to the observation and forecasting of weather. Meteorological radars transmit electromagnetic waves at microwave and radio-wave frequencies. Water and ice particles, inhomogeneities in the radio refractive index associated with atmospheric turbulence and humidity variations, insects, and birds scatter radar waves. The backscattered energy received at the radar constitutes the returned signal. Meteorologists use the amplitude, phase, and polarization state of the backscattered energy to deduce the location and intensity of precipitation, the wind speed along the direction of the radar beam, and precipitation type (for example, rain or hail). *See* METEOROLOGICAL RADAR; RADAR; WEATHER FORECASTING AND PREDICTION.

Echoes from atmosphere. Meteorological radars normally transmit short pulses of energy. The signal received is the combination of electromagnetic waves backscattered toward the radar by each scattering element within the volume illuminated by a pulse. If the scattering elements fill the beam, which is normally true for meteorological applications, this volume is determined from the angular beamwidth

θ and the pulse duration τ, and is given by Eq. (1),

$$V = \frac{\pi}{4}(R\theta)^2 \frac{c\tau}{2} \qquad (1)$$

where c is the speed of propagation of the waves and R is the distance from the radar to the pulse volume.

If the beam width is $1°$, the pulse duration 1 microsecond, and the range 50 km (30 mi), the scattering volume equals about 9×10^7 m^3 (3.2×10^9 ft^3). For precipitation, the volume may contain more than 10^{11} raindrops or ice particles. The contribution of each scattering element in the volume adds in phase to create the returned signal. The returned signal fluctuates from one pulse to the next because the scattering elements move relative to one another. For this reason, the returned signals from many pulses are averaged to determine the average received power, from which the intensity of the precipitation can be estimated. The pulse-to-pulse phase change, due to the average motion of the scattering elements along the radar beam, is also used by Doppler radars to determine air motion toward or away from the radar. *See* DOPPLER RADAR.

The radar range equation for distributed targets such as raindrops may be written as Eq. (2), where

$$P_r = \frac{P_t G^2 \lambda^2}{(4\pi)^3 R^4} V\eta \qquad (2)$$

P_r is the average received power, P_t the transmitted power, G the antenna gain, λ the wavelength, R the range, V the scattering volume, and η the reflectivity, which is assumed to be constant over the scattering volume. The radar cross section σ of a spherical water or ice particle whose diameter D is small compared to the wavelength λ is given by the Rayleigh scattering law, Eq. (3), where $|K|^2$, a dimension-

$$\sigma = \frac{\pi^5}{\lambda^4}|K|^2 D^6 \qquad (3)$$

less factor that depends on the dielectric properties of the particle, is approximately equal to 0.93 for water and 0.21 for ice. The radar reflectivity of clouds and precipitation is obtained by summing the cross sections of all the particles in the scattering volume and is written as Eq. (4).

$$\eta = \frac{\pi^5}{\lambda^4}|K|^2 Z \qquad (4)$$

The radar reflectivity factor Z is defined by Eq. (5), where the summation is over the scattering

$$Z = \frac{\sum D^6}{V} \qquad (5)$$

volume. It is customary to use cubic meters as the unit for volume and to measure the particle diameters in millimeters, so that Z has conventional units of mm^6/m^3. The radar reflectivity factor varies from about 10^{-5} mm^6/m^3 to 10 mm^6/m^3 in nonprecipitating clouds, upward to about 10^6 mm^6/m^3 in rain, and as high as 10^7 mm^6/m^3 in large hail. In a mixture of rain and clouds, the raindrops dominate the returned

signal because of the sixth-power weighting on diameter. This is true even though the number of raindrops in a cubic meter—typically about 10^3—is perhaps 10^5 fewer than the number of cloud droplets. The radar reflectivity factor Z is of significant meteorological importance because of its relationship to the raindrop size distribution within the scattering volume. In meteorological applications, the averaged returned power is measured and converted to values of Z using Eqs. (1)–(4). Since Z varies across many orders of magnitude, a logarithmic scale, defined by Eq. (6), is used to display the radar reflectivity factor.

$$\text{dBZ} = 10 \log_{10} \left(\frac{Z}{1 \text{ mm}^6/\text{m}^3} \right) \qquad (6)$$

Typical radar depictions of weather systems that appear in the media (**Colorplate 1**) display the radar reflectivity factor in logarithmic units. Displays of the radar reflectivity factor overlaid on maps quickly allow the meteorologist to determine the location and intensity of precipitation. The term "radar reflectivity factor" is often shortened to "reflectivity" by meteorologists, although radar specialists reserve this terminology for η.

Raindrops, ice particles, and other objects are not the only source of radar echoes. Reflections from the optically transparent atmosphere can be observed by sensitive radars. These echoes are associated with spatial fluctuations in the atmospheric refractive index created by turbulence in the presence of temperature and humidity gradients. The clear-air radar reflectivity produced by turbulent mixing is given by Eq. (7), where C_n^2 is a quantity that character-

$$\eta = \frac{5}{6} \pi^{4/3} 4^{-5/3} \lambda^{-1/3} C_n^2 = 0.38 \lambda^{-1/3} C_n^2 \quad (7)$$

izes the strength of the refractive index irregularities. The different dependence on wavelength in Eq. (4) and Eq. (7) shows that the radar sensitivity to turbulent scattering decreases at a much slower rate with increasing wavelength than does the sensitivity to scattering by small particles. The different wavelength dependence is the reason why radars with rather long wavelengths are favored for clear-air studies. The clear-air signal is important for retrieving air motions in some Doppler radars.

The Doppler shift refers to a small change in frequency of the returned signal associated with motion of scattering elements toward or away from the radar. Scattering elements move with the ambient wind. In meteorological applications, measurement of this small frequency shift provides an estimate of the along-beam, or radial, wind component, from which information about winds and other motions of interest can be extracted. To obtain Doppler information from a radar echo, the transmitted signal must have suitable characteristics and special signal processing must occur in the receiver. In particular, the pulse-to-pulse change in the phase of the returned signal must be retained and extracted. Ambiguity problems exist in Doppler radar systems because the same series of sampled phase values can be produced by

more than one target radial speed. The maximum unambiguous radial speed measurable by a Doppler radar, termed the Nyquist velocity v_n, is given by Eq. (8), where F is the pulse repetition frequency.

$$v_n = \frac{\lambda F}{4} \qquad (8)$$

Radial wind speeds that exceed the Nyquist velocity are folded back into the observable range. Algorithms exist to "unfold" the radial wind measurements to determine the true radial winds for most meteorological situations.

Conventional radars designed for meteorological measurements radiate and receive electromagnetic waves with a fixed polarization, a single fixed orientation (horizontal) of the electric field. Special polarization diversity radars allow variation of the polarization state (horizontal and vertical) of the transmitted or received signal. Polarization measurements are sensitive to particle orientation. As large raindrops fall, they have a tendency to flatten into an oval shape. Hail tumbles as it falls and generally has no preferred orientation. Ice crystals may or may not have preferred orientation, depending on their shape. The radar cross sections of these precipitation particles depend on the orientation of the impinging electric field. Polarization measurements therefore have the potential to discriminate hail from rain, and to identify predominant types of ice particles in clouds. Radars have been developed that switch polarization between horizontal and vertical on successive pulses, or transmit and receive circular and elliptically polarized waves. Researchers are also exploring methods for using polarization techniques to estimate rainfall. *See* PRECIPITATION (METEOROLOGY).

Applications. Radar meteorologists use analysis and animation techniques to track and predict the movement of storms, estimate winds and precipitation, and identify dangerous weather phenomena, such as tornadoes and hail.

Storm development and motion. Thunderstorms appear on radar as cores of high reflectivity, while widespread precipitation appears as larger areas of weaker reflectivity. Widespread precipitation areas, viewed on a large scale, often are organized in bands with widths ranging from tens to hundreds of kilometers. Meteorologists use displays of reflectivity to observe storms developing and to track storms through their life cycles. Meteorologists use individual radars to determine the location, intensity, and dimensions of individual storms and storm complexes. Data composites from radar networks show how precipitation regions are organized around their parent weather systems, which may extend over thousands of kilometers. From animations of radar images, meteorologists can determine whether storms are intensifying or weakening, and the speed and direction of their motion, enabling the prediction of storm passage over critical areas up to several hours in advance. *See* STORM; STORM DETECTION.

Severe thunderstorm detection. Severe thunderstorms may produce large hail, damaging straight-line winds,

or tornadoes. Radars, particularly Doppler, allow meteorologists to issue timely warnings as destructive thunderstorms approach populated regions. A severe thunderstorm will often develop a hooklike appendage in the reflectivity field. If a tornado develops, it will often appear near the center of the hook. The radial velocity field from a Doppler radar can provide clear indications of storm rotation. Rotation appears as a tight couplet of adjacent strong inbound and outbound radial motions (**Colorplate 2**). Animations of the reflectivity and radial velocity fields are used to track hook positions and radial velocity couplets, permitting meteorologists to issue specific warnings concerning the arrival of dangerous conditions. *See* THUNDERSTORM.

The reflectivity within a storm is another indicator of its severity and potential destructiveness. Values of reflectivity exceeding 55 dBZ are often associated with hail. Large hail, which can reach diameters of 1–6 cm (0.4–2.5 in.), appears on radar with reflectivity approaching 70 dBZ. Radars with polarization switching capability can detect hail by comparing the reflectivity measured at horizontal (Z_H) and vertical (Z_V) polarization. When Z_H is large, and the ratio of $Z_H/Z_V \cong 1$, the precipitation particles are approximately circular and are likely to be hail. For large values of the ratio, the particles are oval shape and are likely to be rain.

Precipitation measurement and hydrology. The value of the radar reflectivity factor Z is a general indicator of precipitation intensity; however, no exact relationship between Z and precipitation rate P exists. Many studies have found that P and Z are approximately related by Eq. (9). The coefficient a and the

$$Z = aP^b \qquad (9)$$

exponent b take on different values depending on the precipitation type. For moderate, widespread rain, for example, a is typically 200 and b is 1.6 if P is measured in mm/h and Z is in mm^6/m^3. Two methods have been used for establishing relationships of this kind. In one method, a radar is used to take frequent measurements of the reflectivity just above one or more rain gages located on the ground. Direct comparisons can then be made between Z and P to determine the most suitable values of a and b in Eq. (9). The second method is indirect, based on measurements of the sizes and concentrations of raindrops in many rainfall samples. For a given sample, both Z and P can be calculated. Many samples corresponding to rainfalls of different intensities yield a number of combinations of Z and P from which a best-fit relation of the form of Eq. (9) can be derived. Owing to uncertainties in the radar calibration, variability in the empirical quantities a and b, and other sources of error, radar estimates of the short-term precipitation rate at a point can deviate by more than a factor of 2 from gage measurements. Some of these errors are random, so when averaged over larger areas and longer times, radar estimates of the total accumulated rainfall are more accurate.

Hydrologists use precipitation estimates from radar to estimate the total precipitation over a watershed and stream runoff. Because radar-based precipitation estimates are almost instantaneously available to hydrologists, radar has become an effective tool for issuing flash flood warnings. *See* HYDROLOGY; HYDROMETEOROLOGY; METEOROLOGICAL INSTRUMENTATION.
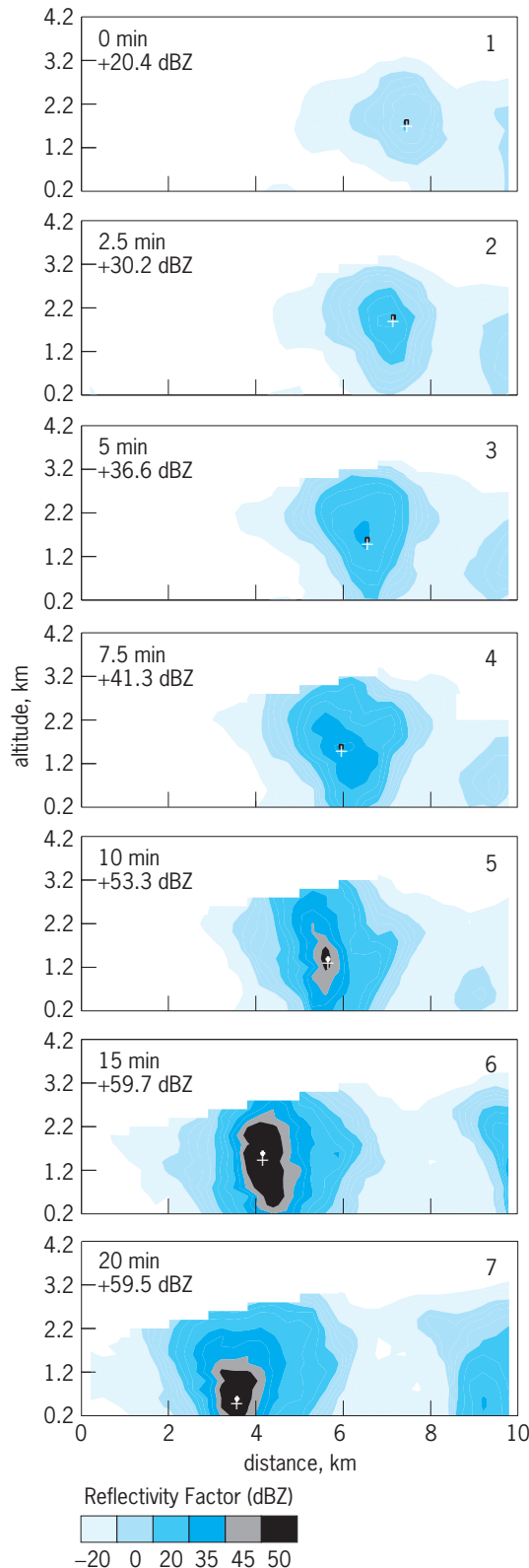
Aviation meteorology. Thunderstorms and associated phenomena such as hail and lightning pose a direct threat to aviation. Radar is used routinely to direct air traffic around and away from thunderstorms. A particular hazard for aircraft approaching or departing an airport is wind shear, a sudden sharp change in wind speed along the flight path. Wind shear occurs along gust fronts, the leading edge of fast-moving rain-cooled air rushing outward from precipitating areas of thunderstorms. Wind shear also occurs in downbursts, localized intense downdrafts that induce an outburst of damaging winds on or near the ground. Downbursts can develop when rain falling from a storm passes though a dry layer and evaporates, cooling air in the dry layer and causing it to descend.

Aircraft encountering wind shear can rapidly lose air speed, which reduces lift and may cause loss of control. Many disastrous aircraft accidents have been attributed to low-level wind shear encountered on approach or departure. Doppler radars are now used at many airports to detect wind shear. Radial velocity measurements from the radar are processed by computer algorithms that rapidly deliver information to air-traffic controllers about the expected air speed loss on approach or departure. Relayed to pilots, this information allows an informed decision about continuing or aborting a takeoff or landing. *See* AERONAUTICAL METEOROLOGY; LIGHTNING.

Cloud and precipitation physics. Current understanding of the physics of precipitation processes derives from direct measurement of cloud properties made by aircraft, remote measurements of clouds made by radar and other remote sensors, laboratory measurements, and theoretical studies. Precipitation forms through two distinct processes, the first involving the collision and coalescence of cloud droplets, and the second involving the formation of ice particles. Radar provides one of the best methods for detecting precipitation particles in clouds because of the sixth-power relationship between particle diameter and reflectivity [Eq. (5)]. Radar measurements, combined with measurements of the temperature and height at which early radar echoes form, have been used to study the relative importance of these two processes in different cloud types. The subsequent growth of precipitation particles can also be inferred from changes in the reflectivity as precipitation falls to the Earth's surface (see **illus.**). Radar can also be used to identify areas of new precipitation formation in widespread clouds. Early studies of radar echoes showed that precipitation in cold stratiform clouds often formed in localized regions near cloud top. These regions, called generating cells by radar meteorologists, produce streams of ice particles which

fall into the lower part of the cloud while growing larger (**Colorplate 3**). *See* CLOUD; CLOUD PHYSICS.

Radar can provide information concerning the type and fall speed of precipitation particles in clouds. The fall speed of precipitation particles is related to their size and shape, and to the motion of the air. In clouds with relatively weak vertical motions of a few cm/s or less, the distribution of fall speeds measured by a vertical-pointing Doppler radar can be related to the particle size distribution. Measurements of the change in the spectrum of fall speeds as particles descend from the cloud top to the ground provide information about precipitation growth.

The most prominent feature of radar echoes in widespread clouds is a "bright band" of high reflectivity at the altitude of the melting level. Several effects explain the intensification of the reflectivity within the bright band. As snowflakes begin to melt, their reflectivity rapidly increases because the dielectric factor $|K|^2$ for water exceeds that for ice [Eq. (3)]. Snowflakes become more compact as they continue to melt and descend, finally collapsing into raindrops. The drops have a smaller radar cross section and fall faster than snowflakes, resulting in a more dilute concentration of smaller particles below the melting layer than within it. These effects reduce the reflectivity below the melting layer.

Polarization diversity radars provide a means of discriminating ice and water in clouds and, in some cases, permit identification of the predominant type of ice particles in different cloud regions. The most important application of polarization techniques is hail detection. A second potential application is the detection of liquid water droplets in regions of widespread clouds where the temperature is below $0°C$ ($32°F$). Detection of these supercooled water droplets is important because they can cause airframe icing.

Wind measurements. Radars operating at longer wavelengths, called wind profilers, are sensitive to turbulent irregularities in the radio refractive index associated with temperature and humidity variations on a scale of half the radar's wavelength. Profilers typically operate with three to five fixed beams, one pointing upward and the others in orthogonal directions. Doppler frequency measurements enable the estimation of drift velocities of the scattering elements, from which the wind velocity can be obtained. Wind profilers operate best in precipitation-free conditions. These radars can measure a vertical profile of the wind speed and direction over the profiler site at time intervals as short as 5 min (standard wind measurements made worldwide from balloon-borne instruments are available only twice each day).

Shorter-wavelength radars used for storm detection also can be used to generate vertical wind profiles. The technique involves analysis of the change in the radial (along-beam) component of the wind on a circle at a fixed distance from the radar. Distant circles correspond to higher elevations. To make these measurements, echoes must cover a sufficient area around the radar. For this reason, the technique works best when precipitation is present. Together, wind profilers and shorter-wavelength radars can provide wind measurements aloft in most atmospheric conditions.

Vertical cross section showing the evolution of the radar reflectivity factor (dBZ) during a 20-min period in a shower located near the coast of the island of Hawaii.

Storm structure studies. Much of the understanding of the structure of storms derives from measurements made with networks of Doppler radars. Two or more Doppler radars simultaneously observing the same volume of space allow the estimation of more than one component of the wind. With appropriate scanning procedures, Doppler radars can scan the entire volume of thunderstorms or other phenomena in a few minutes, so that the entire wind field within the storm can be determined. Techniques have also been developed to estimate pressure and temperature fields within storms from the wind fields once they are deduced. Many research projects since the mid-1970s have employed networks of Doppler radars to investigate the complete three-dimensional wind fields associated with storms, fronts, and other meteorological phenomena.   Robert M. Rauber

Bibliography. D. Atlas (ed.), *Radar in Meteorology*, American Meteorological Society, 1990; L. J. Battan, *Radar Observation of the Atmosphere*, University of Chicago Press, 1973; R. J. Doviak and D. S. Zrnić, *Doppler Radar and Weather Observations*, 2d. ed., Academic Press, 1993; E. E. Gossard and R. G. Strauch, *Radar Observations of Clear Air and Clouds*, Elsevier, 1983.
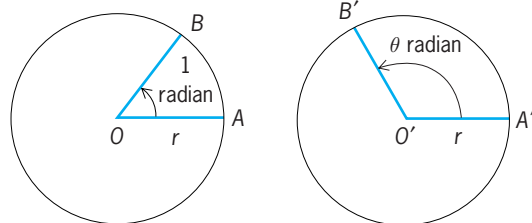
Radian measure

A radian is the angle subtended at the center of a circle by an arc of the circle equal in length to its radius. It is proved in geometry that equal central angles of two circles subtend arcs proportional to their radii; and the converse is true. Hence the radian is independent of the length of the radius. The **illustration** represents two circles of radius r. Arc AB of length r subtends 1 radian (rad) at the center O of the circle, and arc $A'B'$ of length s subtends θ rad at its center. Since arcs on equal circles are proportional to their subtended central angles, $s/r = \theta/1$, or formula (1) holds. If $\theta = 2\pi$, $s = 2\pi r$,

$$S = r\theta \tag{1}$$

the circumference of the circle. Therefore 2π rad is the complete angle about a point or $360°$, and

$$2\pi \text{ rad} = 360°$$
$$1 \text{ rad} = 360°/2\pi = 57.2958° = 57°17'45''$$
$$1° = 2\pi/360 \text{ rad} = 0.0174533 \text{ rad}$$

Diagrams showing radian measurement. Symbols are explained in the text.

Observe that

$$30° = 30\pi/180 \text{ rad} = \pi/6 \text{ rad}$$
$$45° = \pi/4 \text{ rad}$$
$$60° = \pi/3 \text{ rad}$$
$$90° = \pi/2 \text{ rad}$$
$$135° = 3\pi/4 \text{ rad}$$

The degree as a unit of angle has come down from antiquity. However, its use in various theories involves clumsy constants. The use of the radian avoids these constants. The radian is employed generally as a measure of angle in theoretical discussions; when no unit of angle is mentioned, the radian is understood.

The following examples illustrate the simplicity and convenience of radian measure.

The simple formula (1) would be $s = (\pi/180) r\theta$ if degrees were used. A mariner on a ship observes that a lighthouse, known to be 400 ft (120 m) high, subtends 0.064 rad at the eye, and writes from Eq. (1) that

$$400 = r(0.064) \qquad r = \frac{400}{0.064} \text{ ft} = 6250 \text{ ft}$$

The mariner then concludes that it is approximately 6250 ft (1875 m) to the lighthouse. The important ratios $(\sin\theta)/\theta$ and $(1 - \cos\theta)/\theta^2$ approach, when θ approaches zero, respective limits of 1 and $\frac{1}{2}$ when radians are used, but $\pi/180$ and $\pi^2/64{,}800$ when degrees are used. The use of radians avoids these constants throughout the many fields of application of the trigonometric functions. If a particle A moving on a circle with center O and radius r has velocity v, tangential acceleration a_t, normal acceleration a_n, and if OA has angular velocity ω and angular acceleration α, then relations (2) hold when radians are used. Each of these

$$v = r\omega \qquad a_t = r\alpha \qquad a_n = r\omega^2 \tag{2}$$

basic formulas would involve an unwieldy constant if degrees were employed for angle measure. The simplicity arising from use of the radian indicates the importance of a wise choice of basic units in mathematics and its applications. *See* UNITS OF MEASUREMENT.   Lyman M. Kells

Radiance

The physical quantity that corresponds closely to the visual brightness of a surface. A simple radiometer for measuring the (average) radiance of an incident beam of optical radiation (light, including invisible infrared and ultraviolet radiation) consists of a cylindrical tube, with a hole in each end cap to define the beam cross section there, and with a photocell against one end to measure the total radiated power in the beam of all rays that reach it through both holes (see **illus.**). If A_1 and A_2 are the respective areas of the two holes, D is the length of the tube (distance between holes), and Φ is the radiant flux or power
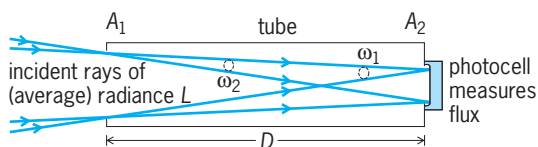
A simple radiometer.

measured by the photocell, then the (average) radiance is approximately given by the equation below.

$$L = \frac{\Phi}{A_1 \cdot A_2/D^2} \, \mathrm{W \cdot m^{-2} \cdot sr^{-1}}$$

$$= \frac{\Phi}{A_1 \cdot \omega_2} = \frac{\Phi}{A_2 \cdot \omega_1}$$

In the alternate forms on the second line, $\omega_2 = A_2/D^2$ is the solid angle subtended at A_1 by A_2 (at a distance D), and vice versa for ω_1. These approximate relations are good to about 1% or better when D is at least 10 times the widest part of either hole, but the accuracy degenerates rapidly for larger holes or shorter distances between them. A more exact formula, in terms of the ray radiance in each direction at every point across an area through which a beam flows, requires calculus. Then the radiance of a ray in a given direction through a given point is defined as the radiant flux per unit projected area perpendicular to the ray at the point and unit solid angle in the direction of the ray at the point. *See* LIGHT.

Power (flux) is given in SI units of watts (W), areas in square meters (m^2), and solid angles in steradians (sr), and so the units of radiance are, as shown, watts per square meter and steradian ($\mathrm{W \cdot m^{-2} \cdot sr^{-1}}$). If the flux of visible radiation is given in units weighted for standardized eye response, called lumens (lm), instead of watts, with everything else exactly the same, the corresponding photometric quantity of luminance is obtained in $\mathrm{lm \cdot m^{-2} \cdot sr^{-1}}$. When concerned with interactions between radiation and matter, flux can also be measured in numbers of photons or quanta per second ($\mathrm{q \cdot s^{-1}}$) rather than watts.

A complete specification of a beam of optical radiation requires the distribution of radiance, not only as a function of ray position and direction, but also as a function of wavelength, time, and polarization. A distribution in space and wavelength λ is spectral radiance L_λ in watts per square meter, steradian, and nanometer of wavelength ($\mathrm{W \cdot m^{-2} \cdot sr^{-1} \cdot nm^{-1}}$). Spectral radiance may also be given in terms of wave number σ in reciprocal centimeters ($\mathrm{cm^{-1}}$) or of frequency ν in terahertz (THz). No special names are given to the distributions with respect to time (or frequency of fluctuation or modulation) and polarization, and they are often overlooked even though their effects may be significant.

All of the foregoing assumes that radiant energy is propagated along the rays of geometrical optics that can form sharp shadows and images. The results may be in error when there are significant diffraction or interference effects, as is often the case with coherent radiation such as that from a laser. *See* DIFFRACTION; GEOMETRICAL OPTICS; INTERFERENCE OF WAVES; LASER.   Fred E. Nicodemus

Bibliography. R. W. Boyd, *Radiometry and the Detection of Optical Radiation*, 1983; R. M. McCluney, *Introduction to Radiometry and Photometry*, 1994; F. E. Nicodemus, Radiance, *Amer. J. Phys.*, 31:368–377, 1963.

Radiant heating

Any system of space heating in which the heat-producing means is a surface that emits heat to the surroundings by radiation rather than by conduction or convection. The surfaces may be radiators such as baseboard radiators or convectors, or they may be the panel surfaces of the space to be heated. *See* PANEL HEATING AND COOLING; RADIATOR.

The heat derived from the Sun is radiant energy. Radiant rays pass through gases without warming them appreciably, but they increase the sensible temperature of liquid or solid objects upon which they impinge. The same principle applies to all forms of radiant-heating systems, except that convection currents are established in enclosed spaces and a portion of the space heating is produced by convection. The radiation component of convectors can be increased by providing a reflective surface on the wall side of the convector and painting the inside of the enclosure a dead black to absorb heat and transmit it through the enclosure, thus increasing the temperature of that side of the convector exposed to the space to be heated.

Any radiant-heating system using a fluid heat conveyor may be employed as a cooling system by substituting cold water or other cold fluid. This cannot be done with electric resistance-type radiant-heating systems. Thermoelectric couples will emit or absorb heat, depending upon the polarity of the direct current applied to them. However, the technique is practical only for very special and small-scale heating and cooling applications, certainly not on the scale required for comfort control of an occupied space. *See* COMFORT HEATING; HEAT RADIATION.   Erwin L. Weber; Richard Koral

Bibliography. M. F. Modest, *Radiative Heat Transfer*, 2d ed., 2003; R. Siegel and J. R. Howell, *Thermal Radiation Heat Transfer*, 4th ed., 2001; R. Watson and K. Chapman, *Radiant Heating and Cooling Handbook*, 2002.

Radiation

The emission and propagation of energy; also, the emitted energy itself. The etymology of the word implies that the energy propagates rectilinearly, and in a limited sense, this holds for the many different types of radiation encountered.

The major types of radiation may be described as electromagnetic, acoustic, and particle, and within these major divisions there are many subdivisions.

For example, electromagnetic radiation, which in the most familiar energy ranges behaves in a manner usually characteristic of waves rather than of

particles, is classified roughly in order of decreasing wavelength as radio, microwave, visible, ultraviolet, x-rays, and gamma rays. In the last three subdivisions, and frequently in the visible, the behavior of the radiation is more particlelike than wavelike.

Since the energy of a photon (light quantum) is inversely proportional to the wavelength, this classification is also on the basis of increasing photon energy. See ELECTROMAGNETIC RADIATION.

Acoustic or sound radiation may be classified by frequency as infrasonic, sonic, or ultrasonic in order of increasing frequency, with sonic being between about 16 and 20,000 Hz. Infrasonic sound can result, for example, from explosions or other sources so loud that exceptional waves are set up because the large amplitudes of the source vibrations exceed the elastic limit of the transmitting medium. Ultrasonic sound can be produced by means of crystals which vibrate rapidly in response to alternating electric voltages applied to them. There is a nearly infinite variety of sources in the sonic range. See SOUND.

The traditional examples of particle radiation are the alpha particles and beta particles of radioactivity. Cosmic rays also consist largely of particles— protons, neutrons, and heavier nuclei, along with beta particles, mesons, and the so-called strange particles. See COSMIC RAYS; ELEMENTARY PARTICLE.
McAllister H. Hull, Jr.

Radiation biology

The study of the action of ionizing and nonionizing radiation on biological systems. Ionizing radiation includes highly energetic electromagnetic radiation (x-rays, gamma rays, or cosmic rays) and particulate radiation (alpha particles, beta particles, neutrons, or heavy charged ions). Nonionizing radiation includes ultraviolet radiation, microwaves, and extralow-frequency (ELF) electromagnetic radiation. These two types of radiation have different modes of action on biological material: ionizing radiation is sufficiently energetic to cause ionizations, whereas nonionizing radiation causes molecular excitations. In both cases, the result is that chemical bonds of molecules may be altered, causing mutations, cell death, or other biological changes. See ELECTROMAGNETIC RADIATION; RADIATION.

Ionizing radiation originates from external sources (medical x-ray equipment, cathode-ray tubes in television sets or computer video displays) or from internal sources (ingested or inhaled radioisotopes, such as radon-222, strontium-90, and iodine-131), and is either anthropogenic (medical, industrial, or military) or natural (atmospheric or terrestrial).

Nonionizing radiation originates from natural sources (sunlight, Earth's magnetic field, lightning, static electricity, endogenous body currents) and technological sources (computer video displays and television sets, microwave ovens, communications equipment, electric equipment and appliances, and high-voltage transmission lines).

Ionizing Radiation

The action of ionizing radiation is best described by the three stages (physical, chemical, and biological) that occur as a result of energy release in the biological target material (**Fig. 1**).

Physical stage. All ionizing radiation causes ionizations of atoms in the biological target material. The Compton effect, which predominates at the energies of electromagnetic radiation that are commonly encountered (for example, x-rays or gamma rays), strips orbital electrons from the atoms. These electrons (Compton electrons) travel through the target material, colliding with atoms and thereby releasing packets of energy. About 60 eV of energy is released with each collision, breaking chemical bonds and causing ionizations known as primary ionizations. The typical initial energy of a Compton electron is 200 keV, and so there may be over 3300 primary ionizations from a single Compton electron. For low-energy x-rays, the photoelectric effect predominates, producing photoelectrons that transfer their energy in the same manner as Compton electrons. See COMPTON EFFECT; ELECTRON; GAMMA RAYS; X-RAYS.

The absorbed dose of ionizing radiation is measured as the gray (Gy, 1 joule of energy absorbed by 1 kilogram of material). In general, the amount of absorption of ionizing radiation is greater for target material of high density and high atomic number, and it is independent of the chemical bonds of the target material. Because of the very localized absorption of ionizing radiation as compared to heat energy, an amount of ionizing radiation energy equivalent to 1/100 the heat energy in a cup of coffee will result in a 50% chance that the person absorbing the radiation will die in 30 days.

Neutrons with energies between 10 keV and 10 MeV transfer energy mainly by elastic scattering, that is, billiard-ball-type collisions, of atomic nuclei in the target material. In this process the nucleus is torn free of some or all of the orbital electrons because its velocity is greater than that of the orbital electrons. The recoiling atomic nucleus behaves as a positively charged particle. Because the mass of the neutron is nearly the same as that of the hydrogen atom, hydrogenous materials are most effective for energy transfer. A unit known as the kerma (acronym for kinetic energy released in material) is used to measure the amount of energy transfer from neutrons and other indirectly ionizing radiation (for example, x-rays and gamma rays). This quantity is frequently equivalent to the absorbed dose of radiation in the material (the gray) because the ranges of the secondary recoiling charged particles are much shorter than the ranges of neutrons, x-rays, or gamma rays. See NEUTRON.

Chemical stage. Chemical changes in biological molecules are caused by the direct transfer of radiation energy (direct radiation action) or by the production of chemically reactive products from radiolysis of water that diffuse to the biological molecule (indirect radiation action). More than half the biological action of low linear-energy-transfer (LET) ionizing

radiation (for example, x-rays and gamma rays) results from indirect radiation action, about 90% of which is due to the action of the hydroxyl radical (OH·). Characteristically, the manifestations of indirect radiation action decrease as the linear energy transfer of the radiation increases. Thus, for high linear-energy-transfer radiation, direct radiation action predominates. Chemicals that react with hydroxyl radicals, rendering them unreactive, provide protection against indirect radiation damage. For example, sulfhydryl compounds (RSH) remove hydroxyl radicals by reaction (1), where R represents

$$2OH \cdot + 2RSH \longrightarrow 2H_2O + RSSR \qquad (1)$$
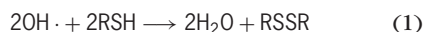
an organic functional group. *See* LINEAR ENERGY TRANSFER (BIOLOGY).

Indirect radiation action is also responsible for the oxygen effect of radiobiology, which describes the increased biological sensitivity when irradiation occurs in the presence of oxygen compared to the absence of oxygen during irradiation. The oxygen must be present during irradiation because of the limited lifetime (less than 1 millisecond) of the chemical species that react with oxygen. The oxygen effect is observed only in biological systems with membranes; the radiosensitivity of viruses and free deoxyribonucleic acid (DNA) is not enhanced by the presence of oxygen. Thus, two types of radiation damage are probable: oxygen-dependent radiation damage, which involves the membrane as the principal target; and oxygen-independent damage, which involves nonmembrane components, including the genome. *See* RADIATION CHEMISTRY.

The most important biological targets for damage from ionizing radiation are probably the plasma membrane and DNA, because there is only one copy, or a few copies, in the cell; because they serve critical roles for the survival and propagation of cells; and because they are large. The last factor is important because ionizing radiation releases its energy in a random manner; thus the larger the target, the more likely that it will be damaged by radiation.

Membranes are composed of protein (50%) and lipids (50%). Biochemical damage in either of these components can cause membrane damage. Consequences of radiation damage in membranes are changes in ion permeability, with leakage of potassium ions; changes in active transport; and cell lysis. *See* CELL MEMBRANES.

The initial direct radiation action results in hydrogen abstraction [reaction (2)]. In the presence of oxygen (O_2), reactions (3)–(6) occur. Oxidative

$$RH \longrightarrow R \cdot + H \cdot \qquad (2)$$

$$H \cdot + O_2 \longrightarrow HO_2 \cdot \qquad (3)$$

$$R \cdot + HO_2 \cdot \longrightarrow ROOH \qquad (4)$$

$$R \cdot + O_2 \longrightarrow RO_2 \cdot \qquad (5)$$
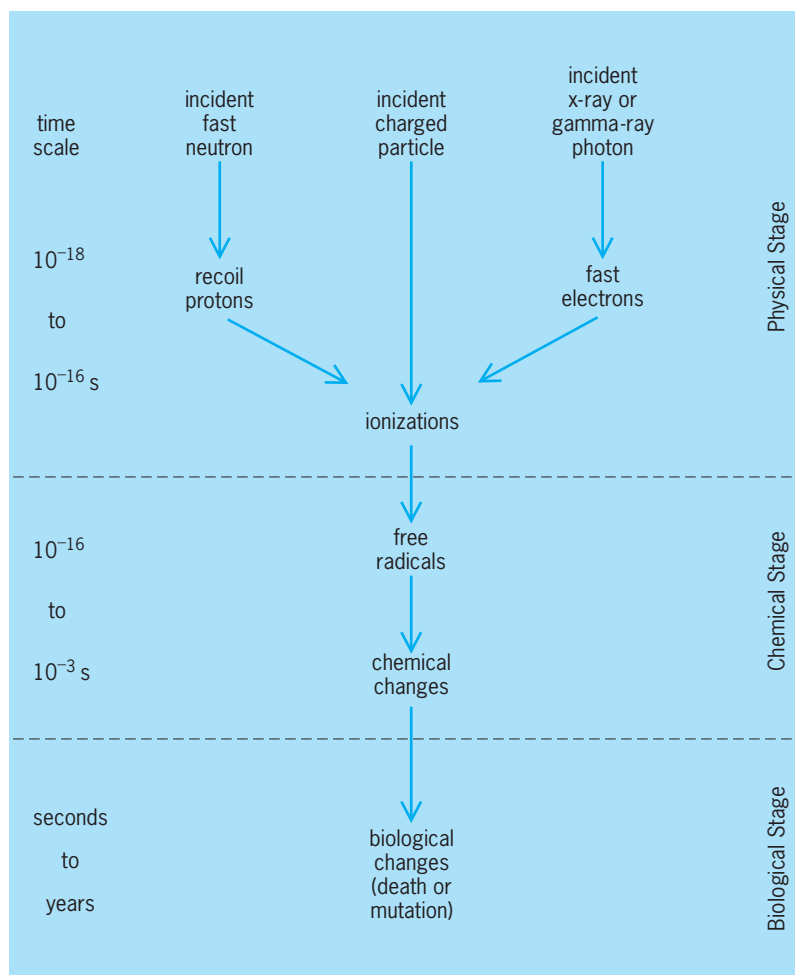
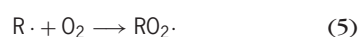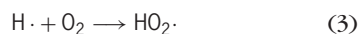$$RO_2 \cdot + H \cdot \longrightarrow ROOH \qquad (6)$$

Fig. 1. Diagram of the three stages for the action of ionizing radiation on biological systems. The duration of the physical stage approximates the time required for the ionizing radiation to traverse an atom. During the chemical stage, molecular and thermal equilibrium is achieved. The biological stage involves metabolism of the chemical damage.

decomposition products of the organoperoxides (ROOH) accumulate, and overall lipid damage is increased approximately threefold when irradiation occurs in the presence of oxygen. Two sensitive sites for protein damage are the free amino and sulfhydryl groups, as observed by the release of ammonia and hydrogen sulfide gases during irradiation of proteins. Another sensitive site is the imidazole ring of the amino acid histidine. The peptide bond is also subject to attack, which causes breakage of the polypeptide chain and a reduction of the molecular weight of the protein. *See* LIPID; PROTEIN.

Membrane components can be protected by reactions of the initially formed organic radical with sulfhydryl compounds [reaction (7)]. The sulfhydryl

$$R \cdot + RSH \longrightarrow RH + RS \cdot \qquad (7)$$
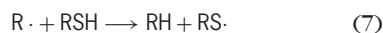
compound restores the initially damaged membrane component to its original state. This radioprotective reaction does not appear to be important for protection of DNA.

Lesions in DNA that is irradiated in aqueous solution include single-strand breaks, double-strand

breaks, base damage, interstrand cross-links, and DNA-protein cross-links. Base damage occurs most in thymine and least in guanine, and the yield of single-strand breaks is about 10 times the yield of double-strand breaks and cross-links. The ratio of the radiosensitivities of thymine in DNA is 1:2:6 when the DNA is in the form of free DNA, active chromatin, and condensed chromatin, respectively. This demonstrates protection of DNA by the proteins of the chromatin against lesions from diffusible water (hydroxyl) radicals. Oxidations of sugars in the DNA backbone and loss of whole nucleotides are the main mechanisms for formation of breaks in strands. *See* DEOXYRIBONUCLEIC ACID (DNA).

Biological stage. Various biological effects can result from the biological actions of ionizing radiation (**Fig. 2**). Reproductive death is most pronounced in mammalian cells that are actively dividing and in nondifferentiated tissue. Thus, dividing tissues (bone marrow and the germinal cells of the ovary and testis) are radiosensitive, and nondividing tissues (liver, kidney, brain, muscle, cartilage, and connective tissue) are radioresistant in animals exposed to whole-body radiation. Developing embryos are quite radiosensitive, as predicted by this principle. The radiosensitivity of organisms varies greatly, being related to their intrinsic sensitivity to radiobiological damage and to their ability to repair the damage. Radiation doses resulting in 10% survival range from 3 Gy (mouse and human cells), to 60 Gy (most bacteria and the fruit fly), to 130 Gy (cabbage looper), to 600 Gy (viruses),
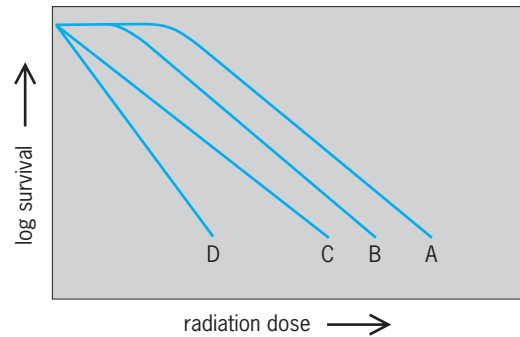

Fig. 3. Biological survival curves. A and B are biphasic with an initial shoulder period followed by exponential inactivation kinetics. C and D are exponential only.

to greater than 1000 Gy (the bacterium *Deinococcus radiodurans*).

Single-break chromosome damage leads to chromosome aberrations in cells irradiated early in the cell cycle (before DNA synthesis has started) and to aberrations in chromatid aberrations in cells irradiated later in the cell cycle (after DNA synthesis). The yield of chromosome breaks increases twofold when oxygen is present during irradiation. *See* CELL CYCLE; CHROMOSOME ABERRATION.

Cells lose their reproductive capacity when exposed to ionizing radiation. Mammalian cells are most resistant to this action during the late S phase of the cell cycle, and most sensitive during mitosis. *See* MITOSIS.

Repair of biological damage involves repair of damage in the DNA. The kinetics of biological inactivation can be biphasic, with an initial "shoulder" period followed by exponential kinetics (**Fig. 3**, A and B), or exponential only (Fig. 3, C and D). Two types of biological repair are involved in the differences in the magnitude of the shoulder region and the slope of the exponential region: the shoulder region arises from the action of Q repair, which diminishes as dose increases, and the value of the exponential slope is affected by the action of P repair, which is not diminished as dose increases. Curves A and B in Fig. 3 exemplify conditions in which the biological system of curve A has a more active Q-repair system than that of curve B while both have the same amount of P repair. Alternatively, the final slope of a curve can be influenced by the intrinsic radiosensitivity of the biological system. Also, the initial shoulder period can result from the necessity that several critical targets must be inactivated before the biological system is inactivated, or several radiation-damaging events (hits) must occur within the same target in order for biological inactivation to occur (Fig. 3, A and B).

The three organ systems that generally contribute to the death of mammals following a single dose of whole-body irradiation are, in decreasing order of radiosensitivity, the hematopoietic system, the gastrointestinal system, and the cerebrovascular system. At very high doses (above 100 Gy), the survival time of the animal is short (from minutes to several days after exposure, depending on the dose), and death results from damage to the cerebrovascular system.
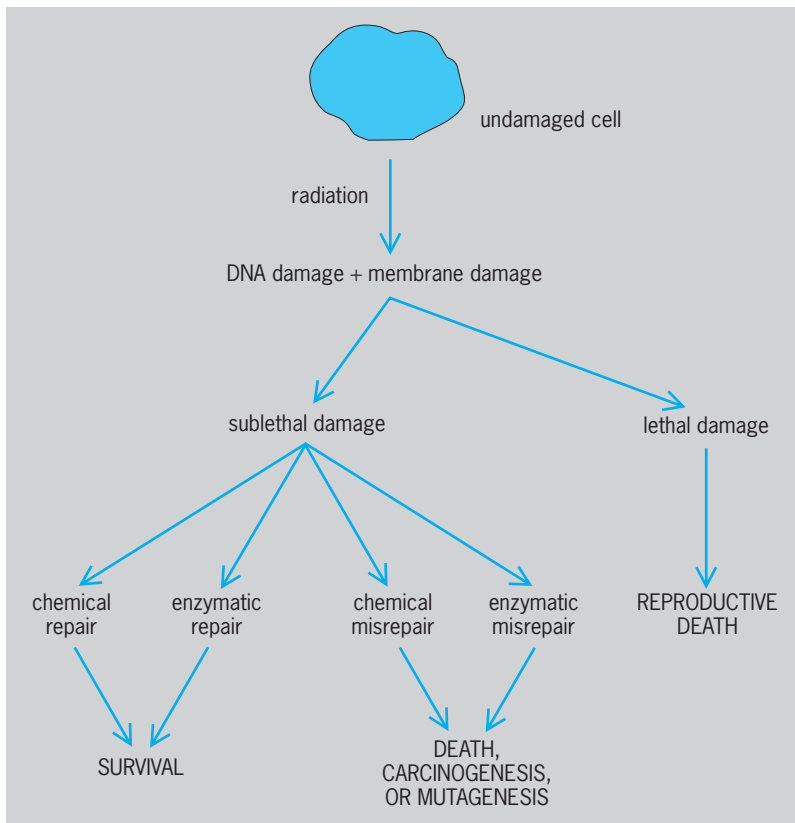

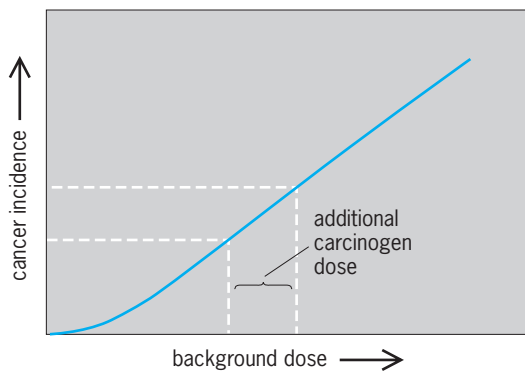Fig. 2. Possible pathways for cellular response to ionizing radiation.

Fig. 4. Additive interaction of carcinogenic agents. (*After A. C. Upton, The question of thresholds for radiation and chemical carcinogenesis, Canc. Invest., 7(3):267–276, 1989***)**

At intermediate doses (between 10 and 100 Gy), survival is longer (between 3 and 4 days), and death occurs because of gastrointestinal damage. The longest duration between exposure and death (between 1 week and 1 month) follows radiation doses of 3 to 10 Gy, and arises from failure of the hematopoietic system.

Late somatic effects may take years or decades to appear and include genetic mutations transmitted to subsequent generations, tumor development and carcinogenesis, and shortening of life span. *See* MUTAGENS AND CARCINOGENS; MUTATION; TUMOR.

An important and controversial issue concerning ionizing radiation is the question of the existence of a threshold dose, below which no biological effects occur. The most conservative model uses a linear equation to relate the incidence of biological effects with increasing dose. Use of a quadratic equation to describe the dose response predicts the presence of a threshold dose. Because radiation damage is cumulative, and this damage can interact with other chemical agents in causing biological effects such as carcinogenesis, the linear model is favored (**Fig. 4**). *See* RADIATION INJURY (BIOLOGY).

Nonionizing Radiation

Of all the nonionizing radiations, only ultraviolet radiation, microwaves, and high-voltage electromagnetic radiation are considered in the study of radiation biology.

Ultraviolet radiation. This type of radiation is much less penetrating than ionizing radiation. Since it can penetrate only several layers of cells, the effects of ultraviolet (UV) radiation on humans are restricted to the skin and the eyes. Ultraviolet radiation is divided into UV-C (wavelength of 200–280 nanometers), UV-B (280–320 nm), and UV-A (320–400 nm). The most biologically damaging is UV-C, and the least damaging is UV-A, with UV-B having intermediate efficiency of biological action. The solar spectrum at the Earth's surface contains only the UV-B and UV-A radiations. Stratospheric ozone strongly absorbs UV-C radiation and the shorter wavelength portion of UV-B radiation, thus providing some biological protection. Depletion of the ozone layer caused by stratospheric pollution dramatically reduces this protective action, resulting in a decreased atmospheric absorption of UV-B radiation. *See* AIR POLLUTION; STRATOSPHERIC OZONE.

Biological effects can arise only when absorption of ultraviolet radiation occurs. Absorption is dependent on the chemical bonds of the material, and it is highly specific. Identification of the critical target for biological action has been facilitated by this specific absorption, using a comparison to the action spectrum, that is the relative efficiency for causing biological effects at different wavelengths. The action spectrum for ultraviolet radiation closely matches the absorption spectrum of nucleic acid, at wavelengths between 240 and 300 nm. At wavelengths greater than 320 nm, ultraviolet-mediated damage is oxygen-mediated, because of the interaction of dioxygen (O_2) with photoexcited biological chromophores, whereas below 300 nm, biological damage occurs principally by photochemical damage in DNA. Exposure dose for ultraviolet radiation is measured in units of joules per square meter; it is a measure of the incident energy per unit area impinging on the target material.

Biological responses. Sunburn is a form of erythema produced by overexposure to the UV-B portion of the solar spectrum (which is not transmitted through window glass). A rare but deadly form of skin cancer in humans, malignant melanoma, is induced by exposure to sunlight, with occurrences localized on those regions of the body that are most frequently exposed.

Ultraviolet light can also cause photochemical damage. Cyclobutane pyrimidine dimers are the main photoproduct following exposure to UV-C and UV-B, and they can lead to cell death and precarcinogenic lesions. Other types of dimers are considered to be especially mutagenic. DNA-protein cross-links that are observed after ultraviolet radiation can be lethal.

Repair systems. Survival from ultraviolet irradiation is reduced as the dose of radiation is increased. The shapes of survival curves are similar to those for lethality from ionizing radiation (Fig. 3); they are dependent on the presence or absence of repair systems. The four repair systems that enhance biological survival are discussed below.

Photoreactivation is an enzymatic repair system that enhances biological survival by splitting cyclobutane dimers in the DNA of cells that have been irradiated by ultraviolet light. The process requires light, and the most effective wavelengths are in the blue region of the visible spectrum. The repair system is error-free, and thus it is nonmutagenic. However, the repair must correct the damage before it is copied by DNA replication.

Another system is excision repair. A region of DNA containing bases that have been damaged by ultraviolet radiation is removed enzymatically, followed by synthesis of new DNA to replace the damaged region. There are no errors, and mutant cells that lack this repair system are highly sensitive to ultraviolet radiation. In fact, cells isolated from individuals with the genetic-recessive disease xeroderma

pigmentosum lack excision repair. The clinical symptoms of this disease include a high incidence of skin lesions and early death from malignant melanoma. Excision repair operates in the dark. The damage must be corrected before DNA replication copies it. Not all regions of DNA are equally accessible to the action of this repair system, and it has been observed that actively transcribed regions of DNA in mammalian cells are more efficiently repaired than are nontranscribed regions.

Recombination repair is another repair system. DNA that contains damage is replicated, and gaps in the newly synthesized DNA appear opposite to the sites where there is damage in the parental strand. These gaps are filled in by recombining the portion of the undamaged complementary parental strand with the daughter strand containing the gaps. This repair system is essentially error-free; it is sometimes known as postreplicational repair, because it occurs after replication of the damaged region. *See* RECOMBINATION (GENETICS).

The system known as SOS repair is an inducible repair system. It is activated by a reactive intermediate of DNA metabolism in irradiated bacteria. SOS repair is associated with a complex of responsive genes known as *din* (damage-inducible) genes. SOS repair acts by inducing synthesis of an alternative excision repair system, and the recombination repair system, which increases survival. The alternative excision repair system involves removal of a longer segment of DNA containing the damaged region; it is error-prone, which leads to increased mutations. Another manifestation of the SOS response is the induction of DNA replication that can bypass the damaged template region of DNA. Thus this is not actually a repair system, since the original damage remains in the DNA, and can best be thought of as a system for tolerating DNA damage. Survival is increased, because DNA replication is not blocked by the presence of the damage. *See* ULTRAVIOLET RADIATION; ULTRAVIOLET RADIATION (BIOLOGY).

Microwaves. Microwaves are electromagnetic radiation in the region from 30 MHz to 300 GHz. They originate from devices such as telecommunications equipment and microwave ovens. Metals reflect microwaves; glass transmits them; and aqueous material absorbs them, accompanied by a rise in temperature of the liquid. The 915- and 2450-MHz bands are used for industrial microwave heating and in microwave ovens. Thermal effects of microwaves occur at exposure rates greater than 10 mW/cm^2 (70 mW/in.^{2}), while nonthermal effects are associated with exposure rates less than 10 mW/cm^2.

Microwave radiation is absorbed unevenly in biological tissue because of the heterogeneity of the dielectric properties of the material. Thus, significant temperature gradients can be established, which may enhance the action of thermal heating from microwaves as compared to that from infrared radiation. Certainly, material with a high water content will have a higher absorption coefficient for microwaves, and thus a greater thermal response to microwave action. Microwave absorption is high in skin, muscle, and internal organs, and lower in bone and fat tissue.

Biological effects. Cultured mammalian cells exposed to microwaves at a high power density show chromosome abnormalities after 15 min of exposure. Progression through the cell cycle is also temporarily interrupted, which interrupts DNA synthesis. Chromosome aberrations in peripheral blood lymphocytes are significantly greater for persons who are occupationally exposed to microwaves.

Microwaves can be lethal when the power intensity and exposure time are sufficient to cause a rise in temperature that exceeds an organism's homeostatic capabilities; this occurs when the temperature rise exceeds approximately 5°C (9°F). For example, cataracts can result from exposure to ultraviolet radiation because the eye is not able to dissipate heat very well since there is not blood circulation. The testicles are at a temperature about 2°C (4°F) below that of body temperature, and spermatogenesis is particularly sensitive to temperature rise. Reversible testicular damage, in the form of reduced spermatogenesis and degeneration of the epithelial lining of the seminiferous tubules, therefore can occur from exposure to microwaves.

There are also some nonthermal effects associated with microwaves. There is significant uncertainty in the actual doses among people exposed as a result of their occupations; this reduces the reliability of the observations of the nonthermal actions of microwaves. However, a list of clinical symptoms includes increased fatigue, periodic or constant headaches, extreme irritability, decreased hearing acuity, and drowsiness during work. Laboratory studies involving exposure of animals to microwaves have produced changes in the electroencephalogram, blood–brain barrier, central nervous system, hematology, and behavior. Cell membrane permeability is also altered.

Exposure limits. It would be desirable to establish limits of exposure to microwaves based on the absorbed dose of microwave energy, similar to the gray unit that is used for ionizing radiation. However, the difficulty associated with measuring the absorbed dose makes this impractical. Therefore, radiation protection standards are based on exposure values instead of the previously used specific absorption rate (SAR, measured as W/kg); exposure dose is measured as W/cm^2. Occupational limits are set at a maximum of 1 mW/cm^2 (7 mW/in.^{2}) for frequencies between 30 and 100 MHz, and 10 mW/cm^2 (70 mW/in.^{2}) for frequencies between 1 and 300 GHz, averaged over a 6-min period. These values are selected to limit the specific absorption rate of the average whole body to 0.4 W/kg (0.2 W/lb), and they take into consideration the variable absorption for microwaves of different frequencies. The maximum acceptable limit for leakage from household microwave ovens is set at 1 mW/cm^2 (7 mW/in.^{2}) or less, measured at a distance of 5 cm (2 in.) from a new oven, and never more than 5 mW/cm^2 measured 5 cm (2 in.) from the oven, during its expected lifetime. *See* INDUSTRIAL HEALTH AND SAFETY; MICROWAVE.

Extremely low frequency electromagnetic fields. This type of radiation is generated by the electric and magnetic fields associated with high-voltage current in power transmission lines, and also some household and industrial electrical equipment. Biological effects from ELF radiation are the least understood, and the potential consequences are the most controversial. The issue of potential biological damage from this type of radiation has arisen only since the introduction of very high voltage electric power transmission lines (440 kV and above) and the occurrence of widespread use of various electrical and electronic equipment. *See* ELECTROMAGNETIC PULSE (EMP).

Biological effects. Biological studies on ELF electromagnetic fields have been performed on cells and whole animals; and epidemiological studies have been carried out on populations exposed occupationally. The results share some common features that had not been expected: (1) There is not always a clear dose response; that is, increasing the exposure does not necessarily give rise to an increased biological effect, as is observed commonly for other types of radiation. (2) Some biological effects are seen only at certain frequencies and dose rates. Some of the reported effects are subjective, and may be related to normal physiological adaptation to environmental changes.

In cellular studies, membrane transport of calcium ions by chick brain cells was altered by exposure to ELF electric fields. Chromosome aberrations have not been observed in lymphocytes after long-term exposure to 60-Hz fields of 50 kV/m (15 kV/ft) and 10 gauss (10^{-3} tesla), suggesting that ELF fields whose strengths are less than these are unlikely to cause cancer or mutations; but this does not provide conclusive evidence that such radiation is noncarcinogenic. The major action of ELF electromagnetic radiation is targeted to the cell membrane, and most likely involves changes in membrane activity.

Controlled laboratory studies on developing chick embryos continuously exposed to 60-Hz electric fields up to strengths of 100 kV/m (30 kV/ft) show no effects on the general health, development, mortality, bone growth, malformations, and behavior for embryos that are hatched and followed for 6 weeks of growth. However, rats exposed to 60-Hz electric fields between 2 and 40 kV/m (0.6 and 12 kV/ft) for 21 days had reached nighttime peaks of pineal gland melatonin secretion.

In humans, qualitative biological effects of low-frequency radiation (0 to 300 Hz) include headaches, lethargy, and decreased sex drive. Humans have been noted to perceive the presence of a 60-Hz electric field when the intensity is in the range of 2 to 12 kV/m (0.6 to 3.6 kV/ft), and animals were observed to avoid entering an area where the electric field was greater than 4 kV/m (1.2 kV/ft).

Some epidemiological studies have shown a pattern of increased cancer (leukemias and brain tumors) among individuals occupationally exposed to increased ELF electromagnetic fields. However, the evidence is weak, because no dose dependency has been found; also the evidence is only associative and not of a cause-and-effect relational type. Some studies have reported no link between exposure to ELF fields and cancer.

Exposure limits. Because of the conflicting experimental results on the biological effects of ELF electromagnetic radiation, the guidelines are of an interim nature. Exposure to electric and magnetic fields gives rise to induced current in the human body. Clearly, exposure should be well below values that would lead to acute effects such as heart fibrillation and pain. Measurements show that a current density of 4 mA/m^2 (0.4 mA/ft^2) averaged over the head and trunk regions occurs when an individual is exposed to an electric field of 10 kV/cm (25 kV/in.), or a magnetic field of 500 G (50 milliteslas). The natural current densities in the body are about 10 mA/m^2 (0.9 mA/ft^2). The rationale for setting exposure limits to radiation from ELF electric and magnetic fields is to limit external exposure to no more than the endogenous values.

Studies on humans exposed to a 60-Hz magnetic field of 5 mT combined with a 60-Hz electric field of 20 kV/m (6 kV/ft) for 4 to 6 hours per day over several days revealed no effects on the health or welfare of the volunteer subjects.

The International Radiation Protection Agency (IRPA), in collaboration with the Environmental Health Division of the United Nations World Health Organization (WHO), has established exposure limits to ELF electric and magnetic fields, based upon available knowledge about the endogenous fields, and the biological effects of ELF electric and magnetic fields. Continuous exposure of the general public should not exceed 2 kV/m (0.6 kV/ft) for 50/60-Hz electric fields, and should not exceed 0.1 mT for 50/60-Hz magnetic fields, which includes a safety factor of 5 with regard to the maximum continuous field exposures allowed for industrial safety [10 kV/m (3 kV/ft) and 0.5 mT]. *See* RADIATION INJURY (BIOLOGY).   Phillip M. Achey

Bibliography. Z. M. Bacq and P. Alexander, *Fundamentals of Radiobiology*, 2d ed., 1961; W. Harm, *Biological Effects of Ultraviolet Radiation*, 1980; IRPA/INIRC Committee Report, Interim guidelines on limits of exposure to 50/60 Hz electric and magnetic fields, *Health Phys.*, 58:113–122, 1990; D. E. Lea, *Actions of Radiations on Living Cells*, 2d ed., 1955; A. H. Nias and R. Dimbleby, *An Introduction to Radiobiology*, 2d ed., 1998; U. L. Prenn, *Introduction to Biological Radiation Effects*, 2d ed., 1994; Special issue on radiation effects on man and animals, *Experientia*, 45:1–114, 1989.

Radiation chemistry

The study of chemical changes resulting from the absorption of high-energy, ionizing radiation. Such radiation includes alpha particles, electrons, gamma rays, fission fragments, protons, deuterons, helium nuclei, and heavier charged projectiles. X-rays are distinguished from gamma rays only as being extranuclear in origin. In absorbing materials of low and

intermediate atomic weight such as aqueous systems and most biological systems, these radiations deposit energy in a largely indiscriminate manner, leaving behind a complex mixture of short-lived ions, free radicals, and electronically excited molecules. This contrasts with the absorption of visible and ultraviolet radiation, in which one or a few specific electronically excited species are formed. Radiation chemical change results from the further reaction of these intermediates. *See* PHOTOCHEMISTRY.

Because of the nature of the absorption process, chemical-product yields are expressed in terms of total energy absorption. The extent of chemical change per unit of absorbed energy is known as the *G* value for the reaction. The SI unit is moles or micromoles of product formed (or reactant decomposed) per joule. The more frequently used unit is molecules changed per 100 electronvolts absorbed: $G(\mu\text{mol/joule}) = 0.103\ G(\text{molecules/100 eV})$.

Despite the initial complexity, a wealth of kinetic information has been obtained from radiation chemical studies.

Sources of high-energy radiation. Sources of high-energy radiations in the laboratory and industry include radioactive nuclides [for example, cobalt-60 (^{60}Co), strontium-90 (^{90}Sr), and hydrogen-3 (^{3}H)] and instruments such as x-ray tubes, Van de Graaff generators, the betatron, the cyclotron, and the synchrotron. An electron accelerator known as the linac (linear electron accelerator) has proved particularly valuable for the study of transient species that have lifetimes as short as 16 picoseconds; and another electron accelerator, known as the Febetron, has been used for the study of the effects of single pulses of electrons with widths of several nanoseconds at very high currents.

Interaction of radiation with matter. The primary absorption processes for high-energy radiation are ionization and molecular excitation. The distribution of the absorbed energy, however, depends significantly upon the nature of the radiation and absorbing medium.

Electrons. If the electron source is a beta-particle-emitting isotope dispersed in the absorbing medium, electron energies (*E*) will range from essentially zero to an $E_{\max}$ characteristic of the decay process. For ^{3}H, $E_{\max}$ is 18 keV; for phosphorus-32 (^{32}P), it is 1.71 MeV. Accelerator-produced electron beams are monoenergetic, and for radiation chemical applications may be from several hundred kiloelectronvolts to 10 MeV. *See* BETA PARTICLES.

Fast electrons undergo energy loss through inelastic scattering by atoms and molecules in which ionization and electronic excitation occur. Except for straggling, which involves a statistical fluctuation in the number of encountered molecules, a charged particle has a definite range in a given medium. This range depends upon the particle charge, mass, and energy and upon the density of the medium. The geometrical distribution of energy deposition along the particle path is determined by its linear rate of energy transfer (LET). The LET will vary along the path of the particle, increasing as the particle velocity decreases. An average LET may be obtained as the total particle energy divided by its range.

An electron of 300 keV will travel, on the average, 80 mm in water and produce approximately 9000 ionization events along its path, some by the secondary electrons released. These events are distributed along the path in small volumes known as spurs. Each spur may contain several ion pairs, electronically excited molecules, and free radicals formed from reaction of these species. Spurs may overlap in tracks of dense ionization, or they may be separated by many molecular diameters in tracks of low LET. Chemical change results from reactions within the spurs and through reaction of intermediates diffusing from the spurs into the bulk medium.

An electron slowed to thermal energy in a liquid medium may be trapped by a small group of solvent molecules that have been oriented by its presence. The hydrated electron (e_{aq}^{-}) is a strong reducing agent, has a mean lifetime of milliseconds in neutral solutions, and has an optical absorption peak at 715 nanometers. It is perhaps the most well characterized intermediate in radiation chemistry. *See* ELECTRON.

Gamma rays. Dependent upon the energy and the absorbing medium, gamma rays transfer energy by three principal interactions.

In photoelectric absorption, the gamma ray imparts all of its energy to an electron that is ejected from the absorbing molecule with an energy equal to that of the gamma ray less the binding energy of the electron. The photoelectron is absorbed, with the production of further ionization and excitation. The photoelectric effect is important for low-energy gamma rays and high-atomic-number absorbers. It is the dominant mode in water for gamma rays or x-rays of less than 20 keV.

In Compton scattering, the gamma ray imparts a portion of its energy to an electron, ejecting it from the molecule. The scattered gamma ray continues with diminished energy, and it may experience further such interactions in the medium. The Compton electron or electrons are responsible for the major energy depositions in the system. Compton scattering is most important for absorbers of low and intermediate atomic numbers. For gamma rays of from 0.1 to several megaelectronvolts, it dominates the absorption process in aqueous and biological media. *See* COMPTON EFFECT.

Pair production may occur with gamma rays of energy greater than 1.02 MeV. The rest-mass energy of the electron is 0.51 MeV; and in an interaction with an atomic nucleus, a gamma ray of twice this energy may be absorbed, with the formation of an electron-positron pair. The pair share the gamma-ray energy in excess of 1.02 MeV and undergo the usual electron absorption behavior. The positron, however, upon reaching thermal energy, will undergo annihilation with an interacting electron, and two 0.51-MeV gamma rays are formed. The probability of pair production increases with gamma-ray energy and with increasing atomic number of the absorber.

It occurs significantly in water and biological material only above 10 MeV. *See* GAMMA RAYS; POSITRON.

Heavy charged particles. These particles interact with matter in qualitatively the same way as electrons, producing ionization and excitation. For a given energy, however, ionization and spur density are much greater for these particles and increase with decreasing particle velocity. Average LET values in water, for 1-MeV particles, are 43, 58, and 190 keV/μm for protons, deuterons, and alpha particles, respectively. *See* ALPHA PARTICLES.

Dosimetry. Evaluation of the yields of radiation-induced reactions requires a knowledge of the energy imparted to the reacting system. The energy deposited in the system is termed the dose, and the measurement process is called dosimetry. This is a distinction from actinometry, referring to the absorption of visible and ultraviolet light. Absorbed energy from ionizing radiation is described in terms of grays (Gy; joule/kg), in rads (100 ergs/g), or in electronvolts per gram or per cubic centimeter.

If the radiation source consists of a uniformly dispersed alpha- or beta-particle-emitting radioactive species, the energy absorption rate may be evaluated from the particle energy and the source intensity. The intensity is determined radiochemically and, for a beta-particle source, the average electron energy must be used. For such as evaluation to be valid, the absorbing sample must be sufficiently large that radiations escaping from the surface can be neglected.

In an aqueous or biological medium containing, for example, 1 millicurie (3.7×10^7 d/s) of iodine-131 (^{131}I; average beta energy, 0.205 MeV) per gram, the dose rate is 1.2×10^{-3} Gy/s or 0.12 rad/s. This will be increased by a small contribution from the accompanying gamma ray.

For systems in which the mechanisms and responses may be determined by the inhomogeneous, microscopic distributions of absorbed energy, for example, biological cells, an average absorbed dose could be inadequate in the interpretation of radiation effects. For the required microdosimetry, specialized techniques and models have been developed.

For a typical radiation-chemical system in which gamma rays such as those from cobalt-60 (^{60}Co) or cesium-137 (^{137}Cs) are the source of ionizing radiation, chemical dosimetry is most usual. The reaction system of interest is replaced in the radiation field with a geometrically similar system having a known chemical response to the absorption of the energy. The latter is a secondary standard that has been calibrated by an absolute method for energy absorption measurement such as calorimetry.

The oxidation of ferrous ion [Fe(II)] to ferric ion [Fe(III)] in aerated solutions containing 0.4 M sulfuric acid (H_2SO_4; the Fricke dosimeter) occurs with a G value of 15.6 molecules per 100 eV for gamma rays and fast electrons. Dependent upon LET, the G values are 5.1 and 3.0 molecules per 100 eV for 5.3-MeV alpha particles and uranium-235 (^{235}U) fission fragments, respectively. The reaction is readily followed by measurement of the Fe(III) absorbance at 305 nm and is used for doses of 50–350 Gy.

Other chemical dosimeters include the cesium [Ce(IV) to Ce(III)] reduction, which occurs in acid solution with a G value of 2.5 molecules per 100 eV and the oxalic acid decomposition in an aerated solution having a G value of 5 molecules per 100 eV. Both values apply to gamma-ray or fast-electron irradiation.
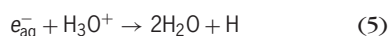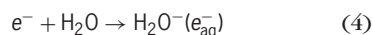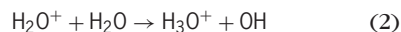
The radiation chemical system being studied will normally be chemically different from the dosimeter solution. The dose as measured in the dosimeter (D_D) must be corrected to apply to the system of interest. With Compton absorption dominant, the appropriate correction factor is the electron density ratio of the system being studied to that of the dosimeter solution. With doses in grays, rads, or electronvolts per gram, the densities have dimensions of electrons per unit mass; for doses in electronvolts per cubic centimeter, electrons per unit volume are required. These are determined by the chemical composition of the dosimeter and system.

For charged-particle radiation, the incident beam is usually totally absorbed in both dosimeter and sample, and energy absorptions are the same. Geometric distributions may differ, however. For thin samples with partial beam absorption, corrections require the use of atomic stopping powers.

Water. Because of its importance in both chemical and biological systems, the radiation chemistry of water has been extensively studied and serves as an example of radiation-induced chemical change. A primary radiation interaction process may be represented by reaction (1), where H_2O^* represents

$$H_2O \rightarrow H_2O^+, e^-, H_2O^* \qquad (1)$$
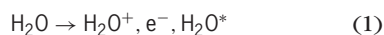
an electronically excited water molecule. The secondary electron (e^-), if formed with sufficient energy, will form its own trail of ionization and excitation. Within 10^{-10} to 10^{-8} s, reactions within spurs form hydrogen (H) atoms, hydroxyl (OH) radicals, hydrated electrons and molecular products, molecular hydrogen (H_2), and hydrogen peroxide (H_2O_2), as shown in reactions (2)–(7).

$$H_2O^+ + H_2O \rightarrow H_3O^+ + OH \qquad (2)$$

$$H_2O^* \rightarrow H + OH \qquad (3)$$

$$e^- + H_2O \rightarrow H_2O^- (e_{aq}^-) \qquad (4)$$

$$e_{aq}^- + H_3O^+ \rightarrow 2H_2O + H \qquad (5)$$

$$2H \rightarrow H_2 \qquad (6)$$

$$2OH \rightarrow H_2O_2 \qquad (7)$$

In pure water, radicals escaping the spurs undergo further radical-radical reactions and reactions with molecular products. In reactions such as (5) and the reverse of (3), molecular water is formed; and upon continuous irradiation, steady-state concentrations of H_2, H_2O_2, and smaller amounts of dioxygen (O_2) result and no further decomposition occurs.

Initial radical and molecular product yields depend upon the ionization density of the radiation and are a function of LET (**Fig. 1**). The initial G value for
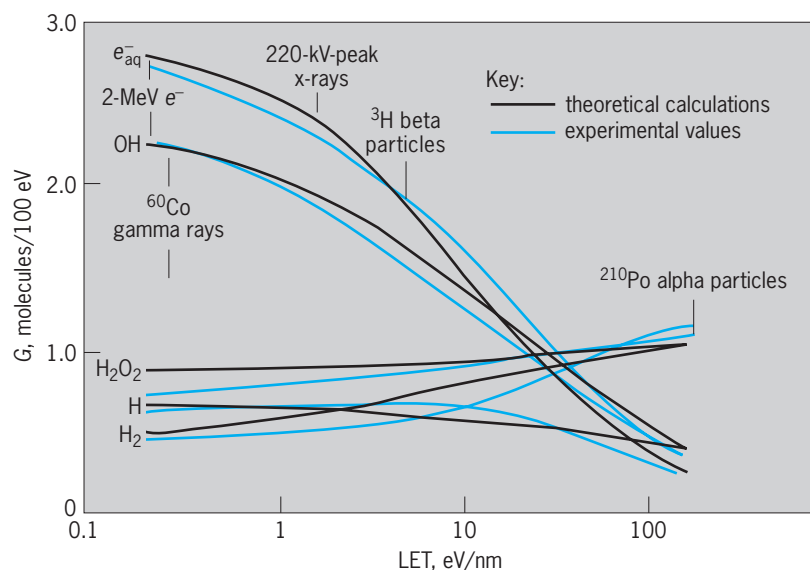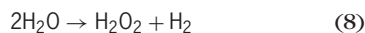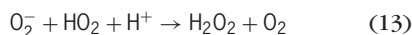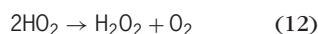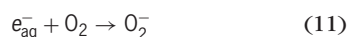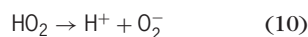
Fig. 1. Variation of *G* with linear energy transfer (LET) for the products solvated electron (e_{aq}^-), OH, H_2O_2, H, and H_2. (*After R. C. Cooper and R. W. Wood, eds., Physical Mechanisms in Radiation Biology, National Technical Information Service, 1974*)

water decomposition, essentially reaction (8), is 4.6

$$2H_2O \rightarrow H_2O_2 + H_2 \qquad (8)$$

molecules/100 eV for gamma rays and fast electrons and 2.9 for 12-MeV alpha particles.

In the presence of oxygen, the reducing species, e_{aq}^- and H, form the hydroperoxyl radical and superoxide ion, respectively, and are converted to H_2O_2, as shown in reactions (9)–(13). The hydroperoxyl

$$H + O_2 \rightarrow HO_2 \qquad (9)$$

$$HO_2 \rightarrow H^+ + O_2^- \qquad (10)$$

$$e_{aq}^- + O_2 \rightarrow O_2^- \qquad (11)$$

$$2HO_2 \rightarrow H_2O_2 + O_2 \qquad (12)$$

$$O_2^- + HO_2 + H^+ \rightarrow H_2O_2 + O_2 \qquad (13)$$

radical behaves as a weak acid [reaction (10)] in aqueous solution. In dilute aqueous systems containing oxygen, the spectrum of radical intermediates is markedly altered, and reactions of the OH radical become most important. This has significant consequences in irradiated biological systems exposed to the atmosphere.

In dilute aqueous solutions, radicals escaping spur recombinations can undergo reaction with the solute. *G* values for these radical yields have been determined from pulsed radiolysis or by a technique in which a specific radical is scavenged by an additive to form a stable molecule. The product is measured, for example, by spectrophotometry. Tetranitromethane [$C(NO_2)_4$] is an effective scavenger for the hydrated electron [reaction (14)]. The product,

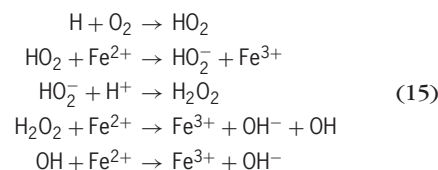$$C(NO_2)_4 + H_2O^- \rightarrow C(NO_2)_3^- + NO_2 + H_2O \qquad (14)$$
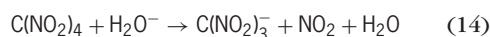
nitroform [$C(NO_2)_3^-$] is a stable anion with a well-known optical absorption spectrum.

The concept of yield for radicals surviving spur reactions is not a precise one. In addition to a physical LET dependence, it varies slightly with the scavenging system used. For gamma rays and fast electrons and in solutions of pH > 4, generally satisfactory correlations with radiolysis yields are obtained by using 2.8, 2.8, and 0.6 molecules, per 100 eV for e_{aq}, OH, and H radicals, respectively. Corresponding molecular yields for H_2 and H_2O_2 are 0.4 and 0.8 molecule per 100 eV.

In acidic solutions, e_{aq}^- is converted to H through reaction (4), and radical yields are 3.7 and 2.9 molecules per 100 eV for H and OH, respectively. The molecular yields are essentially unchanged.

Radiation chemical mechanisms in aqueous solutions are determined by correlations of chemical change with known radical yields. Oxidation of Fe(II) to Fe(III) in an aerated acid solution (the Fricke dosimeter solution) is an example of such a system. The following mechanism [reactions (15)] has

$$H + O_2 \rightarrow HO_2$$
$$HO_2 + Fe^{2+} \rightarrow HO_2^- + Fe^{3+}$$
$$HO_2^- + H^+ \rightarrow H_2O_2 \qquad (15)$$
$$H_2O_2 + Fe^{2+} \rightarrow Fe^{3+} + OH^- + OH$$
$$OH + Fe^{2+} \rightarrow Fe^{3+} + OH^-$$

been developed for radiolysis. The yield for Fe^{3+}, based upon these radical yields, is 15.6 molecules per 100 eV, consistent with the experimental value.

Applications. In addition to basic kinetics and mechanistic studies, the principles of radiation chemistry find application in any process in which ionizing radiation is used to study, treat, or modify a biological or chemical system.

In radiation therapy, tumors are destroyed by the application of ionizing radiation from external or internally administered sources. Either or both free-radical attack and specific local energy deposition may be involved in cell destruction. Gamma rays are used for treatment of internal tumors; electron or charged-particle beams are applied to external or invasively accessible lesions. Radiation LET is an important variable in treating tumors with particle beams.

The physiological concentration of iodine in the thyroid is the basis for the treatment of hyperthyroidism with [131]I. The beta radiation from this isotope is effective in localized tissue destruction.

A goal of any radiation therapy is maximum tumor cell destruction with minimum damage to healthy cells. Chemical sensitization toward, and protection of cells from, radiation damage are areas of continuing research. *See* RADIATION THERAPY.

Disposable medical supplies such as gloves, needles, syringes, and saline solutions are routinely sterilized by radiation. Doses of the order of 10^6 rads (10 kGy) are effective in deactivating most microorganisms.

Radiation processing of sewage sludge enables environmentally safe disposal or conversion to fertilizer. However, while chemically feasible, the process is not economically competitive.

Radiation chemistry and food preservation. Ionizing radiation in sufficient dose is lethal to microorganisms. This was observed shortly after the discoveries of radioactivity and x-rays and was early considered as a means to control the growth of microorganisms in foods.

The inactivation of a microorganism by radiation is a result of reaction by free radicals produced in the organism itself or in the surrounding medium, and by direct action of the radiation on the organism. While all cell constituents are susceptible to radiation damage, it is generally believed that the DNA components in the cell chromosomes are the most critical target. In a direct-action event, nucleic acid molecules are ionized or excited by a primary or secondary electron. A chain of reactions is initiated which may result in damage or death to the organism. In the cytoplasm of cells, which may contain 80% water, free radicals from the water are formed which diffuse to and react with the DNA. Initial chemical events involve radical site formation followed by strand breakage.

The more complex the organism, the greater the sensitivity to radiation damage. The larger DNA units provide a larger target for interaction with radiation and with the radiation-produced intermediates. Viruses, for example, have a greater resistance to radiation than do bacteria. The radiation dose required to kill or inactivate 90% of a specific microbial population is designated as D_{10}. While the effectiveness of a given radiation dose also depends upon the physical state and chemical composition of the surrounding medium, D_{10} values are within the range of 0.1–2 kGy for most bacteria but may be of the order of 10 kGy for some viruses. A dose of 5 kGy has been reported to decrease populations of *Salmonella*, *Staphylococcus aureus*, *Escherichia coli*, *Brucella*, and *Vibrio* by factors of at least 10^6.

The use of ionizing radiation for pathogen control has been approved by most governments for a wide range of foods. In general, limitations on dose have been specified for all products. The Food and Drug Administration (FDA) approved the use of radiation for pathogen control in meat and meat products. A maximum of 4.5 kGy was established for refrigerated meat and meat products and 7.5 kGy for the frozen. The actual dose used for preservation of a commercial product is determined by storage requirements and conditions.

Processing of commercial quantities of food supplies requires a source of stable intensity and a radiation of sufficient penetrating power to deposit energy throughout the product. Current and proposed facilities utilize a conveyor belt on which the product being sterilized is passed for a predetermined exposure time under a cobalt-60 (or cesium-137) gamma-ray source or an electron accelerator. With the accelerator, energetic electrons, for example, 10 MeV, are incident upon a metal film in which the electron energy is converted to the more penetrating x-radiation. With a copper converter, approximately 10% of incident 10-MeV electron energy is converted to x-radiation. If only a surface sterilization is desired, electron beams may be used directly. In the decay of cobalt-60, gammas of 1.17 and 1.33 MeV are emitted; a beta particle emitted in the same decay process is largely absorbed within the source and plays essentially no role in the sterilization process. Cesium-137 emits a single gamma of 0.67 MeV in energy but has a longer half-life of 30 years compared to 5.3 years for cobalt-60.

The radiation processing of food products on a commercial scale requires sources of sufficient intensity to treat large quantities of produce in an economic, brief time period. To achieve this, 10^4–10^6-curie quantities of cobalt-60 or electron accelerators of 10–100-kW beam power are required. While both facility types require extensive shielding for radiation protection, the electron accelerator can be shut off, but the radioisotope source will require more involved and strictly monitored access limitation.

Despite the physical feasibility and federal approval, concern over public acceptance has delayed the implementation of large-scale radiation processing of food supplies. The term "cold pasteurization" has been used in some cases to label food so treated.

Scientific concerns involving possible radiolytic formation of toxic products, destruction of nutritional components of the food, and a detrimental effect on taste or odor have been addressed in over 50 years of extensive research and testing. Radiolysis products in foods are in most cases similar but not identical to those resulting from cooking with heat or by microwave. In addition to water, the main components of foods are carbohydrates, lipids, and proteins. Extensive studies involving analyses for specific products formed in these materials have been carried out. Representative materials, such as starch, corn oil, and alanine, were subjected to radiation doses equal to and greater than that recommended for treatment. In carbohydrates, shorter-chain aldehydes, alcohols, and ketones are the major products of irradiation. From proteins, amino acids and NH_3 and non-nitrogen-containing simpler acids and aldehydes are formed. From lipids, hydrocarbons and aldehydes are major products as well as a variety of esters. Such studies have given no evidence for significant chemical toxins being formed in food subjected to sterilizing doses of radiation.

Most studies of toxicological safety of irradiated foods have involved in vitro experiments with animals. In such studies, an animal population is fed a diet containing a food subjected to a dose of radiation usually many times that proposed for sterilization and increased shelf-life requirements. Comparisons with populations fed a regular diet have demonstrated no adverse effects attributable to the irradiation.

Some vitamin loss does occur in irradiated food samples. Loss normally occurs in the cooking process, and this is sometimes enhanced by irradiation. Thiamine is particularly sensitive. The FDA has concluded that other dietary factors are largely unaffected in nutritional value.

Milk and most dairy products will develop an off-flavor from 10-kGy doses of radiation. Radiation will not be usable for their processing. No

significant tastes or odors have been reported for other foods subjected to sterilization or preservation doses. *See* FOOD PRESERVATION; ISOTOPIC IRRADIATION.   Francis J. Johnston

Pulse radiolysis. The chemical effects produced by the absorption of ionizing radiation involve, to a large extent, the production and reaction of free radicals, that is, molecules containing unpaired electrons. In most cases, these free radicals are very reactive and have very short lifetimes, so that they cannot be studied directly by the usual chemical approaches. However, pulse radiolysis methods can be very effectively employed to observe the production and subsequent reaction of free radicals. By using these methods, both chemical and instrumental, the properties of many radicals and other transient intermediates have been described in detail. In many cases, these properties are unavailable from more conventional experiments. A major fraction of pulse radiolysis studies have been carried out on radicals produced in aqueous solutions. Here, the interest is usually in the secondary radicals resulting from reaction of organic or inorganic substrates with the radicals produced in the radiolysis of water, that is hydrogen atoms, hydrated electrons, and hydroxyl radicals. The radiation chemical yields of these initial radicals are known quite well so that quantitative measurements are possible. Since excited states also play an important role in the radiolysis of organic systems, they can be studied by pulse radiolysis techniques. *See* FREE RADICAL.

Pulse radiolysis usually involves observing optical absorption changes following irradiation with nanosecond or microsecond pulses of megaelectronvolt electrons. Once the spectral characteristics of the transient molecules (free radicals) are established, measurements on the time dependence of their optical absorptions allow determination of the rate for their production and reaction. Changes in absorption spectra with time frequently allow details of the mechanisms involved in radical reactions to be elucidated. While optical absorption studies are usually carried out on the nanosecond or microsecond time scale following pulse irradiation, measurements can also be made at times as short as a few picoseconds or at millisecond and longer times. *See* ULTRAFAST MOLECULAR PROCESSES.

Spectroscopic methods. Electron spin resonance and laser resonance Raman spectroscopic methods are also employed in pulse radiolysis. Both approaches provide information on the structure of the transients that otherwise would not be available. Because these approaches are sensitive to the structure of the intermediate, they provide fingerprint identification of the intermediate being examined and allow discrimination against unwanted species. They can be an important adjunct to spectrophotometric experiments. However, very sophisticated equipment is required for such experiments. Conductivity measurements are usually used to provide information on the production of acid or base in oxidation or reduction processes. However, the background conductance of ions effectively restricts these studies to solutions of low ionic strength in the pH range 4–10. Nonpo-

lar media such as hydrocarbons can also be examined by time-resolved conductance methods. In this case, microwave methods are particularly attractive since electrodes are not necessary and polarization problems are absent. *See* ELECTRON PARAMAGNETIC RESONANCE (EPR) SPECTROSCOPY; RAMAN EFFECT.

High-energy electrons. Pulse radiolysis experiments are usually based on pulses of electrons having energies of a few megaelectronvolts from a Van de Graaff or linear accelerator. At these energies the electrons have sufficient penetration that it is possible to irradiate a sample of 1 cm thickness reasonably uniformly. Beam currents during the pulse are usually of the magnitude of amperes, so that a 1-nanosecond pulse of electrons having an energy of 2 MeV produces an initial radical concentration in water of a few micromolar. Many detection methods are sufficiently sensitive that this concentration is adequate for appropriate measurements. It is possible to produce considerably higher initial free-radical concentrations by using longer or more intense pulses, and to produce initial free-radical concentrations in excess of 10 millimolar by using the intense pulses from a commercially available field emission accelerator. However, it is generally desirable to limit the radical concentrations in order to avoid complicating reactions between radicals.

Electron transfer kinetics. Pulse radiolysis is particularly valuable in providing information on electron transfer kinetics and equilibria. In particular, pulse radiolysis allows direct observation of electron transfer reactions. Most data on electron-transfer rates and one-electron reduction potentials of free radicals have been derived from pulse radiolysis. The results obtained represent very significant additions to the information available on a type of reaction which is of considerable importance to many aspects of chemistry. A different approach does not involve free radicals but a determination of the rate for acid dissociation of naphthol (**Fig. 2**). In this case, hydroxide ion, which is a by-product of water radiolysis, attacks the naphthol to produce a naphtholate anion which has a greater absorption than does the naphthol.

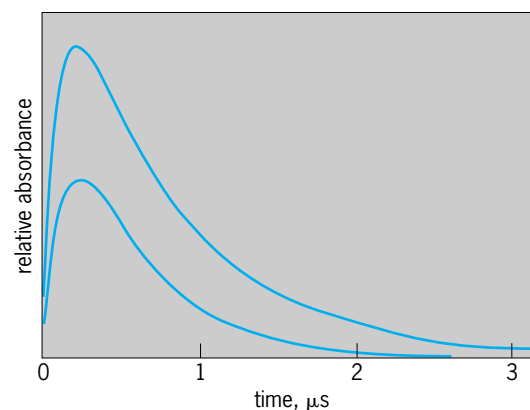

Fig. 2. Growth and decay of the signal of naphtholate anion in the pulse irradiation of a naphthol solution. Determination of the decay rate gives a spontaneous lifetime of 30 milliseconds for loss of the naphthol proton in aqueous solution.
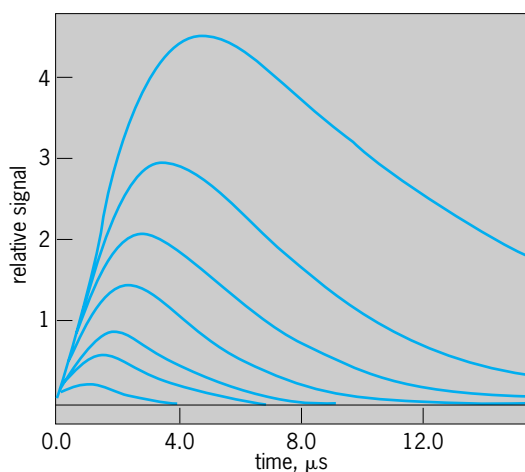
Fig. 3. Time dependence of the electron spin resonance signals of hydrogen atoms in the presence of 50–600-micromolar azide ion. The signal intensities and decay periods decrease with increasing azide ion concentration.

Observations on the recovery in acidic solution provides information on the rate for protonation of the phenolate anion and allows derivation of the spontaneous lifetime for acidic dissociation of the naphthol. Time-resolved electron spin resonance are similarly valuable in providing kinetic information on the reactions of free radicals. In optical studies, hydrogen atom signals are usually masked by other absorptions so that they cannot be observed directly. They are, however, uniquely observed in techniques (**Fig. 3**). Electron spin resonance measurements have provided a significant fraction of the available information on the reaction of hydrogen atoms with organic substrates. Determination of the rates for reaction of radicals with spin traps is another important application of time-resolved electron spin resonance measurements. Time-resolved Raman studies provide information on the vibrational structure of radicals, which is otherwise unavailable. Since the concentrations required for detection can be maintained for only brief periods, pulse techniques are required.   Robert H. Schuler

Bibliography. G. V. Buxton et al.,Critical review of rate constants for reactions of hydrated electrons, hydrogen atoms and hydroxyl radicals ($\cdot$OH/$\cdot$O$^-$) in aqueous solution, *J. Phys. Chem. Ref. Data*, 17(2):513–886, 1988; W. C. Dewey et al. (eds.), *Radiation Research: A Twentieth Century Perspective*, 1992; J. F. Diehl, *Safety of Irradiated Foods*, 1995; M. Ebert et al., *Pulse Radiolysis*, 1965; *Federal Register*, 62(232):64107–64121, 1997; R. H. Schuler, Three decades of spectroscopic studies of radiation produced intermediates, *Radiat. Phys. Chem.*, 43(5):417–423, 1994; J. W. T. Spinks and R. J. Woods, *An Introduction to Radiation Chemistry*, 3d ed., 1990; Y. Tabata, Y. Ito, and S. Tagawa (eds.), *Handbook of Radiation Chemistry*, 1991; P. Wardman, Reduction potentials of one-electron couples involving free radicals in aqueous solution, *J. Phys. Chem. Ref. Data*, 18(4):1637–1755, 1989; V. M. Wilkinson and G. W. Gould, *Food Irradiation:, A Reference Guide*, 1996.

Radiation damage to materials

Harmful changes in the properties of liquids, gases, and solids, caused by interaction with nuclear radiations. For a discussion of radiation damage in minerals. For a description of damage caused to biological systems by radiation *see* METAMICT STATE; RADIATION INJURY (BIOLOGY)

The interaction of radiation with materials often leads to changes in the properties of the irradiated material. These changes are usually considered harmful. For example, a ductile metal may become brittle. However, sometimes the interaction may result in beneficial effects. For example, cross-linking may be induced in polymers by electron irradiation leading to a higher temperature stability than could be obtained otherwise. *See* RADIOACTIVITY AND RADIATION APPLICATIONS.

Radiation damage is usually associated with materials of construction that must function in an environment of intense high-energy radiation from a nuclear reactor. Materials that are an integral part of the fuel element or cladding and nearby structural components are subject to such intense nuclear radiation that a decrease in the useful lifetime of these components can result. *See* NUCLEAR REACTOR.

Radiation damage will also be a factor in thermonuclear reactors. The deuterium-tritium (D-T) fusion in thermonuclear reactors will lead to the production of intense fluxes of 14-MeV neutrons that will cause damage per neutron of magnitude two to four times greater than damage done by 1–2 MeV neutrons in operating reactors. Charged particles from the plasma will be prevented from reaching the containment vessel by magnetic fields, but uncharged particles and neutrons will bombard the containment wall, leading to damage as well as sputtering of the container material surface which not only will cause degradation of the wall but can contaminate the plasma with consequent quenching. *See* NUCLEAR FUSION; PLASMA (PHYSICS).

Superconductors are also sensitive to neutron irradiation; hence the magnetic confinement of the plasma may be affected adversely. Damage to electrical insulators will be serious. Electronic components are extremely sensitive to even moderate radiation fields. Transistors malfunction because of defect trapping of charge carriers. Ferroelectrics such as $BaTiO_3$ fail because of induced isotropy; quartz oscillators change frequency and ultimately become amorphous. High-permeability magnetic materials deteriorate because of hardening; thermocouples lose calibration because of transmutation effects. In this latter case, innovations in Johnson noise thermometry promise freedom from radiation damage in the area of temperature measurement. Plastics used for electrical insulation rapidly deteriorate. Radiation damage is thus a challenge to reactor designers,
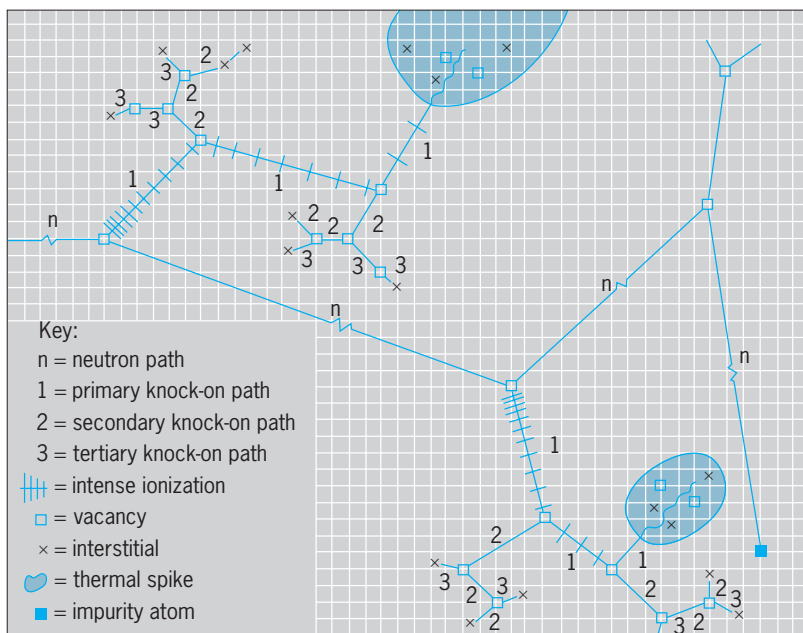
Fig. 1. The five principal mechanisms of radiation damage are ionization, vacancies, interstitials, impurity atoms, and thermal spikes. Diagram shows how a neutron might give rise to each in copper. Grid-line intersections are equilibrium positions for atoms. (*After D. S. Billington, Nucleonics, 14:54–57, 1956*)

materials engineers, and scientists to find the means to alleviate radiation damage or to develop more radiation-resistant materials.

Damage mechanisms. There are several mechanisms that function on an atomic and nuclear scale to produce radiation damage in a material if the radiation is sufficiently energetic, whether it be electrons, protons, neutrons, x-rays, fission fragments, or other charged particles.

Electronic excitation and ionization. This type of damage is most severe in liquids and organic compounds and appears in a variety of forms such as gassing, decomposition, viscosity changes, and polymerization in liquids. Rapid deterioration of the mechanical properties of plastics takes place either by softening or by embrittlement, while rubber suffers severe elasticity changes at low fluxes. Cross-linking, scission, free-radical formation, and polymerization are the most important reactions. *See* RADIATION CHEMISTRY.

The alkali halides are also subject to this type of damage since ionization plays a role in causing displated atoms and darkening of transparent crystals due to the formation of color centers. *See* COLOR CENTERS.

Transmutation. In an environment of neutrons, transmutation effects may be important. An extreme case is illustrated by reaction (1). The ^{6}Li isotope is ap-

$$^6\text{Li} + n \rightarrow \ ^4\text{He} + \ ^3\text{H} + 4.8 \text{ MeV} \tag{1}$$
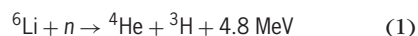
proximately 7.5% abundant in natural lithium and has a thermal neutron cross section of 950 barns (1 barn $= 1 \times 10^{-24}$ cm^2). Hence, copious quantities of tritium and helium will be formed. (In addition, the kinetic energy of the reaction products creates many defects.) Lithium alloys or compounds are consequently subject to severe radiation damage. On the

other hand, reaction (1) is crucial to success of thermonuclear reactors utilizing the D-T reaction since it regenerates the tritium consumed. The lithium or lithium-containing compounds might best be used in the liquid state.

Even materials that have a low cross section such as aluminum can show an appreciable accumulation of impurity atoms from transmutations. The capture cross section of ^{27}Al (100% abundant) is only 0.25×10^{-24} cm^2. Still reaction (2) will yield several per-

$$^{27}\text{Al} + n \rightarrow \ ^{28}\text{Al}^{\beta-} \ \xrightarrow{\ 2.3 \text{ min}\ } \ ^{28}\text{Si} \tag{2}$$
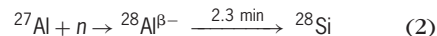
cent of silicon after neutron exposures at fluences of $10^{23}n/\text{cm}^2$.

The elements boron and europium have very large cross sections and are used in control rods. Damage to the rods is severe in boron-containing materials because of the ^{10}B(n,α) reaction. Europium decay products do not yield any gaseous elements. At high thermal fluences reaction (3) is most important in

$$^{58}\text{Ni} + n \rightarrow \ ^{59}\text{Ni} + n \rightarrow \ ^{56}\text{Fe} + \alpha \tag{3}$$
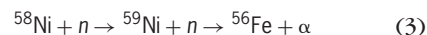
nickel-containing materials. The reaction $(n,n') \rightarrow \alpha$ at 14 MeV takes place in most materials under consideration for structural use. Thus, transmutation effects often can be a problem of great importance. *See* NEUTRON SPECTROMETRY; NUCLEAR REACTION; TRANSMUTATION.

Displaced atoms. This mechanism is the most important source of radiation damage in nuclear reactors outside the fuel element. It is a consequence of the ability of the energetic neutrons born in the fission process to knock atoms from their equilibrium position in their crystal lattice, displacing them many atomic distances away into interstitial positions and leaving behind vacant lattice sites. The interaction is between the neutron and the nucleus of the atom only, since the neutron carries no charge. The maximum kinetic energy ΔE that can be acquired by a displaced atom is given by Eq. (4), where M is mass of

$$\Delta E = \frac{4Mm}{(M + m)^2} \cdot E_N \tag{4}$$

the primary knocked-on atom (PKA), m is the mass of the neutron, and E_N is the energy of the neutron.

The energy acquired by each PKA is often high enough to displace additional atoms from their equilibrium position; thus a cascade of vacancies and interstitial atoms is created in the wake of the PKA transit through the matrix material. Collision of the PKA and a neighbor atom takes place within a few atomic spacings or less because the charge on the PKA results in screened coulombic-type repulsive interactions. The original neutron, on the other hand, may travel centimeters between collisions. Thus regions of high disorder are dispersed along the path of the neutron. These disordered regions are created in the order of 10^{-12} s. The energy deposition is so intense in these regions that it may be visualized as a temporary thermal spike.

Not all of the energy transferred is available for displacing atoms. Inelastic energy losses (electronic

excitation in metals and alloys, and excitation plus ionization in nonmetals) drain an appreciable fraction of the energy of the knocked-on atom even at low energies, particularly at the beginning of its flight through the matrix material. The greater the initial energy of the PKA, the greater is the inelastic energy loss; however, near the end of its range most of the interactions result in displacements. **Figure 1** is a schematic representation of the various mechanisms of radiation damage that take place in a solid.

A minimum energy is required to displace an atom from its equilibrium position. This energy ranges from 25 to 40 eV for a typical metal such as iron; the mass of the atom and its orientation in the crystal influence this value. When appropriate calculations are made to compensate for the excitation energy loss of the PKA and factor in the minimum energy for displacement, it is found that approximately 500 stable vacancy-interstitial pairs are formed, on the average, for a PKA in iron resulting from a 1-MeV neutron collision. By multiplying this value by the flux of neutrons [10^{14-15} n/(cm^2)(s)] times the exposure time [3×10^7 s/yr] one can easily calculate that in a few years each atom in the iron will have been displaced several times.

In the regions of high damage created by the PKA, most of the vacancies and interstitials will recombine. However, many of the interstitials, being more mobile than the vacancies, will escape and then may eventually be trapped at grain boundaries, impurity atom sites, or dislocations. Sometimes they will agglomerate to form platelets or interstitial dislocation loops. The vacancies left behind may also be trapped in a similar fashion, or they may agglomerate into clusters called voids. *See* CRYSTAL DEFECTS.

Effect of fission fragments. The fission reaction in uranium or plutonium yielding the energetic neutrons that subsequently act as a source of radiation damage also creates two fission fragments that carry most of the energy released in the fission process. This energy, approximately 160 MeV, is shared by the two highly charged fragments. In the space of a few micrometers all of this energy is deposited, mostly in the form of heat, but a significant fraction goes into radiation damage of the surrounding fuel. The damage takes the form of swelling and distortion of the fuel. These effects may be so severe that the fuel element must be removed for reprocessing in advance of burn-up expectation, thus affecting the economy of reactor operation. However, fuel elements are meant to be ultimately replaced, so that in many respects the damage is not as serious a problem as damage to structural components of the permanent structure whose replacement would force an extended shutdown or even reconstruction of the reactor. *See* NUCLEAR FISSION; NUCLEAR FUELS.

Damage in cladding. Swelling of the fuel cladding is a potentially severe problem in breeder reactor design. The spacing between fuel elements is minimized to obtain maximum heat transfer and optimum neutron efficiency, so that diminishing the space for heat transfer by swelling would lead to overheating of the fuel element, while increasing the spacing to allow for the swelling would result in lower efficiencies. A possible solution appears to be in the development of low-swelling alloys.

Damage in engineering materials. Most of the engineering properties of materials of interest for reactor design and construction are sensitive to defects in their crystal lattice. The properties of structural materials that are of most significance are yield strength and tensile strength, ductility, creep, hardness, dimensional stability, impact resistance, and thermal conductivity. Metals and alloys are chosen for their fabricability, ductility, reasonable strength at high temperatures, and ability to tolerate static and dynamic stress loads. Refractory oxides are chosen for high-temperature stability and for use as insulators. **Figure 2** shows relative sensitivity of various types of materials to radiation damage. Several factors that enter into susceptibility to radiation damage will be discussed. *See* METAL, MECHANICAL PROPERTIES OF.

Temperature of irradiation. Nuclear irradiations performed at low temperatures (4 K) result in the maximum retention of radiation-produced defects. As the temperature of irradiation is raised, many of the defects are mobile and some annihilation may take place at 0.3 to 0.55 of the absolute melting
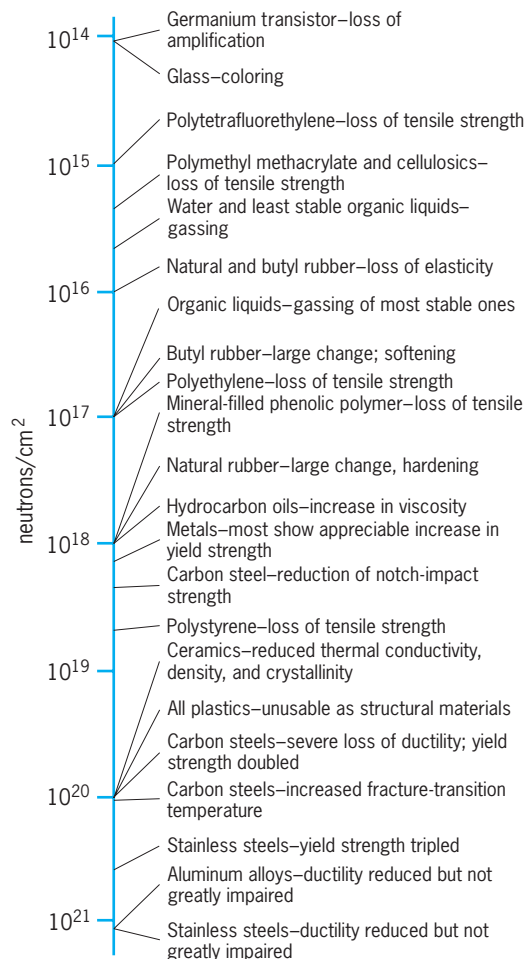

Fig. 2. Sensitivity of engineering materials to radiation. Levels are approximate and subject to variation. Changes are in most cases at least 10%. (*After O. Sisman and J. C. Wilson, Nucleonics, 14:58–65, 1956*)

point T_m. The increased mobility, particularly of vacancies and vacancy agglomerates, may lead to acceleration of solid-state reactions, such as precipitation, short- and long-range ordering, and phase changes. These reactions may lead to undesirable property changes. In the absence of irradiation many alloys are metastable, but the diffusion rates are so low at this temperature that no significant reaction is observed. The excess vacancies above the equilibrium value of vacancies at a given temperature allow the reaction to proceed as though the temperature were higher. In a narrow temperature region vacancy-controlled diffusion reactions become temperature-independent. When the temperature of irradiation is above $0.55T_m$, most of the defects anneal quickly and the temperature-dependent vacancy concentration becomes overwhelmingly larger than the radiation-induced vacancy concentration. However, in this higher temperature region serious problems may arise from transmutation-produced helium. This gas tends to migrate to grain boundaries and leads to enhanced intergranular fracture, limiting the use of many conventional alloys.

Nuclear properties. Materials of construction with high nuclear-capture cross sections are to be avoided because each neutron that is captured in the structural components is lost for purposes of causing additional fissioning and breeding. The exception is in control rods as discussed earlier. Moderator materials, in particular, need to have low capture cross sections but high scattering cross sections. Low atomic weight is an important feature since moderation of fast neutrons to thermal energies is best done by those elements that maximize the slowing-down process. [See Eq. (2).] Beryllium and graphite are excellent moderators and have been used extensively in elemental form. Both elements suffer radiation damage, and their use under high-stress conditions is to be avoided.

Fluence. The total integrated exposure to radiation (flux × time) is called fluence. It is most important in determining radiation damage. Rate effects (flux) do not appear to be significant. The threshold fluence for a specific property change induced by radiation is a function of the composition and microstructure. One of the most important examples is the appearance of voids in metals and alloys. This defect does not show up in the microstructure of irradiated metals or alloys until a fluence of 10^{19} n/cm^2 or greater has been achieved. Consequently, there was no way to anticipate its appearance and the pronounced effect in causing swelling in structural components of a reactor. This and other examples point to the importance of lifetime studies to establish the appearance or absence of any unexpected phenomenon during this time.

Lifetime studies in reactors are time-consuming and are virtually impossible if anticipated fluences far exceed the anticipated lifetime of operating test reactors. A technique to overcome this impasse is to use charged-particle accelerators to simulate reactor irradiation conditions. For example, nickel ions can be used to bombard nickel samples. The bombarding ions at 5 to 10 MeV then simulate primary knocked-on atoms directly and create high-density damage in the thickness of a few micrometers. Accelerators are capable of producing beam currents of several μA/cm^2; hence in time periods of a few hours to a few days ion bombardment is equivalent to years of neutron bombardment. Correlation experiments have established that the type of damage is similar to neutron damage. Moreover, helium can be injected to approximate n,α damage when these reactions do not occur in accelerator bombardments. However, careful experimentation is required to obtain correlation between results obtained on thin samples and thicker, more massive samples used in neutron studies. *See* CHARGED PARTICLE BEAMS; PARTICLE ACCELERATOR.

Pretreatment and microstructure. Dislocations play a key role in determining the plastic flow properties of metals and alloys such as ductility, elongation, and creep. The yield, ultimate and impact strength properties, and hardness are also expressions of dislocation behavior. If a radiation-produced defect impedes the motion of a dislocation, strengthening and reduced ductility may result. On the other hand, during irradiation, point defects may enhance mobility by promoting dislocation climb over barriers by creating jogs in the dislocation so that it is free to move in a barrier-free area. Moreover, dislocations may act as trapping sites for interstitials and gas atoms, as well as nucleation sites for precipitate formation. Thus the number and disposition of dislocations in the metal alloy may strongly influence its behavior upon irradiation.

Heat treatment prior to irradiation determines the retention of both major alloying components and impurities in solid solution in metastable alloys. It also affects the number and disposition of dislocations. Thus heat treatment is an important variable in determining subsequent radiation behavior.

Impurities and minor alloying elements. The presence of small amounts of impurities may profoundly affect the behavior of engineering alloys in a radiation field. It has been observed that helium concentrations as low as 10^{-9} seriously reduce the high-temperature ductility of a stainless steel. Concentrations of helium greater than 10^{-3} may conceivably be introduced by the n,α reaction in the nickel component of the stainless steel or by boron contamination introduced inadvertently during alloy preparation. The boron also reacts with neutrons via the n,α reaction to produce helium. The addition of a small amount of Ti (0.2%) raises the temperature at which intergranular fracture takes place so that ductility is maintained at operating temperatures.

Small amounts of copper, phosphorus, and nitrogen have a strong influence on the increase in the ductile-brittle transition temperature of pressure vessel steels under irradiation. Normally these carbon steels exhibit brittle failure below room temperature. Under irradiation, with copper content above 0.08% the temperature at which the material fails in a brittle fashion increases. Therefore it is necessary to control the copper content as well as the phosphorus and nitrogen during the manufacture and heat treatment of these steels to keep the transition temperature

at a suitably low level. A development of a similar nature has been observed in the swelling of type 316 stainless steel. It has been learned that carefully controlling the concentration of silicon and titanium in these alloys drastically reduces the void swelling. This is an important technical and economic contribution to the fast breeder reactor program.

Beneficial effects. Radiation, under carefully controlled conditions, can be used to alter the course of solid-state reactions that take place in a wide variety of solids. For example, it may be used to promote enhanced diffusion and nucleation, it can speed up both short- and long-range order-disorder reactions, initiate phase changes, stabilize high-temperature phases, induce magnetic property changes, retard diffusionless phase changes, cause re-solution of precipitate particles in some systems while speeding precipitation in other systems, cause lattice parameter changes, and speed up thermal decomposition of chemical compounds. The effect of radiation on these reactions and the other property changes caused by radiation are of great interest and value to research in solid-state physics and metallurgy.

Radiation damage is usually viewed as an unfortunate variable that adds a new dimension to the problem of reactor designers since it places severe restraints on the choice of materials that can be employed in design and construction. In addition, it places restraints on the ease of observation and manipulation because of the radioactivity involved. However, radiation damage is also a valuable research technique that permits materials scientists and engineers to introduce impurities and defects into a solid in a well-controlled fashion. *See* ION IMPLANTATION.

Douglas S. Billington

Bibliography. F. L. Bouquet, *Radiation Damage in Materials*, 4th ed., 1994; S.-H. Chen and M. Kotlarchyk, *Interaction of Radiation with Matter and Applications*, 1994; A. Holmes-Siedle and L. Adams, *Handbook of Radiation Effects*, 1993; J. Koutsky and J. Kocik, *Radiation Damage of Structural Materials*, 1994; A. S. Kumar et al., *Effects of Radiation on Materials: 16th International Symposium*, 1993; K. E. Stahlkopf and L. E. Steele (eds.), *Assuring Structural Integrity of Steel Reactor Pressure Boundary Components: Proceedings of the 5th International Seminar*, 1989; L. E. Steele (ed.), *Radiation Embrittlement of Nuclear Reactor Pressure Vessel Steels*, vols. 1–4, 1983–1993.

Radiation hardening

The protection of semiconductor electronic devices and electronic systems from the effects of high-energy radiation. Applications for such devices are in three major areas: (1) satellites, which are exposed to natural space radiation from the Van Allen belts, solar flares, and cosmic rays; (2) electronics, especially sensor and control electronics for commercial nuclear power-generating plants; and (3) perhaps most important, equipment designed to survive the radiation from nuclear explosions.

The expanding technology base has developed with emphasis shifting from one particular type of radiation to another. An early emphasis on neutron fluence and displacement damage effects was followed by a concentration on dose-rate effects from short x-ray and gamma-ray pulses. Next came a large effort to deal with the effects of electromagnetic pulses (EMP) from nuclear weapons on electronic systems and components. The preponderant utilization of complementary metal-oxide-semiconductor (CMOS) devices for satellite electronics increased concern about ionizing-radiation response. This has been incorrectly described as total-dose response, but it usually depends strongly on dose rate as well as total accumulated ionizing dose. The evolution to very large scale integrated-circuit (VLSI) devices has reduced the active volume of devices to levels where a single cosmic ray can produce sufficient ionization to change the logic state of a gate or flip-flop circuit; and this phenomenon, called single-event upset, has aroused strong concern. *See* ELECTROMAGNETIC PULSE (EMP); INTEGRATED CIRCUITS.

Although most radiation effects have been explained and are well understood, much remains to be done. Among the better-understood phenomena are displacement damage from neutrons and photocurrent transients produced by ionizing-radiation pulses. Basic electromagnetic-pulse interactions are also well understood, although their effects on complex electronic systems are extremely difficult to predict because such prediction requires that points of entry and transfer functions be characterized, which is an extremely complex boundary-value problem.

Ionizing dose is discussed with respect to rads as the customary unit of absorbed energy. One rad (material) is the absorption of 100 ergs per gram (10^{-2} joule per kilogram or 10^{-2} gray). Similarly, dose rate is measured in rads (material) per second or ergs per gram per second. Radiation effects depend on both the accumulated ionizing dose (so-called total dose) and the rate at which the dose was accumulated. *See* UNITS OF MEASUREMENT.

Ionizing dose effects. The quasipermanent effects of exposure to ionizing radiation are least understood, and understanding the response of semiconductor devices to this radiation is probably the single most important remaining radiation-hardening problem. Emphasis has been on electrically measurable manifestations such as accumulation of positive charge in silicon dioxide and buildup of negative charge at the oxide–semiconductor interface. The chemical or physical nature of traps has not been adequately explained, although many theories have been set forth. Hardening of metal-oxide-semiconductor (MOS) devices has been accomplished by lower-temperature processing, which probably reduces physical or crystalline defects, and by developing extremely clean processes, which probably reduce chemical defects. The dearth of basic knowledge about damage from ionizing radiation is so serious that presently known electrical, chemical, or physical measurements on semiconductor devices cannot be used to predict behavior in

an ionizing-radiation environment. *See* CRYSTAL DEFECTS; TRAPS IN SOLIDS.

The problem of creating hardened systems for use in ionizing-radiation environments is further exacerbated by the common practice of specifying the environment only in terms of total dose. Semiconductor devices do not have a unique response to total dose; rather they show a complex time-dependent response, which depends on both dose rate and accumulated dose. The damage depends on operating bias and temperature. Thus, it is necessary to specify the environment in terms of dose, dose rate, and temperature.

The most important electrical manifestations of ionizing dose damage in MOS devices are an increase in leakage current; a shift in threshold voltage; and a decrease in speed, transconductance, and channel conductance. The threshold voltage shift is predominantly negative for unhardened devices, and negative at low doses followed by positive at high doses for hardened devices.

Dose-rate effects. Dose-rate effects are well understood. The generation of electron-hole pairs in semiconductors is proportional to the dose; the generation constant for silicon is 4.3×10^{13} electron-hole pairs per rad (4.3×10^{15} pairs per gray). Carriers generated in or near a *pn* junction result in a transient photocurrent proportional to dose rate and the effective volume of the junction. Effective junction volume is the product of junction area and junction width plus a diffusion length on each side of the junction. High dose rates can damage semiconductor devices through logic upset, latch-up, and burnout. In logic upset, logic and memory cells are switched. In latch-up, susceptible devices are triggered into a high current–high voltage state via silicon controlled rectifier action, producing circuit incapacitation and in some cases burnout. Finally, the photocurrents can cause enough energy dissipation, primarily in *pn* junction regions, to cause device burnout. *See* PHOTOVOLTAIC EFFECT.

Displacement damage effects. Displacement damage effects are caused by neutrons, protons, electrons, and other high-energy particles. The production of lattice defects is proportional to the nonionizing energy absorbed by the lattice. The irradiation itself produces Frenkel defects, which are unstable at room temperature and diffuse to produce room-temperature-stable defects, including divacancies, E-centers (donor vacancy defects), and A-centers (oxygen vacancy defects). The defects act as recombination and trapping centers, producing a minority-carrier lifetime decrease proportional to particle fluence. A second-order effect is an increase in resistivity in silicon semiconductors, which is caused by a decrease in mobility and majority-carrier concentration. The dominant effect in bipolar silicon devices is a reduction in common-emitter current gain. This effect is quantitively described by the Messenger-Spratt equation, which shows that the increase in reciprocal common-emitter current gain is directly proportional to particle fluence and inversely proportional to gain-bandwidth product. *See* ELECTRON-HOLE RECOMBINATION.

Single-event phenomena. Single-event upsets are caused at a very low rate in logic and memory circuits by cosmic rays. Rates are low enough (less than 10^{-3} upset per bit per day) that error-detection-and-correction (EDAC) software can be effectively used. This problem is a result of the extremely small effective volumes of large-scale integrated circuits. Integrated circuits with feature sizes greater than 10 micrometers do not show significant single-event upset rates. The upset rate is a function of the maximum linear dimension of the critical device volume, is inversely proportional to the area of the collecting junction, and is proportional to the flux of cosmic-ray particles that exceed the critical linear energy transfer required to produce critical charge in logic devices. This charge is the charge required to change a 0 to a 1 or vice versa, whichever is lower. Single-event upsets may trigger latch-up or second breakdown and hence may also cause catastrophic failure.

Single-event phenomena have led to the development of fault-tolerant architectures to mitigate the effects of random digital upsets, and to the design modification of integrated circuits to prevent an upset even when the active volume is struck by a cosmic ray. Single-event research and development programs are expected to become increasingly important since the problem becomes progressively more serious as integrated circuits utilize smaller active volumes, allowing cosmic rays of lower energy to produce upsets. *See* FAULT-TOLERANT SYSTEMS.

System hardening. Hardening of electronic systems is accomplished by a combination of selecting hardened components, designing circuits more tolerant to radiation-induced degradations, and shielding. Shielding is effective against x-rays and electrons, which have short ranges in materials, and relatively ineffective against gamma radiation, neutrons, and cosmic rays, which have long ranges in materials. Shielding is also effective in mitigating the effect of electromagnetic pulse. *See* RADIATION SHIELDING.

SDI environments. Another major radiation problem results from Strategic Defense Initiative (SDI) environments. Damage effects could result from high-energy-beam weapons, which include lasers, microwave beams, and neutral-particle beams. The frequency range of electromagnetic pulse and the energy range of atomic particles would be extended orders of magnitude beyond the limits thus far considered. Thus, a major research and development program in radiation hardening will be necessary to support the SDI program. *See* RADIATION DAMAGE TO MATERIALS; SEMICONDUCTOR.   George C. Messenger

Bibliography. W. R. Dawes, Jr., *Hardening Semiconductor Components Against Radiation and Temperature*, 1990; R. N. Ghose, *EMP Environments and System Hardness Design*, 1984; F. Larin, *Radiation Effects in Semiconductor Devices*, 1968; T. Ma and P. Dressendorfer, *Ionizing Radiation Effects in MOS Devices and Circuits*, 1989; G. Messenger and M. Ash, *The Effects of Radiation on Electronic Systems*, 2d ed., 1992; Summaries of Annual Conferences on Nuclear and Space Radiation Effects, *IEEE Trans. Nucl. Sci.*, December issues.

Radiation injury (biology)

The harmful effects of ionizing radiation on plant and animal cells and tissues. Although the mechanisms of these changes are still poorly known, a wealth of information has been acquired about them. *See* GAMMA RAYS; X-RAYS.

In multicellular organisms, effects on cells are complicated by the interaction of injured and intact cells. Consequently, proper understanding of radiation injury in such organisms calls for appreciation not only of the reaction of individual cells but also of groups of cells in organs and tissues.

Effects of Radiation on Cells

Radiation is known to alter the genetic apparatus of the cell, to interfere with cell division, and to cause cell death.

Effects on genes and chromosomes. The most important action of radiation on the cell is the production of changes in the genetic or chromosomal apparatus. Once established, these changes are largely irreversible and, because of the strategic importance of each gene, may alter the fate of the cell and of its progeny as well. The changes include mutations and chromosome breaks.

Mutation, or gene alteration, consists of a change in the nucleic acid that constitutes the gene. Once established, it may be reversed or otherwise altered only by further mutation. A mutation induced by irradiation is essentially indistinguishable from a natural one. *See* MUTATION.

Chromosome breakage is visible microscopically as a break in the continuity of the chromosomal fiber. The ends of the broken chromosome often reunite within a few minutes, restituting the original structure of the chromosome; but they may unite with fragments of other chromosomes to cause various genetic rearrangements (**Fig. 1**), or they may not rejoin at all.

Radiation may also cause chromosome stickiness and clumping, which presumably affect the chromosomal surface, and may cause chromosomes to adhere to one another and fail to separate normally at cell division. *See* CHROMOSOME ABERRATION.

Effects on cell division. Relatively small amounts of radiation postpone cell division, the delay varying with the dose. When division is eventually resumed in a population of affected cells, the percentage dividing may temporarily exceed, or overshoot, normal; this is often associated with extensive degeneration of the dividing cells. If a large enough dose is administered, the proliferative ability of the cell is permanently abolished. *See* MITOSIS.

Killing of cells. Cells differ markedly in their susceptibility to radiation-induced death, but any cell may be killed if it receives enough radiation. In general, susceptibility varies in proportion to the rate of cell division. Few rapidly dividing mammalian cells are able to survive 5 grays (Gy) of x-rays. (A gray is a unit of absorbed radiation dose, 1 gray equaling 1 joule per kilogram of absorbing tissue.) Although radiation death may occur during or immediately after irradiation, more often it ensues when the cell
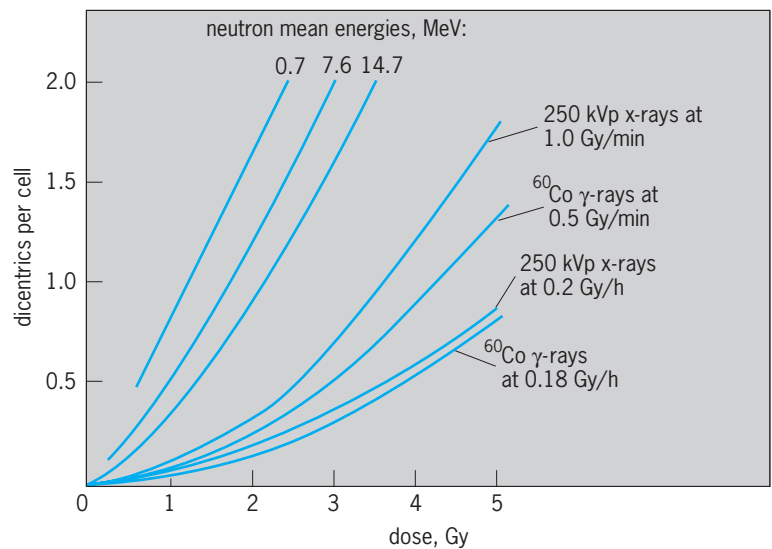

Fig. 1. Frequency of dicentric chromosome aberrations (dicentrics) in human lymphocytes irradiated in the laboratory as a function of the dose and dose rate of neutrons, x-rays, and gamma rays. kVp = kilovolt peak. (*From D. C. Lloyd and R. J. Purrott, Chromosome aberration analysis in radiological protection dosimetry, Radiat. Protect. Dosimetry, 1:19–28, 1982*)

attempts to divide or after it has divided several times. The relation between the percentage of cells surviving and the radiation dose implies that the dose required to kill all cells in a given population varies with the total number of cells in that population. Hence, although the median lethal dose per cell may be less than 1 Gy, several hundred grays may be required to kill all the billions of cells in a large organ or tumor. This is attributable to the spatial distribution of radiation-induced ion pairs in the absorbing medium, the probability of all cells being appropriately affected decreasing as the total number of cells irradiated is increased.

The mechanism by which radiation kills cells is unknown, but the high radiosensitivity of dividing cells, with their tendency to die during cell division, suggests that the most critical type of injury involves the reproductive or chromosomal apparatus of the cell. This is also suggested by the fact that many irradiated cells rendered incapable of division retain their ability to differentiate into more highly specialized forms and to synthesize nucleic acid and protein. Enormous amounts of radiation (many hundred grays) are generally required to stop metabolic activity in such nondividing cells.

Effects of Radiation on Animal Tissues

Tissues differ widely in radiosensitivity, but in general their susceptibility to radiation varies with the rate at which their component cells divide. Accordingly, the most radiosensitive tissues of the body, composed of cells that divide rapidly, are the blood-forming organs, gonads, skin, and intestine.

Radiation injury is not, however, limited to the killing of radiosensitive cells. Invariably, the initial cellular destruction leads to secondary disturbances and reparative processes, often through systemic mechanisms, which modify the primary lesion. In this respect, radiation injury simulates other types of injury and is not unique or specific.

Blood-forming tissues. Since circulating blood cells live only a few days or weeks, they must be replaced continually. The new cells are produced in the bone marrow, spleen, and lymphoid organs through division of blood-cell precursors, which are highly radiosensitive and hence readily destroyed by radiation. Because destruction of these cells leads to shortage of mature blood cells, which may have drastic consequences, damage to the blood-forming organs is one of the most important types of radiation injury. *See* HEMATOPOIESIS.

If only a part of the body is irradiated, blood-forming cells from nonirradiated tissue are carried by the bloodstream to the damaged organs, where they multiply and replace destroyed cells. Consequently, greater quantities of radiation are tolerated if applied to part of the body than if applied to the whole.

The rate of fall in blood count after irradiation varies with the type of blood cell in question, the species, and the radiation dose. In general, the number of cells is depleted more rapidly in small laboratory animals than in humans, the depletion being accelerated in any given species with increasing dose of radiation. In the following remarks, the blood changes are characterized according to their time sequence after a single exposure to whole-body radiation in humans.

The lymphocyte is the first blood cell to be depleted. Diminution in the number of these cells is usually evident within only a few hours after irradiation and becomes maximal within several days. One reason for this may be the relatively short lifespan of such cells; another is the high radiosensitivity of lymphocytes, which is puzzling since they are not rapidly dividing cells. An added factor causing lymphocyte destruction is the nonspecific stress reaction of the body in response to irradiation. Minute amounts of radiation less than 10 millisieverts (mSv); the sievert is a standard unit of radiation dose] cause injured lymphocytes to appear in the blood. The loss of lymphocytes is believed to be responsible to some extent for the depression of immunity that occurs soon after intensive irradiation and that predisposes the affected individual to infection.

The granulocyte, another type of white blood cell, disappears more slowly from the bloodstream. Paradoxically, the number of such cells in the blood may be increased for several days after irradiation, owing to a constitutional reaction of the body to stress. Since mature granulocytes are nondividing cells and are relatively radioresistant, the gradual depletion of their number reflects predominantly the loss of aging cells from the total population, the maximal depression of the granulocyte count not being reached until 3–5 weeks after exposure. Because these cells constitute the body's most important defense against bacterial invasion, infection is a common cause of death in individuals that have been heavily irradiated.

Simultaneously with the fall in granulocyte count, the number of circulating blood platelets also declines, maximal deficiency of these corpuscles also occurring 3–5 weeks after irradiation. Depression of the platelet count below a certain level leads to hemorrhage. The bleeding, which may be either localized or generalized, varies in severity and can be fatal.

The red blood cell count falls gradually in the absence of hemorrhage, severe anemia occurring 5 or more weeks after irradiation. The anemia results primarily from underproduction of new red blood cells owing to destruction of precursors in the bone marrow. Although red-blood-cell-forming elements are usually among the first to regenerate after radiation injury, regeneration may be delayed, and the anemia may therefore be severe and prolonged, even in the absence of bleeding. *See* BLOOD.

Skin. The response of the skin to radiation has been studied extensively, because the skin is the most exposed structure of the body and because its radiosensitivity at one time seriously limited the amount of radiation that could be applied to underlying tumors. The extent of skin injury varies from transient reddening, a reaction formerly used to measure dose in radiotherapy, to ulceration and sloughing. Development of injury is characteristically slow, maximal damage not being evident until weeks after irradiation. Months later, permanent effects consisting of thinning of the skin, scarring of the underlying connective tissue, and dilation of cutaneous blood vessels may gradually appear. This sequence of changes, known collectively as radiation dermatitis, predisposes to subsequent development of skin cancer.

Injury of accessory skin structures is manifested in loss of the hair. The epilation, or loss of hair, appears several weeks after irradiation and may be temporary or permanent, depending on the amount of radiation absorbed.

Gastrointestinal tract. Because the dividing cells lining the small intestine are extremely radiosensitive, relatively small amounts of radiation applied to the abdomen elicit profound effects. These vary from slight disturbances of motility and secretion to ulceration and sloughing of the lining of the bowel.

Nausea and vomiting usually ensue within a few hours after exposure to doses above 2 Gy. After several times this much radiation, sloughing of the intestinal lining may lead to ulceration, intractable diarrhea, dehydration, and invasion of the bloodstream by bacteria that normally inhabit the lumen of the bowel. This sequence is usually fatal and constitutes one of the major causes of death after massive irradiation of the whole body.

Gonads. Since the developing germ cells are highly radiosensitive, their irradiation may result in sterility. Sterility does not usually occur immediately, however, owing to survival of more radioresistant mature eggs or sperm, but only after the supply of these cells is exhausted. Even then, sterility is transitory unless too few precursors survive to resume adequate production of germ cells. In humans, as in most other mammals studied, permanent sterilization requires amounts of radiation that are lethal when absorbed by the entire body; hence sterility is not generally a complication of whole-body irradiation.

Apart from killing of germ cells, mutations may be induced by radiation and be passed on via the

eggs and sperm of successive generations of progeny. The risk of these genetic disturbances is thought to be inceased by even the minutest amounts of radiation, and is therefore considered to be one of the limiting factors in the radiation dose permissible for humans.

Eye. The part of the eye most easily injured by radiation is the lens, opacification of which (cataract formation) has been observed after acute exposure to as little as 2 Gy of x-rays. Smaller doses of neutrons are estimated to have caused cataracts, examples of which have been noted among early neutron physicists. The cornea, conjunctiva, and the retina withstand much more radiation. The retina, however, is highly radiosensitive early in its embryonic development. Minute amounts of ionizing radiation are visible through radiochemical reactions in the retina, which are harmless. *See* EYE (VERTEBRATE).

Nervous system. Although the developing nervous system is highly radiosensitive, only relatively large amounts of radiation will kill nerve cells in the adult. Even in the adult, however, transitory functional disturbances may be elicited by relatively low doses, and after intensive exposure of the brain to a dose in excess of 50 Gy, incapacitating neurological effects may lead to death within minutes or hours. *See* NERVOUS SYSTEM (VERTEBRATE).

Bones and teeth. A dose as low as a few grays applied to bone- and tooth-forming cells in infancy or early childhood cause disturbances of dentition and skeletal growth. In contrast, mature bones and teeth are relatively radioresistant. Large amounts of radiation, however, such as may accumulate from locally deposited radioisotopes or from the treatment of cancer, produce demineralization and necrosis of bone that can lead to fractures, loosening of the teeth, bone cancer, and other complications. *See* BONE.

Vascular system. Transitory dilation of blood vessels, causing erythema or reddening of the skin, is one of the earliest known reactions to ionizing radiation. It occurs after only a few grays and may be accompanied by increased permeability of blood capillaries. Larger amounts of radiation may severely injure or kill cells that line the walls of blood and lymph vessels, giving rise to rupture, occlusion, permanent dilation, or scarring of vessels. Adverse effects on the blood supply may lead to secondary effects, such as metabolic disturbances and atrophy, in involved tissues. *See* CARDIOVASCULAR SYSTEM; LYMPHATIC SYSTEM.

Endocrine glands. The glands of internal secretion have traditionally been regarded as radioresistant because of their ability to withstand relatively large amounts of radiation without developing morphological lesions. There is growing evidence, however, that smaller doses (<10 Gy) may elicit changes in endocrine function and may, in some instances, induce lasting functional impairment. Apart from radiation injury itself, the endocrine system's adaptational response to acute effects of radiation resembles its response to other types of stress. *See* ENDOCRINE SYSTEM (VERTEBRATE).

Urinary system. The kidney and lower urinary tract are relatively radioresistant. Doses in excess of 10–20 Gy may, however, cause gradually progressive scarring and atrophy of the kidney, which can lead to fatal loss of renal function. *See* URINARY SYSTEM.

Lungs. Although relatively radioresistant, the lungs may be injured by intensive irradiation. The result is a chronic, pneumonialike disease, with scarring of lung tissue and blood vessels. Another complication is cancer of the lung, which has been noted in miners of radioactive ore and other irradiated populations.

Effects of Whole-Body Irradiation

When the entire body is irradiated, killing of cells in the various radiosensitive organs causes a complex series of disturbances, the predominant signs and symptoms of which (**Table 1**) are referable to injury of the intestinal tract, blood-forming organs, and skin (acute radiation syndrome). The injury, if severe enough, is fatal within 30 days in most laboratory animals, but in humans death may be delayed until the second month after irradiation.

Susceptibility to death varies from species to species, the average median lethal radiation dose for mammals being about 5 Gy (**Table 2**). As with most other lethal agents, radiation in doses below a certain minimum threshold causes negligible mortality, but radiation in doses above a maximum level is uniformly lethal (**Fig. 2**).

TABLE 1. Symptoms of acute radiation syndrome*			
Time after exposure	Supralethal dose, 10 Gy	Median lethal dose, 3–5 Gy	Sublethal dose, 2 Gy
First week Second week	Nausea and vomiting, first day Continued nausea, vomiting, diarrhea, fever, inflammation of throat, prostration, emaciation, leading to death	Nausea and vomiting, first day	
Third week		General malaise, loss of appetite, loss of hair, hemorrhage, pallor, diarrhea, fever, inflammation of throat, emaciation leading to death in 50% of victims	Loss of appetite, loss of hair inflammation of throat, pallor, hemorrhage, diarrhea; recovery begins; no deaths in absence of complications

*After United Nations Scientific Committee on the Effects of Atomic Radiation, *Sources, Effects and Risks of Ionizing Radiation: Report to the General Assembly with Annexes*, 1988.

TABLE 2. Median lethal dose of x-rays for various species of organisms*

Organism	Median lethal dose, Gy
Viruses	
Tobacco mosaic	2000
Rabbit papilloma	1000
Bacteria	
? *coli*	50
? *mesentericus*	1300
Algae	
Mesotenium	85
Pandorina	40
Protozoa	
Colpidium	3300
Paramecium	3000
Vertebrates	
Goldfish	7.5
Mouse	4.5
Rabbit	8.0
Rat	6.0
Monkey	4.5
Human	4.0(?)

*After R. Paterson (ed.), *The Treatment of Malignant Disease by Radium and X Rays,* E. Arnold, 1948.

The cause of death varies with the species, dose, dose rate, and type of radiation. The predominant cause of death in most mammals is injury of blood-forming tissues, which results in infection, hemorrhage, and anemia. Lethal injury of the intestine, usually requiring somewhat higher doses than lethal injury of the marrow, is another major cause of death. In this case, death results from diarrhea, dehydration, loss of body salts, and massive bacterial invasion from the lumen of the bowel. With even larger doses of radiation, death may ensue rapidly from injury of the brain, as mentioned earlier. When the latter mode of death is prevented by shielding the brain, other lethal mechanisms are encountered. Hence, it would appear that sufficient injury of any one of a variety of organs is potentially lethal.

Individuals surviving the acute radiation syndrome usually appear outwardly to have recovered and to be normal by the sixth month after irradiation. That such individuals are not fully recovered, however, is indicated by their subsequent increase in mortality from cancer.

Radiation-Induced Cancer

It is paradoxical that ionizing radiation, which is effective in the treatment of cancer, should also be capable of causing cancer. As such, however, it is but one of many agents, including ultraviolet rays, viruses, and a variety of chemicals, that are known to have cancer-inducing potency. The earliest known example of radiation-induced cancer was reported in 1902, less than 10 years after the discovery of x-rays. Since then, numerous instances have been observed in humans, and the process has been studied extensively in experimental animals. *See* ONCOLOGY.

Although susceptibility to cancer induction varies widely among species and organs, virtually all types of cancer have been induced experimentally, and it is clear that all types of ionizing radiation share cancer-forming potency. The radiation-induced cancers that have been noted most often in humans are leukemia and cancers of the skin, breast, lung, thyroid, stomach, colon, and bone.

Cancer of the skin was encountered early in the twentieth century as a complication of radiation dermatitis, developing on the hands and fingers of many of the pioneer radiologists. This effect resulted from prolonged manipulation of radiation equipment, the hands being exposed to large cumulative doses of radiation in the era before the attendant hazards were suspected. Although this disease claimed the lives of more than 100 of the first workers in radiology, it is no longer an occupational threat because of safeguards.

Another malignant disease encountered with unusual frequency in radiologists is leukemia. This disease is also abnormally prevalent in other populations exposed to radiation, such as the Japanese atomic bomb survivors (**Table 3**).

Radiation-induced bone cancer was first noted in painters of luminous watch dials who inadvertently ingested toxic quantities of radium-containing paint by habitually drawing their paint brushes to a point between their lips. Deposition of the radium in the skeleton, where it delivered relatively large doses of radiation to small foci of bone, led to the development of skeletal cancer. Osseous tumors are also abnormally prevalent in those who have consumed radium for medicinal purposes.

Cancer of the lung in miners employed in the pitchblende mines of Saxony, traditionally the predominant cause of death among these workers, has been attributed to their prolonged exposure to radon, which is present in high concentrations in the air of such mines (2–200 becquerels per liter). It is noteworthy, however, that other complicating factors also may have contributed to the effects of radiation in causing lung cancer in this population. *See* RADON.

Cancer incidence versus radiation dose. Although irradiation increases the incidence of many types of cancer, the precise relation between incidence and dose is not known for any type of cancer, especially
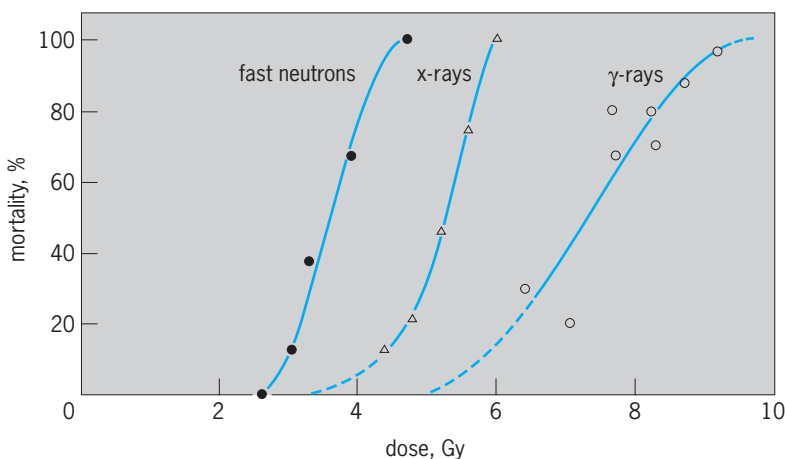

Fig. 2. Thirty-day mortality of mice exposed acutely to x-rays, fast neutrons, and gamma rays. (*After A. C. Upton et al., Radiation Research, vol. 4, Academic Press, 1956*)

TABLE 3. Mortality from leukemia in atomic bomb survivors of Hiroshima and Nagasaki, 1950*

Radiation dose to bone marrow, Gy	Number of survivors	Number of leukemias	Relative risk[†]
0	35,290	61	1.0
0.01–0.05	19,740	33	1.1
0.06–0.09	4,059	5	1.4
0.10–0.19	5,210	11	1.7
0.20–0.40	6,375	23	2.7
0.50–0.99	3,042	24	4.6
1.0–1.99	1,578	24	8.2
2.0–2.99	412	15	13.4
3.0–3.99	130	2	18.9
4.0 +	155	4	25.0

[*] After Y. Shimizu, H. Kato, and W. J. Schull, *Life Span Study Report 11,* Part 2: *Cancer Mortality in the Years 1950–1985 Based on the Recently Revised Doses* (DS86), RERF TR 5–88, Radiation Effects Research Foundation, 1988.
[†] Fitted relative risk, values rounded.

at low radiation dose levels. The yield of tumors per unit dose cannot, therefore, be accurately estimated, nor can it be assumed that any amount of radiation, however small, will increase the cancer rate. At the same time it cannot be assumed that a threshold dose exists below which no cancers are induced. For most experimentally induced cancers the incidence is not linearly proportional to dose, and the evidence suggests that radiation is appreciably less effective at low dose levels than at high dose levels. Paradoxically, moreover, irradiation actually reduces the frequency of certain types of tumors in experimental animals. Accordingly, no generalization about the relation between cancer incidence and radiation dose can be applied to all types of cancer. Estimates of lifetime risks of the major types of radiation-induced cancer in humans are shown in **Table 4**.

Effect of constitutional variables. Susceptibility to cancer induction in experimental animals varies widely with constitutional variables, such as genetic inheritance, age at irradiation, sex, and hormonal activity. Furthermore, the influence of these variables on susceptibility to tumor formation differs depending on the type of tumor and the variables in question. As yet, little is known about the action of these variables on susceptibility in experimental animals and almost nothing about their action in humans.

Mechanisms of cancer induction. The essential change caused by radiation that leads ultimately to the development of cancer is unknown, as is the na-

ture of the cancer change itself. There is growing evidence, however, that the transformation of a normal cell into a cancer cell entails a series of alterations and is not merely a one-step process. This is suggested by the relatively long induction period intervening between irradiation and the onset of malignant growth, a period averaging 5–20 years in humans, depending on the type of cancer induced. It is also suggested by the tendency of naturally occurring cancer to develop late in life and to become progressively more frequent with advancing age.

One theory of cancer formation ascribes the cancer change to one or more mutations in the cell that predispose it to unrestrained growth, perhaps by activating oncogenes or by inactivating tumor-suppressor genes. According to this view, the cancer-forming action of ionizing radiation stems from its mutagenic potency. In some situations, radiation appears to promote the development of cancer through other mechanisms as well, including the activation of latent cancer viruses. *See* CANCER (MEDICINE); ONCOGENES.

Effects on the Embryo and Fetus

Radiation injury of human embryos has been amply documented, but most knowledge of the effects of radiation on embryonic and fetal development comes from experiments with laboratory animals. These studies have shown that the individual is more vulnerable to killing by radiation early in development than at any other time of life. The most sensitive period is that preceding implantation of the fertilized egg in the uterus. At this stage, death may be caused by less than one-third the dose required to kill the adult. After implantation and with increasing maturation, susceptibility to killing declines, and death tends to be postponed until birth or thereafter.

The teratogenic effects of radiation depend on the stage of development of the embryo at the time of irradiation, because different parts of the body are formed according to a definite order and time schedule. Malformation of almost any organ may be induced by irradiation at the appropriate moment during or preceding the development of that organ. Since the major period of organ formation occurs early in embryonic life (during the first trimester of

TABLE 4. Estimated lifetime risks of different forms of cancer attributable to whole-body irradiation

Type of cancer	Additional fatal cases of cancer per 10,000 persons per gray
Leukemia	90–100
Respiratory tract	50–200
Gastrointestinal tract	130–300
Female breast	40–70
Urinary tract	20–40
Genital organs	20–30
All other	100–300
Total	450–1040
Percentage of normal lifetime cancer risk	2.5–6.0

pregnancy in humans), susceptibility to gross malformation decreases in the middle and latter part of gestation. Even late in gestation, however, irradiation may cause abnormalities in the development of slowly maturing organs such as the brain, eye, and gonads. In atomic bomb survivors who were irradiated between the eighth and fifteenth week of embryonic development, the frequency of mental retardation increased with increasing dose. Subtle degrees of abnormality may not be evident at birth.

How radiation causes malformations is not yet fully known, although interference with organ development through the killing of embryonal cells must play an important role in the process. Whatever the mechanism of malformation, the sensitive period for induction of a given abnormality is frequently very short. With increasing radiation dose, however, the sensitive period tends to be prolonged and the incidence and severity of abnormalities increased. Doses as low as 0.25 Gy have been shown to cause developmental disturbances in laboratory animals if applied at the critical time. For this reason, care should be taken to avoid exposing the abdomen of a pregnant woman to radiation early in her pregnancy.

In addition to malformations, other delayed injuries of the type induced by irradiation in adult life may be produced by exposure of the embryo. These, however, tend to manifest themselves long after birth.

Heritable Effects of Radiation

Since radiation-induced mutation may alter characteristics inherited through eggs or sperm and may be passed on through successive generations, such effects are of greater potential significance than those affecting only the exposed individual. Although there is virtually no information about the genetic effects of radiation in human beings, certain facts established by investigations with other organisms appear applicable to all species, including humans.

Natural mutation rate. Through various mechanisms, genes are altered at random during the course of time. These alterations vary from gross chromosomal changes, affecting many genes, to changes affecting only one gene at a time (point mutations). The former are, in general, more deleterious in their effects than the latter, but even the latter are preponderantly detrimental and, in certain instances, lethal. Since genes are present in homologous pairs and one gene of each pair is inherited from each parent, the effects of a mutation depend on whether the mutant gene behaves as dominant or recessive or interacts with other genes. In humans, certain dominant mutations are estimated to appear naturally in 4–40 of every million eggs or sperm. *See* LETHAL GENE.

Radiation-induced mutations. For a given dose of radiation the yield of detectable mutations in germ cells varies from gene to gene, and the relative frequencies of the various mutants differ from those obtained with mutagenic chemicals. Because of this variation and because mutations have been observed thus far in very few of the thousands of genes present in any cell, only crude estimates of the effects of radiation on the overall mutation rate can be made.

Although the frequency of point mutation in some types of irradiated cells varies in direct proportion to the dose, regardless of the dose rate, experiments with mice suggest that the latter may not be true of the complex populations of germ cells in mammals. These experiments disclose that highly fractionated or chronic gamma-irradiation causes fewer detectable mutations in germ cells than the same total dose delivered in a brief single exposure. Because of these complexities, and because of the absence thus far of any evidence of radiation-induced mutations in irradiated human populations, including the children of the atomic bomb survivors, any estimates of the mutation-inducing effectiveness of radiation in humans are of necessity highly speculative. Such estimates have, nonetheless, been made in an effort to assess the genetic hazards of radiation, the mutation rate-doubling dose for humans being considered to be no smaller than 1 Gy.

Practical significance of heritable effects. Until more is known about the extent to which the mind and body are normally handicapped by unfavorable genes, the effects of added deleterious mutants may be only guessed at. Consequently, even if the mutagenic effectiveness of radiation were known accurately, the diverse biological and social consequences of a given radiation-induced increase in the mutation rate could not be estimated at present. It is generally accepted, however, that the genetic burden is regulated by the rates at which mutant genes are produced and subsequently eliminated from the population through natural selection. Hence it is thought that any increase in the mutation rate above natural levels augments the genetic burden.

In the absence of more adequate knowledge of human genetics, attempts have been made to assess the genetic burden in terms of the morbidity from specific genetic traits, such as albinism, achondroplasia, and aniridia. Of such traits, which are detectable in about 4% of all live births, only about one-fourth appear to be attributable to simple genetic mechanisms such as a single mutant gene. Accordingly, if the mutation rate were doubled by radiation, although the latter group would presumably be doubled, the frequency of the other groups would not be increased so greatly. The total incidence then, of all such traits, although elevated from approximately 4% to a level above 5%, would probably still remain below 8%, or twice normal.

In addition to the aforementioned traits, however, the occurrence of which is determined by all-or-none mechanisms, there are characteristics that appear to be quantitatively influenced by heredity, such as general vigor, length of life, fertility, and perhaps, intelligence. Animal experimentation has provided evidence that vigor, length of life, and fertility are reduced in the offspring of irradiated parents, but the data are still fragmentary. Nevertheless, because of the importance of these characteristics, similar radiation effects in humans would be very significant.

Effects of Radiation on Plants

Plant cells have been used extensively for studying the genetic effects of radiation and for investigating the effects of radiation on chromosomes and on cell division. Relatively little work has been done in plants, however, on radiation injury at the tissue or organ level. From available evidence, it appears that radiation effects on plant and animal tissues are not qualitatively different if allowances are made for physiological discrepancies between the two types of organisms.

Irradiation has produced a wide variety of disturbances in plants. These vary from subtle changes resulting from mutations in seeds to marked inhibition of growth and killing. The cause of death is not known, but it is noteworthy that the process of photosynthesis seems relatively insensitive to radiation. Many of the effects of radiation on the growth of higher plants can be attributed to destruction of growth hormone or depression of its synthesis.

Induction of mutations by irradiation of seed has been used to good advantage in breeding new varieties of plants; however, it must be remembered that the mutations so induced occur more or less at random. But since the great majority of mutations are deleterious, one improved plant is gained only at the cost of thousands of others. *See* BREEDING (PLANT).

Factors Affecting Radiation Response

The biological effects of radiation, that is, the response of a cell or tissue to radiation, depend on a variety of direct and indirect factors including radiosensitivity, physical characteristics of the radiation, reaction of surrounding tissues to irradiated cells, and the effectiveness of agents capable of modifying radiosensitivity.

Radiosensitivity. The great variation among cells in susceptibility to radiation injury is largely unexplained, although a number of factors are known to influence radiosensitivity. Of these, perhaps the most important is position on the phylogenetic scale, for example, more than 10,000 Gy are required to kill certain bacteria, whereas less than 10 Gy kill most mammals and less than 1 Gy is lethal for many types of mammalian cells (Table 2).

Another determinant of radiosensitivity is the number of sets of chromosomes present in the cell. This is logical, since each set of chromosomes contains genes duplicating those of another. Hence, in any given family of cells, those with multiple sets of chromosomes are more resistant than those with only a single set. On the other hand, radiosensitivity in certain diploid plants increases in relation to chromosomal content. The radiosensitivity of the cell also varies with the stage in the division cycle at which it is irradiated, as well as with its rate of division, as mentioned earlier. Similarly, the radiosensitivity of the organism varies at different stages in its development.

In addition to the aforementioned intrinsic factors affecting radiosensitivity, extrinsic variables such as temperature, moisture, light, and oxygen tension influence susceptibility. Cells irradiated in the frozen state are, in general, relatively radioresistant, presumably because of the reduced formation and diffusion of radicals at low temperature. The influence of temperature is complex, however; it varies with the time the temperature is altered in relation to irradiation and with the type of cells studied. In most cells, vulnerability to radiation varies with oxygen tension, presumably because oxidizing radicals, such as hydrogen peroxide, are important in the production of radiation injury. Oxygen also exerts other effects, however, since in some types of cells the healing of broken chromosomes requires energy obtained through oxidative pathways.

Physical factors of radiation. The biological effects of radiation depend not merely on the amount of radiant energy absorbed but also on the distribution of the radiation in time (dose rate, or radiation intensity) and space (linear ion density, or linear energy transfer). Certain effects, such as chromosomal rearrangements, generally require more than one ionizing particle or radical to act simultaneously within a small volume of the cell. For multihit effects of this type, densely ionizing radiations such as protons have a high relative biological effectiveness in comparison with sparsely ionizing radiations such as gamma rays.

The dose rate may influence the effects of radiation in other ways too, since some types of injury undergo repair with the passage of time. For such injuries, a given total dose is less effective if fractionated into successive widely separated increments than if administered in a single brief exposure. Paradoxically, however, fractionation in some instances increases the effectiveness of radiation if the successive doses are of appropriate size and periodicity. The influence of fractionation depends, therefore, on many variables, including the total dose, total duration of irradiation, number of exposures, dose per exposure, dose rate per exposure, interval between exposures, and recovery capacity of the system irradiated.

Indirect biological effects of radiation. In addition to affecting the cells irradiated, radiation indirectly alters neighboring or distant cells. These indirect effects may conceivably result from toxic substances liberated by the dying cells or from reactive changes occurring as part of the body's adaptation to injury.

Attempts to demonstrate the liberation of toxic materials from irradiated cells have yielded inconsistent results. Hence it is not established whether such materials are produced in significant quantities. It would appear, however, that they are rarely, if ever, of major importance in the reaction to ionizing radiation.

On the other hand, indirect effects resulting from adaptive or reparative processes are well documented, although their relative importance is not always known. These include alterations caused by local inflammation, scarring, and occlusion of blood vessels, in addition to constitutional changes such as immunological depression, hormonal disturbances, and debilitation.

Modification of radiation injury. It has long been the goal of radiobiologists and radiotherapists to be able to increase or decrease radiosensitivity at will. Radiotherapists have sought ways of increasing destruction of cancer cells without damaging normal tissues. Although their objective is yet to be fully attained, methods for modifying radiosensitivity are now known and are being tested at the experimental and clinical levels.

One of these methods consists in administering drugs that reduce the effectiveness of radiation, principally by blocking or inactivating radiation-induced radicals. The most potent of such drugs yet discovered are cysteine and its derivatives, some of which lower the effectiveness of x-radiation in microorganisms and laboratory animals by a factor of 2 or more. Although most of the chemicals of this type thus far tested have proved too toxic for human use, a few of the newer agents appear promising.

The opposite approach, enhancement of the effectiveness of radiation, is likewise still in the experimental stage. Perhaps the most notable of the radiosensitizing agents studied is oxygen. Irradiation of cancer under increased oxygen tension has been prompted by the fact that many tumors are relatively poorly oxygenated in contrast to normal tissues and to that extent relatively radioresistant.

Among other modifying agents under investigation are chemicals such as nitrogen mustard, urethane, and fluorinated pyrimidines, some of which themselves possess radiomimetic, or radiationlike, activity. These compounds, which are effective by themselves in curbing the growth of cancer cells, exert additive or synergistic effects when administered concomitantly with radiation.

In addition to attempts at preventing or modifying radiation injury by certain processes applied before or during exposure, treatments after irradiation are being explored. Some cellular effects hitherto considered irreversible may be inhibited if appropriate measures are taken promptly enough after irradiation. A particularly notable advance is the discovery that otherwise lethal injury of the bone marrow may be repaired by transplantation of nonirradiated blood-forming cells. Unfortunately, however, since immunological reactions usually occur when the donor of the marrow is genetically different from the recipient, use of this method in humans will be greatly restricted until the immunological complications can be overcome. *See* TRANSPLANTATION BIOLOGY.

Effects of Internally Deposited Radioelements

The radiation injuries caused by radioisotopes within the body are basically no different from those caused by radiation penetrating from without. Because of inhomogeneities of distribution, however, and other variables peculiar to internally deposited emitters, the effects of radioisotopes should be considered separately. *See* RADIOACTIVITY; RADIOISOTOPE.

Anatomic distribution of isotope. In distribution of injury, the area damaged depends on the localization of the radioelement in the organism and the energy of the emitted radiation.

The localization of the isotope is determined by its chemical nature, physical properties, and route of entry into the body. Radioactive iodine-131, for example, behaves chemically like stable iodine, being concentrated from the bloodstream in the thyroid gland. Strontium-90, on the other hand, behaves chemically like calcium, being deposited primarily in the mineral salts of bone. In contrast to these two elements, which become localized in specific organs if dissolved in the bloodstream, tritium, like hydrogen,

TABLE 5. Distribution and excretion of some radioactive fission products and some other radioelements

Radionuclide	Organ or tissue of maximum deposition	f^*	Half-life, days Biological	Half-life, days Physical
Carbon-14	Total body	1.0	40	2,090,000
Calcium-45	Bone	1.0	(18,000)[†]	165.2
Cobalt-60	Liver	0.6	6	1,924
		0.2	60	—
		0.2	800	—
Strontium-90	Bone	1.0	(3,900)[†]	10,512
Iodine-131	Thyroid	1.0	120	8.04
Cesium-137	Muscle	0.1	2	11,012
		0.9	110	—
Barium-140	Bone	1.0	(200)[†]	12.79
Polonium-210	Liver[‡]	1.0	50	138.38
	Kidney	1.0	50	—
	Spleen	1.0	50	—
Radium-226	Bone	1.0	(16,000)[†]	584,000
Uranium-233	Bone	0.9	20	159,000
		0.1	5,000	—
	Kidney	0.996	6	—
		0.004	1,500	—
Plutonium-239	Liver	1.0	14,600	8,800,000
	Bone	1.0	36,500	—

[*] f = fraction of organ burden retained with designated biological half-life.
[†] Retention is better represented by a complex exponential plus power function equation.
[‡] When more than one organ has been given, the initial deposition has been determined to be equally distributed.

permeates the entire body. If, conversely, these same elements are incorporated into the body as insoluble particulates, their distribution is altered in that they are then concentrated in phagocytic cells, the location of which depends on the portal of entry of the particulates.

The microscopic distribution of a radioelement—whatever its organ distribution—is usually irregular, with the result that its radioactivity tends to be concentrated in foci. For this reason and since the intensity of emitted radiation diminishes with the square of the distance from the radioactive atom, the distribution of radiation damage tends to be patchy, especially when the emission consists of alpha particles or low-energy beta particles. On the other hand, if the emission consists of penetrating gamma rays, distant organs that contain no radioisotope may be injured.

Elimination of isotope from the body. The degree of injury caused by a given radioelement depends on the amount of radiation it delivers to surrounding tissue. This, in turn, depends on the disintegration rate of the isotope (physical decay) and on its rate of elimination from the tissue by metabolism or excretion (biological decay). Since the physical decay of an isotope occurs at an exponential rate and since for practical purposes its elimination from the body also occurs at a more or less exponential rate after its initial uptake and distribution, the amount remaining in the body at any given time (the effective concentration) will vary, depending on the physical and biological half-lives (**Table 5**). There is, of course, no consistent correlation between the two half-lives; both types of values vary enormously from one isotope to another. In general, however, elements that are concentrated in bone tend to be eliminated very slowly and are therefore particularly hazardous. Included in this group are many heavy elements and fission products.

Practical Hazards of Ionizing Radiation

It is evident that, barring occasional massive exposures from nuclear accidents or atomic warfare, the hazards of radiation result from small doses accumulated over a long period of time. The effects of such doses fall into two categories, genetic effects and delayed somatic effects.

Because of concern about the potential danger of fallout from nuclear weapon tests, systematic studies have been made to ascertain the amount of radiation to which humans are now exposed. This radiation comes from (1) naturally occurring radioelements present in the Earth, atmosphere, and human bodies; (2) cosmic rays; and (3) human-made devices, such as x-ray machines, watch dials, nuclear weapons, and television apparatus. It is now apparent that the natural background level has been materially increased by human-made radiation, particularly medical exposure; however, the contribution from weapon fallout is almost negligible to date, despite its notoriety.

Although, as mentioned above, the biological effects of small doses of radiation cannot yet be estimated accurately, certain recommendations have been made in an effort to safeguard humans against exposure to levels of radiation likely to cause detectable injury. The National Council on Radiation Protection and Measurements in the United States has therefore recommended that for the general population the maximum permissible dose of radiation from human-made sources other than medical exposure not exceed 1 mSv per year.

Concerning the hazard of delayed somatic radiation injury such as cancer, there is no conclusive evidence as yet that doses only slightly above the natural background are damaging. It is noteworthy, however, that on the basis of extrapolation from the experimental and clinical data now available, natural background radiation cannot conceivably account for more than a small fraction of the natural incidence of cancer. *See* RADIATION BIOLOGY; RADIOACTIVE FALLOUT.   Arthur C. Upton

Bibliography. C. B. Meinhold (ed.), *Limitation of Exposure to Ionizing Radiation*, 1993; National Academy of Sciences–National Research Council, Committee on Biological Effects of Ionizing Radiation, *Health Effects of Exposure to Low Levels of Ionizing Radiation* (BEIR V), 1990; United Nations Scientific Committee on the Effects of Atomic Radiation, *Sources, Effects and Risks of Ionizing Radiation: Report to the General Assembly with Annexes*, 1988.

Radiation pressure

The force on an object exposed to electromagnetic radiation. It has been known since the days of J. C. Maxwell in the nineteenth century that electromagnetic radiation (which includes visible light) carries both energy and momentum. If radiation impinges on a material body and becomes absorbed, the energy gives rise to heat and is readily detectable. When radiation interacts with an object and is absorbed or scattered, there is also a change in the momentum of the light. By conservation of momentum, this gives rise to a force on the object. This is called radiation pressure. The magnitude of this momentum for visible light is quite small and is difficult to detect. Only near or inside stars, where the intensity is enormous, do light forces have large effects. *See* ELECTROMAGNETIC RADIATION; LIGHT; MAXWELL'S EQUATION; POYNTING'S VECTOR.

This situation is different with lasers. Lasers generate coherent light beams and can be focused to spot sizes of about one wavelength, giving rise to very high intensities and intensity gradients, using modest total powers. Using Planck's law, the energy of a light photon is $E_{photon} = h\nu$, where h is Planck's constant and ν is the frequency. From Einstein's law, the photon's energy in terms of its effective mass is $E_{photon} = m_{effective}\, c^2$, where c is the velocity of light. The photon momentum then is $m_{effective}\, c = h\nu/c$. If, for example, a laser beam strikes a 100% reflecting mirror, it generates a radiation pressure force $F_{rad} = (P/h\nu)\,(2h\nu/c) = 2P/c$, where P is the power carried by the beam. For $P = 1$ watt,

$F_{rad} \cong 10^{-8}$ N $= 10^{-3}$ dynes. *See* LASER; PHOTON; PLANCK'S CONSTANT; QUANTUM MECHANICS; RELATIVITY.

For a macroscopic mirror of mass of about 1 gram, the acceleration due to the radiation is $A = F/m \cong 10^{-3}$ cm/s$^2 \cong 10^{-6}$ g, where g is the acceleration of gravity, about 10^3 cm/s^2. This is quite small. For a 1 micrometer-sized particle, however, with a density of about 1 g/cm^3, the acceleration is $F/m \cong 10^{-3}/10^{-12} = 10^9$ cm/s$^2 \cong 10^6$ g, which is very large and should be readily observable.

Particle guidance and trapping. Based on these considerations, Arthur Ashkin in 1970 showed experimentally that a fraction of a watt of laser power striking a small transparent polystyrene particle several micrometers in diameter was sufficient to propel it many diameters per second through water in the direction of the light. This force in the direction of the light, called the scattering force, agreed with the above calculation. However, it was discovered that, when the particle was located off the beam axis in the region of a strong gradient of light intensity, there was an additional force component pushing the particle into the high-intensity region at the center of the beam. This component of force was termed the gradient component. It served to guide the particle and keep it on axis as it moved along the beam.

Figure 1 shows that both these force components do indeed originate from radiation pressure. It shows a sphere with high index of refraction, many wavelengths in diameter, placed off axis in a mildly focused beam with a Gaussian intensity profile. A typical pair of rays, *a* and *b*, are shown, striking the sphere symmetrically about its center. Neglecting relatively minor surface reflections, most of the rays refract through the particle, giving rise to forces F_a and F_b in the direction of the momentum change. Since the intensity of ray *a* is higher than ray *b*,
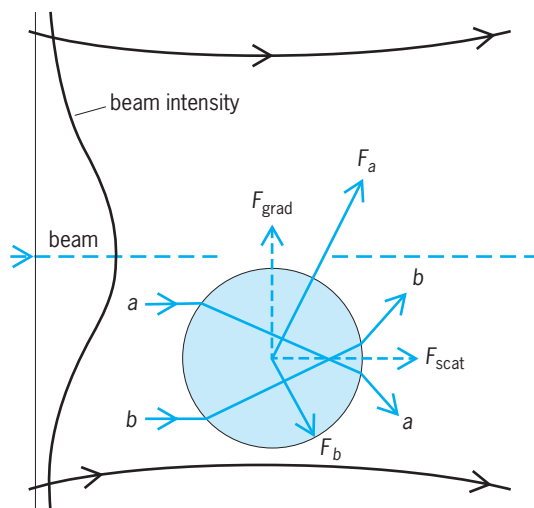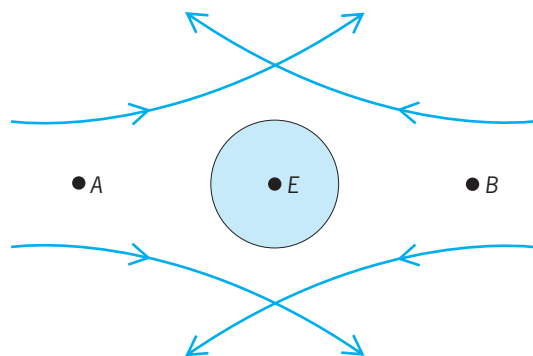

Fig. 2. **Geometry of a two-beam trap. Curved lines with arrows indicate rays of the two beams.**

the force F_a is greater than F_b. Adding all such symmetrical pairs of rays striking the sphere, one sees that the net force can be resolved into two components: the scattering force component F_{scat}, in the direction of the light; and the gradient component F_{grad} arising from the gradient in light intensity and pointing transversely toward the high-intensity region of the beam. For a particle on axis or in a plane wave, $F_a = F_b$, and there is no net gradient force component.

The understanding of the magnitude and properties of these two basic force components made it possible to devise the first stable three-dimensional (3D) optical trap for single neutral particles. The trap consisted of two opposing moderately diverging Gaussian beams focused at points *A* and *B* (**Fig. 2**). The predominant effect in any axial displacement of a particle from the equilibrium point *E* (that is, in a direction parallel to the beam axes) is a net opposing scattering force. Any radial displacement (that is, in a direction perpendicular to the beam axes) is opposed by the gradient force of both beams. The trap was filled by capture of randomly diffusing small particles which wandered into the trap.

At this time it was proposed that similar stable trapping of atoms was possible using the same two force components. Much progress was made in the 1970s in understanding light forces on macroscopic transparent dielectric particles and atoms. In 1972 Ashkin and colleagues stably levitated transparent particles against gravity using a single, mildly divergent, vertically directed laser beam in air and in vacuum. In vacuum an optical feedback system was developed to lock levitated particles to an external reference and also to provide optical damping to avoid fluctuation in position and possible escape of particles in the absence of viscous air damping. It acted as a sensitive force-measuring device as well, since the application of any type of external force resulted in a change in the laser power needed to maintain a fixed position. High-resolution measurement of the charge of the electron was made using an applied electric field in a modern version of the Millikan oil-drop experiment. With a tunable dye laser with feedback, extremely high Q resonant responses of a transparent

Fig. 1. **Origin of scattering force component F_{scat} and gradient force component F_{grad} for a sphere with high index of refraction displaced from the axis of a mildly focused Gaussian beam. Curved lines with arrows indicate rays of the beam.**

oil droplet were observed for the first time. These resonances corresponded to optical whispering gallery modes of the spherical droplet. They gave a means of determining particle size and the index of refraction to unprecedented accuracy. Highly resonant quartz spheres can be used as filters in optical communications, as resonators in quantum electrodynamics experiments, and as high-Q resonators in nonlinear optics experiments.

Atom traps. With atoms, the scattering force can be understood as the absorption of resonant photons and their subsequent reemission. The gradient force on atoms can be thought of as force on an optically induced dipole in the gradient of the optical electric field. In the early 1970s, experiments were done on the deflection of an atomic beam by a resonant light beam for purposes of isotope separation. This geometry avoids the effects of Doppler shifting.

In 1973 Theo W. Hänsch and Arthur L. Shawlow proposed using the Doppler shift to provide optical damping or cooling of an atomic gas. If a pair of opposing laser beams in one dimension is tuned below an atomic resonance, then an atom moving in either direction feels a net damping because it is Doppler-shifted to a higher frequency, closer to resonance for the opposing beam, and to a lower frequency, away from resonance, for the trailing beam. With three pairs of beams in three dimensions, all the degrees of freedom can be damped. This creates what is known as an optical molasses. Residual fluctuations of the scattering force limit the cooling to a temperature known as the Doppler limit. For sodium atoms this is about 240 microkelvin. Attempts to design traps to capture atoms on the intensity maxima of the standing waves using the gradient force have problems because of excessive saturation, or the ability of the atom to absorb and emit only a finite number of photons. In 1978 Ashkin solved this problem by designing gradient force traps detuned far below resonance to reduce absorption. Two-beam traps and a single-beam, strongly focused gradient trap were proposed for holding the atoms, cooled by optical molasses. The single-beam gradient trap, later referred to as a tweezers, is the simplest of all traps, consisting of a single strongly focused Gaussian beam. *See* DOPPLER EFFECT; LASER COOLING; PARTICLE TRAP.

The proposed technique to increase the gradient force was checked in 1978 in experiments showing focusing and guiding of atoms in detuned light beams. This marked the beginning of what is known as atom optics, and also suggested that stable atom traps were feasible. By 1984 William D. Phillips and Wolfgang Ertmer succeeded in overcoming the strong Doppler shifts involved in slowing thermal atomic beams to a temperature of about 1 K, using the scattering force of an opposing light beam. Such atoms provided an excellent source for further three-dimensional cooling. In 1985 Steven Chu entered the field and, with J. E. Bjorkholm, Ashkin, and L. W. Hollberg, demonstrated the first molasses cooling of sodium atoms to around 240 microkelvin. These atoms were used in 1986 to demonstrate the first stable trapping of neutral atoms using the off-resonance single-beam gradient trap. Later in 1986 another type of large-volume, molasses-cooled trap called a magneto-optic-trap (MOT) was demonstrated. This could trap many more atoms than the first tweezer trap.

Overcoming the Doppler limit. In 1988 Phillips and his colleagues showed that by changing the molasses detuning, it was possible to reach a new cooling regime, where the atom's temperature was around 20 microkelvin, ten times lower than the Doppler limit. The theory of this process was explained principally by Claude Cohen-Tannoudji and his colleagues and was called Sisyphus cooling. Applications of these ultracold atoms soon appeared, such as improved atomic clocks using atomic fountains, atom interferometers to measure the gravitational force with much improved accuracy, and the study of cold-atom collisions in the regime of large de Broglie wavelengths. *See* ATOMIC CLOCK.

Biological applications. In 1986 Ashkin and his colleagues showed that tweezer traps could also trap submicrometer and micrometer-sized dielectric particles. In further experiments in 1987, Ashkin and J. M. Dziedzic serendipitously discovered that infrared tweezers could trap living bacteria, viruses, and other biological organisms without causing optical damage. It was even possible to optically manipulate organelles inside larger cells and observe the mechanical properties of cytoplasm. This initiated the application of laser tweezers to biology.

Study of motor molecules. A major application of tweezers is the study of motor molecules that drive organelle movement, cell locomotion, and muscle action. Steven M. Block and his colleagues improved earlier in-vitro motility assays by optically manipulating single molecules attached to dielectric spheres in what is known as the handles technique. For kinesin they resolved the detailed motion along a microtubule into a sequence of 8-nanometer steps. They measured the complete force-velocity relationship with a maximum force of about 5 to 6 piconewtons. For myosin, Jeffrey T. Finer and his colleagues used a new feedback-enhanced dual optical trap to measure stepwise motion on actin filaments of about 11 nm and forces of 3 to 4 pN. These studies often used mutant forms of actin and kinesin molecules. They measured the ATP hydrolysis cycle and single-enzyme kinetics, and the efficiency of kinesin motors. Another advance was the development of a novel feedback-controlled molecular force clamp that maintains constant force on moving single molecules. Such force clamps were used to study another class of motors, the nucleic acid motor enzymes such as RNA polymerase, which synthesizes a RNA replicate of DNA, and DNA polymerase, which catalyzes DNA replication. These motors are slow and powerful. Resolutions down to about a single DNA base pair were achieved, and novel behavior only previously

postulated was observed, such as the backtracking by RNA polymerase during long pauses in transcription, during which the proofreading and error-correcting mechanisms can cleave and discard flawed sections. Block and his colleagues, using identically prepared strands, showed that RNA polymerase tracks the DNA groove during the search for the promoter sequence that initiates transcription. Carlos Bustamante introduced the study of DNA polymerase. Data on replication in base pairs per second taken at different tensions using constant force feedback gave the kinetic tension or stall force at which replication ceases. *See* CELL MOTILITY; CELL ORGANIZATION; MITOSIS; MUSCLE PROTEINS; NUCLE-OPROTEIN.

Other DNA-associated molecules were studied with tweezers, including chromatin, protamines, arginines, MukBEF, and RecBCD. Many structural measurements were made on DNA and other macromolecules, including stretching and applying controlled torques. Using traps, David G. Grier and others have made the surprising observation of confinement-induced attractive forces between like colloidal particles in charge-stabilized suspensions. Despite considerable theoretical attention, this effect remains unexplained.

Optically driven motors. The rotation of particles by light forces has been much studied using laser beams. The rotation of shaped (vanelike) particles caused by the linear momentum of light was observed. Also, rotation induced by the intrinsic angular momentum of the photon and the orbital angular momentum of higher-order rotating light beam modes was investigated. Such optically driven motors are expected to be important for study of physical and biological materials.

Fluidic sorting. Another topic studied was the fluidic sorting of particles and cells using dynamic holographic tweezers. With nematic spatial light modulators it is possible to form arrays with hundreds of adjustable tweezer traps. One can spatially sort mixtures of particles by flowing them through fixed arrays of weak potentials or by capturing them and transporting them through space in what is called optical peristalsis. Or one can drive particles through lithographically formed microchannels using optically driven gear-type fluid pumps and gate valves. Sorters of this type have advantages over large-volume conventional-flow cytometric sorters.

Bose-Einstein condensation. A major advance involves evaporative cooling of atoms from magnetic quadrupole traps to submicrokelvin temperatures (the coldest temperatures in the universe) at which Bose-Einstein condensation (BEC) can occur. In Bose-Einstein condensation, the de Broglie wavelengths of atoms overlap, and they occupy the lowest energy state of the trap, where they behave as a single quantum particle. Experiments have shown the coherence of atoms in such condensates, and crude atom lasers were demonstrated. Applications of Bose-Einstein condensates are possible: in Josephson devices, coherent atom lithography, quantum computers and cryptography, and even more accurate atomic clocks. *See* ATOMLASER; BOSE-EINSTEIN CONDENSATION.

Subsequently, fermionic atoms were cooled well below the Fermi temperature. Using all-optical traps, it is possible to magnetically tune the interaction strength across Feshbach magnetic resonances for all hyperfine levels without affecting the trapping ability. In this way molecular Bose-Einstein condensates were formed for the first time, when tuned above a Feshbach resonance. Strongly interacting fermionic pairing was also observed for tunings below resonance where the gas acts as a superfluid. Weakly interacting Cooper pairs, as in superconducting solids or superfluids, have not as yet been achieved. Using all-optical traps, it was possible to demonstrate five signatures of strongly interacting pairs. It is now clear that optical traps are superior to magnetic traps for making and exploiting fermionic Bose-Einstein condensates. Observations on fermionic pairing may shed light on the phenomenon of high-temperature superconductivity. Radiation pressure forces from lasers can be expected to play an important role in physics, chemistry, and the biological sciences in future work on small particles. *See* RESONANCE (QUANTUM MECHANICS); SUPERCONDUCTIVITY.   Arthur Ashkin

Bibliography. A. Ashkin, Applications of laser radiation pressure, *Science*, 210:1081–1088, 1980; A. Ashkin, History of optical trapping and manipulation of small neutral particle, atoms, and molecules, *IEEE J. Selected Topics Quant. Electr.*, 6(6):841–855, 2000; K. C. Neuman and S. M. Block, Optical tapping, *Rev. Sci. Instrum.*, 75(9):2787–2809, 2004.

Radiation shielding

Physical barriers designed to provide protection from the effects of ionizing radiation; also, the technology of providing such protection. Major sources of radiation are nuclear reactors and associated facilities, medical and industrial x-ray and radioisotope facilities, charged-particle accelerators, and cosmic rays. Types of radiation are directly ionizing (charged particles) and indirectly ionizing (neutrons, gamma rays, and x-rays). In most instances, protection of human life is the goal of radiation shielding. In other instances, protection may be required for structural materials which would otherwise be exposed to high-intensity radiation, or for radiation-sensitive materials such as photographic film and certain electronic components.

For the effects of radiation on living organisms *see* HEALTH PHYSICS; NUCLEAR RADIATION (BIOLOGY); RADIATION BIOLOGY; RADIATION INJURY (BIOLOGY) For the effects on inanimate materials *see* RADIATION DAMAGE TO MATERIALS; NUCLEAR EXPLOSION.

Radiation sources. Nuclear-electric generating stations present a variety of shielding requirements. In the nuclear fission process, ionizing radiation is
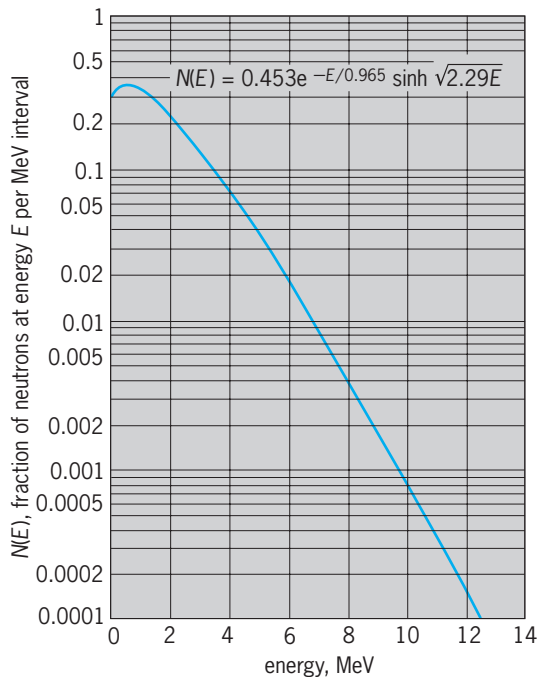
Fig. 1. Neutron spectrum as produced from the fission of ^{235}U by thermal neutrons. (*After L. Canberg et al., Phys. Rev., 103:662–670, 1956*)

$$N(E) = 0.453e^{-E/0.965} \sinh \sqrt{2.29E}$$

released not only at the instant of fission but also through radioactive decay of fission products and products of neutron absorption. Of the energy released in the fission process, about 2.5% is carried by prompt fission neutrons and 3% by delayed gamma rays. About 2.5 neutrons are released per fission with average energy of 2 MeV. About 90% of the prompt neutrons have energies less than 4 MeV, but because of their greater penetrating ability the remaining 10% are of greater concern in radiation shielding. Neutrons are distributed in energy approximately as shown in **Fig. 1**, which gives the distribution for fission of uranium-235. The distribution for other fissionable isotopes is similar. *See* NUCLEAR FISSION; NUCLEAR REACTOR; REACTOR PHYSICS.

Fission of one atom releases about 7 MeV of energy in the form of prompt gamma rays. This energy is distributed over some 10 gamma-ray photons with the energy spectrum as shown in **Fig. 2**. Evaluation of the intensity of delayed gamma-ray sources within a nuclear reactor requires knowledge of the operating history of the reactor. Absorption of neutrons in structural or shielding materials results in the emission of capture gamma rays. Likewise, gamma rays result from inelastic scattering of neutrons. Because these gamma rays (especially capture gamma rays) are generally of higher energy than prompt fission gamma rays and because they may be released deep within a radiation shield, they require careful consideration in shielding design for nuclear reactors. *See* GAMMA RAYS; NEUTRON; NEUTRON SPECTROMETRY; NUCLEAR REACTION.

Controlled thermonuclear reactors, deriving energy from the nuclear fusion of deuterium and

tritium, present radiation shielding requirements similar in kind to those of fission reactors. Highly penetrating (14-MeV) neutrons are released in the fusion process along with charged particles and photons. Capture gamma rays again require careful consideration. *See* NUCLEAR FUSION.

X-ray generators vary widely in characteristics. Typical units release x-rays with maximum energies to 250 keV, but high-voltage units are in use with x-ray energies as high as tens of megaelectronvolts. The dominant nature of x-ray energy spectra is that of bremsstrahlung. X-ray generators are but one of many types of charged-particle accelerators and, in most cases, the governing shielding requirement is protection from x-rays. For very high-energy accelerators, protection from neutrons and mesons produced in beam targets may govern the shield design. *See* PARTICLE ACCELERATOR; RADIOLOGY; X-RAYS.

Although many different radionuclides find use in medical diagnosis and therapy as well as in research laboratories and industry, radiation shielding requirements are of special importance for gamma-ray and neutron sources. Alpha and beta particles from radionuclide sources are not highly penetrating, and shielding requirements are minimal. Space vehicles are subjected to bombardment by radiation, chiefly very high-energy charged particles. In design of the space vehicle and in planning of missions, due consideration must be given to radiation shielding for protection of crew and equipment. *See* COSMIC RAYS; RADIOISOTOPE; RADIOISOTOPE (BIOLOGY).

Attenuation processes. Charged particles lose energy and are thus attenuated and stopped primarily as a result of coulombic interactions with electrons
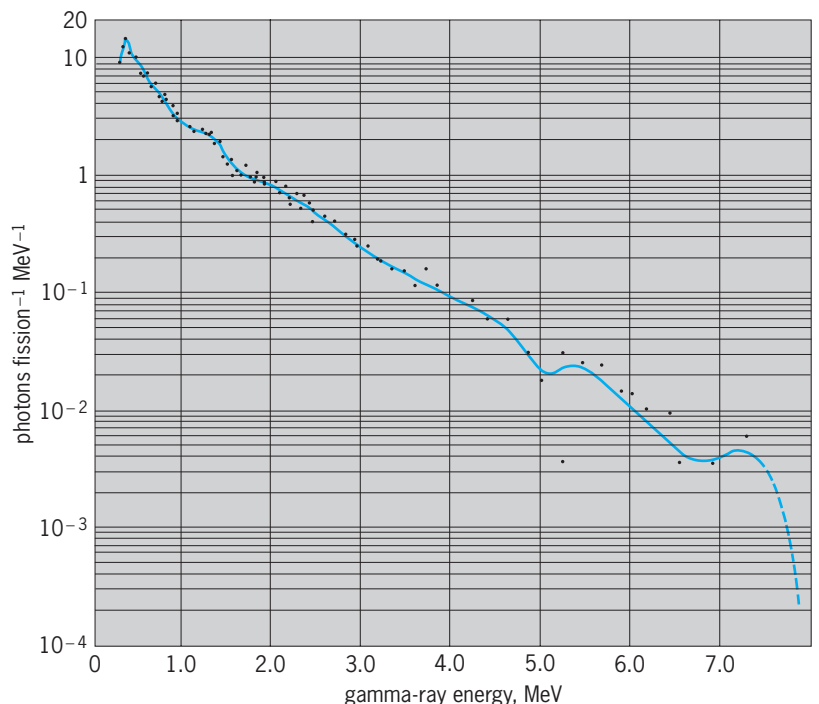

Fig. 2. Energy spectrum of gamma rays observed within 10^{-7} s after fission.

of the stopping medium. For heavy charged particles (protons, alpha particles, and such), paths are nearly straight and ranges well defined. Electrons may suffer appreciable angular deflections on collision and may lose substantial energy radiatively. Very high-energy charged particles may lose energy through nuclear interactions, resulting in fragmentation of the target nuclei and production of a wide variety of secondary radiations.

Gamma-ray and x-ray photons lose energy principally by three types of interactions: photoemission, Compton scattering, and pair production. In photoemission, or the photoelectric effect, the photon transfers all its energy to an atom, and an electron is emitted with kinetic energy equal to the original energy of the photon less the binding energy of the electron in the atom. In Compton scattering, the photon is deflected from its original course by, and transfers a portion of its energy to, an electron. In pair production, the gamma ray is converted to a positron-electron pair. At least 1.02 MeV of gamma-ray energy is required for the rest mass of the pair, and any excess appears as kinetic energy. Ultimately, the positron and an electron recombine and, in annihilation, release two 0.505-MeV gamma rays. Photoemission is especially important for low-energy photons and for stopping media of high atomic number. Compton scattering usually dominates at intermediate photon energies, and pair production at high photon energies. *See* COMPTON EFFECT; ELECTRON-POSITRON PAIR PRODUCTION; PHOTOEMISSION.

Neutrons lose energy in shields by elastic or inelastic scattering. Elastic scattering is more effective with shield materials of low atomic mass, notably hydrogenous materials, but both processes are important, and an efficient neutron shield is made of materials of both high and low atomic mass. The fate of the neutron, after slowing down as a result of scattering interactions, is absorption frequently accompanied by emission of capture gamma rays. Suppression of capture gamma rays may be effected by incorporating elements such as boron or lithium in the shield material. The isotopes ^{10}B and ^{6}Li have large cross sections for neutron capture without gamma-ray emission.

Shielding concepts. The cross section of an atom or electron for interaction with radiation is the effective "target" area presented for the interaction. It is usually given the symbol σ and the units cm^2 or barns (1 barn = 10^{-24} cm^2). The cross section depends on the type of interaction and is a function of the energy of the radiation. When the cross section is multiplied by the number of atoms or electrons per unit volume, the product, identified as the linear attenuation coefficient μ or macroscopic cross section Σ, has the units of reciprocal length and may be interpreted as the probability per unit distance of travel that the radiation experiences in an interaction of a given type. The total attenuation coefficient for radiation of a given energy is the sum of attenuation coefficients for all types of interactions in the shielding medium. The quotient of the linear attenuation coefficient and
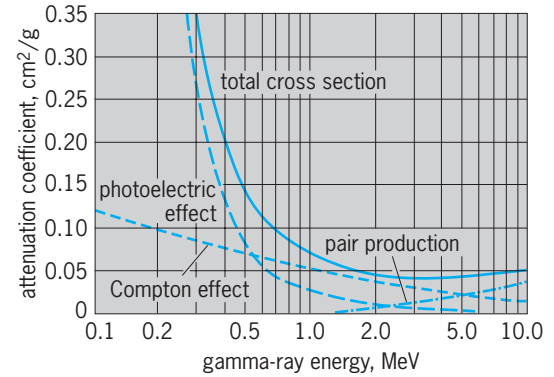

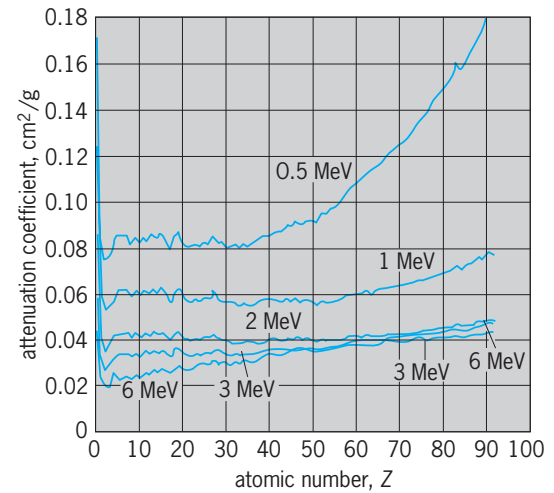Fig. 3. Gamma-ray attenuation coefficients for lead.

Fig. 4. Gamma-ray attenuation coefficients versus atomic number of absorbing material for various energies.

the density, μ/ρ, is called the mass attenuation coefficient. If μ/ρ for gamma rays is weight-averaged by the fraction of the gamma-ray energy locally dissipated subsequent to an interaction, the average, μ_a/ρ, is called the mass energy absorption coefficient. The total and component parts of the mass attenuation coefficients for lead are illustrated in **Fig. 3**. In **Fig. 4** are shown total mass attenuation coefficients for all elements at several photon energies.

The flux density, or fluence rate, characterizes the intensity of radiation. It may be thought of as the path length traveled by radiation per unit volume per unit time. It is usually given the symbol ϕ and has units cm^{-2} s^{-1}.

To illustrate these concepts, consider the flux density at a distance r from a point source isotropically emitting S monoenergetic gamma rays of energy E per second in a uniform medium with total attenuation coefficient μ. If μ were zero, the flux density would be determined just by the inverse square of the distance from the source. If μ were not zero, attenuation of the gamma rays would also be exponential with distance. Thus the equation shown below

$$\phi = \frac{S}{4\pi r^2} e^{-\mu r}$$

would hold. This is the flux density of gamma rays which have traveled distance r without having experienced any interactions in the stopping medium. The energy locally dissipated per unit mass per unit time due to these gamma rays as they experience their first interactions is the product $E(\mu_a/\rho)\phi$. However, some gamma rays reach distance r already having experienced scattering interactions. To account for these secondary gamma rays, a buildup factor B is employed. B is a function of the energy of the source gamma rays and the product μr and depends on the attenuating medium. The total energy locally dissipated per unit mass per unit time is thus the product $BE(\mu_a/\rho)\phi$. Similar concepts apply to neutron shielding; however, the treatment of scattered neutrons is considerably more complicated, and the buildup-factor concept is not well established. In the final phases of shielding design, digital computers are usually employed to carry out the required calculations. It is also common practice to base shielding design, at least in part, on measurements made on a prototype.

Shielding materials. The most common criteria for selecting shielding materials are radiation attenuation, ease of heat removal, resistance to radiation damage, economy, and structural strength.

For neutron attenuation, the lightest shields are usually hydrogenous, and the thinnest shields contain a high proportion of iron or other dense material. For gamma-ray attenuation, the high-atomic-number elements are generally the best. For heat removal, particularly from the inner layers of a shield, there may be a requirement for external cooling with the attendant requirement for shielding the coolant to provide protection from induced radioactivity.
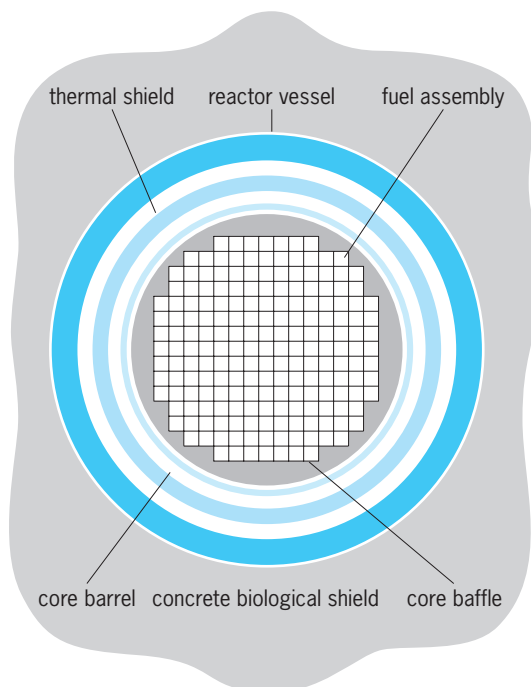
Metals are resistant to radiation damage, although there is some change in their mechanical properties. Concretes, frequently used because of their relatively low cost, hold up well; however, if heated they lose water of crystallization, becoming somewhat weaker and less effective in neutron attenuation.

If shielding cost is important, cost of materials must be balanced against the effect of shield size on other parts of the facility, for example, building size and support structure. If conditions warrant, concrete can be loaded with locally available material such as natural minerals (magnetite or barytes), scrap steel, water, or even earth.

Typical shields. Radiation shields vary with application. The overall thickness of material is chosen to reduce radiation intensities outside the shield to levels well within prescribed limits for occupational exposure or for exposure of the general public. The reactor shield is usually considered to consist of two regions, the biological shield and the thermal shield. The thermal shield, located next to the reactor core, is designed to absorb most of the energy of the escaping radiation and thus to protect the steel reactor vessel from radiation damage. It is often made of steel and is cooled by the primary coolant. The biological shield is added outside to reduce the external dose rate to a tolerable level (**Fig. 5**).   Richard E. Faw

Bibliography. Center for Occupational Research and Development Staff, *Course 19: Radiation Shielding*, 1984; A. B. Chilton, K. Shultis, and R. E. Faw, *Principles of Radiation Shielding*, 1984; R. G. Jaeger et al. (eds.), *Engineering Compendium on Radiation Shielding*, vol. 1, 1968, vol. 2, 1975, vol. 3, 1970; International Atomic Energy Agency, *Reactor Shielding*, 1981; M. F. Kaplan, *Concrete Radiation Shielding*, 1989; A. E. Profio, *Radiation Shielding and Dosimetry*, 1979; J. Wood, *Computational Methods in Reactor Shielding*, 1982.

Radiation therapy

The use of ionizing radiation to treat disease; the method is also known as radiotherapy and therapeutic radiology. Radiotherapy was widely used in the past to treat diseases of the skin, lymph nodes, and other organs. However, because radiation can cause cancer and because alternative treatments for these diseases have been discovered, radiation therapy is now mainly limited to treating malignant tumors: the medical specialty is called radiation oncology. *See* ONCOLOGY; RADIATION INJURY (BIOLOGY).

Radiobiologic basis. The exact mechanisms by which radiation kills cells remain uncertain. Most likely, electrons dislodged from water or biological molecules disrupt the bonds between atoms of the nuclear deoxyribonucleic acid (DNA), resulting in double-strand breaks. Although usually not immediately fatal, such damage may cause the death of cells when they attempt to divide. Complex enzymatic mechanisms can repair some of the damage if given sufficient time.

thermal shield   reactor vessel   fuel assembly

core barrel   concrete biological shield   core baffle

Fig. 5. Typical power reactor shield configuration.

The fraction of cells surviving after irradiation depends on many factors. Radiation affects both normal and cancerous cells in a similar manner qualitatively, but different cell lines vary greatly in their quantitative sensitivity. Rapidly dividing tissues—such as the skin, bone marrow, and gastrointestinal mucosa—usually display the greatest sensitivity experimentally and the most immediate side effects clinically. The cellular environment is important; for example, hypoxic cells that receive less-than-normal amounts of oxygen from the circulatory system are often found in large tumors, and are less sensitive to x-rays than well-oxygenated cells. Particles that produce a dense path of ionizations (such as pions) are more effective at killing cells than is sparsely ionizing radiation (such as photons). Treatment parameters such as the energy of the radiation, the dose given in a single treatment (fraction size), the speed with which it is delivered (dose rate), the interval between treatments (fractionation), the number of days of the treatment course, and the total dose critically influence outcome.

Clinical aspects. The goal of radiation therapy is either to cure the disease permanently (radical treatment) or to reduce or eliminate symptoms (palliative treatment) by destroying tumors without causing unacceptable injuries to normal tissues. For many cancers—such as cancers of the reproductive organs, lymphomas, and small head and neck tumors—a cure usually is possible, and the chance of functionally significant complications is small.

Some tumors, however, contain too many cells to be entirely destroyed by tolerable doses. Also, some neoplasms, such as sarcomas and glioblastomas, are relatively resistant to irradiation and are difficult to eradicate even when only small numbers of cells are present. Such situations are best handled by using surgery to remove all visible tumor. Microscopic deposits of tumor cells often remain in and near the surgical field in tissues that cannot be removed without risk of further injury. Radiotherapy reduces the chance of a local recurrence following surgery. This approach is used successfully in managing the early stages of cancer of the breast (following limited, breast-preserving surgery, or lumpectomy) and in treating tumors in many other body sites, including the brain, head and neck, lung, gastrointestinal tract, and gynecological organs. Radiotherapy is sometimes used before surgery (rather than after surgery) to increase the ease of surgical removal or to lessen the chance of tumor cells being spread by surgical manipulations.

Also, radiation often is used to treat tumors that cannot be removed because of extensive invasion into surrounding structures, that are located in sites where surgery would cause unacceptable complications, or that occur in patients unable to tolerate the stresses of surgery. Except for small or sensitive tumors, success rates with conventional radiotherapy usually have been moderate or poor. Chemotherapy given in combination with radiation (sequentially or simultaneously) may reduce tumor bulk and sensitize cells to irradiation, as well as destroy tumor cells outside the radiation fields. Total body irradiation is used in some diseases (mainly leukemias and lymphomas) to destroy both tumor cells and the bone marrow in preparation for bone marrow transplantation. *See* CHEMOTHERAPY.

Technical aspects. The ionizing radiations most commonly used clinically are photons in the x-ray and gamma-ray energy range (10,000 eV to 45 MeV) and electrons (3 to 20 MeV) produced either by linear or circular particle accelerators or by radioactive isotopes (predominantly radium-226, cobalt-60, iridium-192, cesium-137, and iodine-125). Beams of particles or photons may be directed at the patient from a distance (teletherapy or external-beam therapy). In general, higher energy results in deeper tissue penetration and relatively lower doses to superficial structures such as the skin. Radioisotopes may be placed on the body surface (mold techniques), inside body cavities (intracavitary application), or temporarily or permanently implanted into tissues (interstitial therapy). Some radioisotopes, such as phosphorus-32, gold-198, and iodine-131, have also been administered orally, by intravenous injection, or by instillation into body cavities. Dose is defined as the energy absorbed per unit tissue mass; its unit is the gray (1 Gy = 1 joule/kg = 100 rad). *See* RADIOISOTOPE.

The radiotherapist locates tumor masses by using physical examination, imaging studies, and surgical findings. Knowledge of the patterns of spread and natural history of the disease is used to determine which apparently uninvolved areas might contain tumor cells. These areas are usually irradiated for at least part of the overall course of treatment. Careful planning of the directions of entry and sizes of the radiation beams (fields), the use of multiple converging fixed or rotating fields, and the use of intracavitary and interstitial techniques (brachytherapy) ordinarily permit delivery of much higher doses to the tumor than to surrounding organs. Manipulating the parameters of treatment, particularly dose rate, fraction size, and fractionation, can increase the chance of tumor eradication while lowering the likelihood of complications. The total dose prescribed depends on the exact type of tumor and clinical situation, but it is usually 20–80 Gy delivered over 1–8 weeks.

Technique refinements. Many improvements to existing radiotherapy techniques and practices have been explored. Computed tomography and magnetic resonance imaging, which visualize tumors and organs in three dimensions with great precision, are increasingly used in combination with sophisticated treatment planning and dosimetry programs. Computer control of accelerator motion and dose output, intraoperative radiotherapy, and elaborate localization and immobilization (stereotactic) techniques help to concentrate radiation on tumors and spare normal tissues. Certain drugs (radiosensitizers) and increasing tumor temperatures (hyperthermia) may increase tumor-cell killing while adding few or no complications. Radio-protective drugs for normal tissues are also under investigation. Neutrons, protons, pions, and charged atomic nuclei such as helium,

which interact with matter differently than photons, may be useful in certain clinical situations. *See* CANCER (MEDICINE); RADIATION BIOLOGY.

Abram Recht

Bibliography. G. C. Bentel et al., *Treatment, Planning, and Dose Calculation in Radiation Oncology,* 4th ed., 1989; B. Blackburn (ed.), *Blackburn's Introduction to Clinical Radiation Therapy Physics,* 1989; V. T. DeVita, S. Hellman, and S. A. Rosenberg (eds.), *Cancer: Principles and Practices of Oncology,* 6th ed., 2001; J. M. Vaeth and J. L. Meyer (eds.), *Radiation Tolerance of Normal Tissues,* Frontiers of Radiation and Oncology Series, vol. 23, 1990.

Radiative transfer

The study of the propagation of energy by radiative processes; it is also called radiation transport. Radiation is one of the three mechanisms by which energy moves from one place to another, the other two being conduction and convection. *See* ELECTROMAGNETIC RADIATION; HEAT TRANSFER.

The kinds of problems requiring an understanding of radiative transfer can be characterized by looking at meteorology, astronomy, and nuclear reactor design. In meteorology, the energy budget of the atmosphere is determined in large part by energy gained and lost by radiation. In astronomy, almost all that is known about the abundance of elements in space and the structure of stars comes from modeling radiative transfer processes. Since neutrons moving in a reactor obey the same laws as radiation being scattered by atmospheric particles, radiative transfer plays an important part in nuclear reactor design.

Each of these three fields—meteorology, astronomy, and nuclear engineering—concentrates on a different aspect of radiative transfer. In meteorology, situations are studied in which scattering dominates the interaction between radiation and matter; in astronomy, there is more interest in the ways in which radiation and the distribution of electrons in atoms affect each other; and in nuclear engineering, problems relate to complicated, three-dimensional geometry.

Physical processes. Radiative transfer is a complicated process because matter interacts with the radiation. This interaction occurs when the photons that make up radiation exchange energy with matter. These processes can be understood by considering the transfer of visible light through a gas made up of atoms. Similar processes occur when radiation interacts with solid dust particles or when it is transmitted through solids or liquids. *See* PHOTON.

If a gas is hot, collisions between atoms can convert the kinetic energy of motion to potential energy by raising atoms to an excited state. Emission is the process which releases this energy in the form of photons and cools the gas by converting the kinetic energy of atoms to energy in the form of radiation. The reverse process, absorption, occurs when a pho-
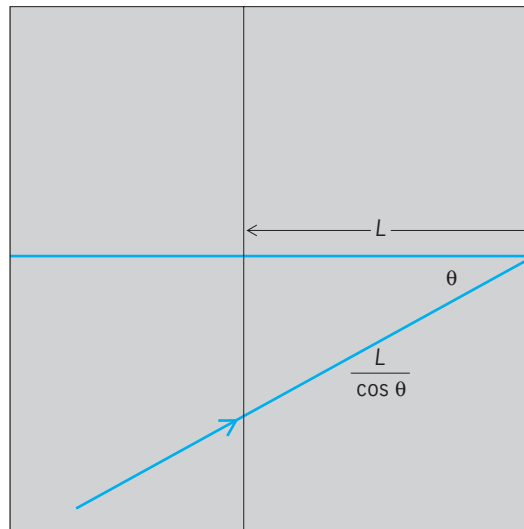

Fig. 1. Beam of light illuminating a slab of gas of thickness *L*, shown in cross section.

ton raises an atom to an excited state, and the energy is converted to kinetic energy in a collision with another atom. Absorption heats the gas by converting energy from radiation to kinetic energy. Occasionally an atom will absorb a photon and reemit another photon of the same energy in a random direction. If the photon is reradiated before the atom undergoes a collision, the photon is said to be scattered. Scattering has no net effect on the temperature of the gas. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; ATOMIC STRUCTURE AND SPECTRA; SCATTERING OF ELECTROMAGNETIC RADIATION.

Equation of radiative transfer. The equation of radiative transfer is built up by considering the three processes—absorption, emission, and scattering. First, however, it is necessary to define some geometry and, in particular, to consider an infinitely wide slab of gas (**Fig. 1**). Radiation enters the slab at some angle θ. If the slab has a thickness L, the radiation must cross a distance L divided by the cosine of the angle θ; that is, the distance traveled is $L/\cos\theta$. It is common to write the cosine of the incidence angle as μ; that is, $\mu = \cos\theta$. In this notation, the distance crossed is L/μ.

Absorption. First, one may consider a gas that absorbs radiation but does not scatter or emit it. This model is an idealization since such a gas is gaining energy which it must eventually radiate away. However, it is a reasonably good approximation of a gas that emits radiation at a much lower energy than the incident radiation.

Figure 2 shows a slab of gas divided into 5 layers; each layer is assumed to absorb half the radiation impinging on it. If the gas is illuminated from the left with a beam of light of some intensity, for example, I_0, the solid dots show how much radiation reaches each layer. If there are 10 layers each absorbing 25% of the incident energy, then the energy falls off as shown by the circles.

The curve shows the intensity at each point if this process is continued to the limit of extremely
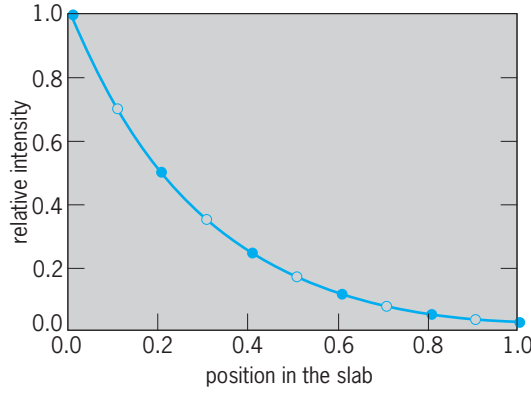
Fig. 2. Absorption of radiation in a slab of gas. The solid circles show the amount of radiation reaching each interface if each layer absorbs 50% of the radiation reaching it. The open circles are for 10 layers, each absorbing 25% of the radiation. The curve shows the result in the limit of a continuous medium.

thin layers absorbing a constant fraction of the incident energy per unit length.

The absorption illustrated in Fig. 2 can be expressed as Eq. (1) where I_n is the intensity leaving

$$I_{n+1} - I_n = -I_n \kappa \frac{\Delta x}{\mu} \qquad (1)$$

layer n, $\Delta x/\mu$ is the distance the radiation travels in crossing the layer, and κ is the absorption coefficient or opacity, a quantity that measures the fraction of the energy absorbed per unit length. On the average, a photon will travel a distance $1/\kappa$ before being absorbed, and $1/\kappa$ is called the mean free path. The quantity I is called the specific intensity and measures the amount of energy crossing a unit area per unit time per unit solid angle in a given direction. Typical units are joules per square centimeter per second per steradian.

Taking the limit of Eq. (1) as the layers get thinner yields the differential equation (2). If the opac-

$$\mu \frac{dI}{dx} = -\kappa I \qquad (2)$$

ity κ does not depend on position in the slab, the gas is said to be homogeneous, and Eq. (2) has the solution given by Eq. (3). In other words, the inten-

$$I = I_0 e^{-\kappa x/\mu} \qquad (3)$$

sity decreases exponentially with distance from the boundary of the slab; the opacity determines how rapidly the intensity falls off.

Since the solution depends only on the product of the opacity and distance, the optical depth of a point a distance x into the gas is defined as $\tau(x) = \kappa x$. The optical thickness of a layer [$\tau(L)$ in the example of Fig. 1] is a measure of the absorbing power of the gas. A thin layer of gas with a large absorption coefficient can absorb as much energy as a thick layer of gas with a small absorption coefficient.

Even if the opacity depends on position, Eq. (2) can be solved by combining κ and dx and calling the product $d\tau$. Equation (2) becomes Eq. (4), which has the solution given by Eq. (5). In order to compute

$$\mu \frac{dI}{d\tau} = -I \qquad (4)$$

$$I = I_0 e^{-\tau/\mu} \qquad (5)$$

the optical depth from the opacity and the physical depth, it is necessary to add the contributions of each thin piece of the gas; that is, integration is carried out over distance to obtain Eq. (6).

$$\tau(x) = \int_0^x \kappa(z)\, dz \qquad (6)$$

All the details of the absorption process are contained in the absorption coefficient κ. For example, the opacity normally depends on wavelength, which means that both the optical depth and specific intensity depend on wavelength.

Emission. It follows from Kirchhoff's law of radiation that if matter absorbs energy it must either radiate away the energy or heat up. In many radiative transfer problems the emission of energy can be ignored because it is in a different wavelength domain than the one under study. For example, in the study of the transfer of visible radiation through the Earth's atmosphere, the emitted radiation, which is primarily in the infrared part of the spectrum, is usually of no concern. However, there are many situations in which both the absorption and emission of radiation must be considered.

In the simplest approximation, the hot gas radiates like a blackbody. The emission is isotropic and has an intensity, denoted $B_\lambda(T)$, given by the Planck radiation law. The subscript λ indicates that the emission is wavelength-dependent, while the argument T indicates the temperature dependence of the emission. The equation of radiative transfer now becomes Eq. (7). In the following discussion, the subscript λ

$$\mu \frac{dI_\lambda}{d\tau_\lambda} = -I_\lambda + B_\lambda(T) \qquad (7)$$

will be omitted, although most quantities depend on wavelength. *See* HEAT RADIATION.

If the temperature through the gas is constant, Eq. (7) has the solution given by Eq. (8). This equa-

$$I = I_0 e^{-\tau/\mu} + B(T)(1 - e^{-\tau/\mu}) \qquad (8)$$

tion is used in astronomy to determine the properties of interstellar clouds, and in the laboratory to find the atomic properties of gases.

If the temperature is not constant, the situation is more complicated. The properties of the gas can still be analyzed, however, if Eq. (8) is replaced by Eq. (9). The equation indicates that the radiation at

$$I = I_0 e^{-\tau/\mu} + \int_0^\tau B(T) e^{-t/\mu} \frac{dt}{\mu} \qquad (9)$$

a point in the slab is made up of two parts. The

first part is the incoming radiation that has not yet been absorbed ($I_0 e^{-\tau/\mu}$). The second part is the radiation emitted by the hot gas. Since radiation emitted far away will be attenuated more than that emitted nearby, Eq. (9) contains a term $e^{-\tau/\mu}$. In addition, the radiation reaching a point comes from all over the gas so it is necessary to add up the contributions, that is, to integrate over all optical depths.

Scattering. When radiation interacts with matter, there is a probability that it will simply change direction. If there is no exchange of energy between matter and radiation, the scattering is called conservative.

In its simplest form, scattering is isotropic; that is, the direction of the radiation after the scattering is independent of its incident direction. If the probability that the radiation is scattered is denoted by ω, then the probability that it is absorbed is $1 - \omega$. The equation of radiative transfer becomes Eq. (10), where the source function S is defined by Eq. (11), and $d\Omega$ denotes that the integration is over all directions.

$$\mu \frac{dI}{d\tau} = -I + S \qquad (10)$$

$$S = (1 - \omega)B(T) + \omega \int I \, d\Omega \qquad (11)$$

The source function S contains two contributions. The first is the thermal emission from the hot gas, $(1 - \omega)B(T)$. The second term is needed because the specific intensity I is a measure of energy traveling in a specific direction. The integral in Eq. (11) counts the contribution of radiation that was traveling in any direction before it was scattered but is now traveling in the particular direction being considered. Since the source function depends on the intensity, there is no simple analytic solution to Eq. (10).

Nature of solutions. The solutions to the transfer equation are quite different, depending on what processes are significant. If the slab is very thick, the amount of radiation at a given layer decreases exponentially with depth for pure absorption, approaches $B_\lambda(T)$ if both absorption and emission are considered, and decreases linearly with depth if only scattering is important.

Only the simplest form of the equation of radiative transfer has been presented. Since the problems are so formidable, each discipline has concentrated on a particular set of approximations and there is relatively little crossover from one field to another.

Scattering dominated problems. As noted above, the transfer of visible radiation through the Earth's atmosphere is dominated by scattering. In other words, the single-scatter albedo ω is near one; that is, the scattering often is nearly conservative. If the radiation left a particle in a random direction on each scattering, that is, if isotropic scattering prevailed, then the solution of Eq. (10) would be tedious but not difficult. In fact, there is an analytic solution in this case if the gas is an infinitely wide slab.

The probability that a photon will be scattered into

a given direction is described by the scattering phase function $P(\Theta)$, where Θ is the angle between the incident and outgoing radiation. The phase function is determined by the size of the particles relative to the wavelength of light being scattered and by the composition of the particles. Metallic particles have very different properties from the nonconducting particles that are found in the atmosphere.

If the scattering is not isotropic, Eq. (11) is written as Eq. (12). The factor $P(\Theta)$ under the integral

$$S = (1 - \omega)B(T) + \omega \int P(\Theta)I \, d\Omega \qquad (12)$$

accounts for the fact that photons traveling in certain directions have a greater probability of being scattered into the direction of interest than do others.

If the scattering particles are very small compared to the wavelength of the radiation, the scattering is described by the Rayleigh phase function, Eq. (13),

$$P(\Theta) = {}^3/_4 (1 + \cos^2 \Theta) \qquad (13)$$

shown in **Fig. 3**. Equal amounts of energy are scattered into the forward ($\Theta = 0°$) and backward ($\Theta = 180°$) directions, and less energy is scattered perpendicular to the incident direction ($\Theta = 90°$).

When the scattering particles are approximately the same size as the wavelength of the light being scattered, the phase function is more complicated. Figure 3 also shows a phase function typical of rain clouds. The value of the phase function at $\Theta = 0°$ is nearly 25, well off the scale used for the figure. In other words, the scattering changes the direction of travel of only a small amount of radiation; most of it is scattered straight ahead. Since absorption in clouds is quite low, even very thick clouds can transmit an appreciable amount of light. *See* METEOROLOGICAL OPTICS.

Radiation and atomic physics. A completely different set of problems arises if the radiation changes the absorption and scattering properties of the gas. The effect is most apparent when atomic or molecular transitions in diffuse gases are considered.
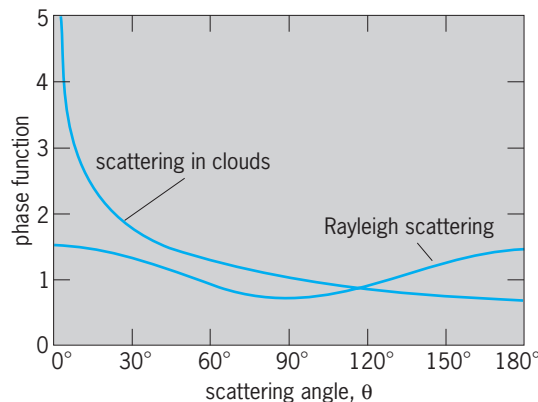

Fig. 3. Representative scattering phase functions, representing the probability that a photon will be scattered through an angle Θ.

If the gas is dense, collisions between atoms are frequent, and the distribution of electron energies depends only on the temperature. Such a gas is said to be in thermodynamic equilibrium. If the conditions of thermodynamic equilibrium are approximately satisfied in some region, the gas is said to be in local thermodynamic equilibrium, and the effect of radiation on the energy distribution of the electrons can be ignored. *See* KINETIC THEORY OF MATTER.

If the gas density is low enough, the distribution of electron energies depends on the radiation as well as the temperature. Since the absorption and scattering properties of the gas depend on the electron distribution, the radiation and electron distribution equations are coupled. Such a gas is said to be in a state of nonlocal thermodynamic equilibrium.

To further complicate matters, atoms can be perturbed while in an excited state. Therefore, electrons in different atoms jumping between the same pair of energy levels emit photons at slightly different wavelengths. The spectral line shape must then be taken into account, and the radiative transfer equation must be solved for a number of frequencies simultaneously. In the simplest situation, Eq. (10) becomes Eq. (14), where ϕ_λ, the spectral line shape, is

$$\mu \frac{dI_\lambda}{d\tau_\lambda} = -I_\lambda + \int_0^\infty \phi_\lambda S_\lambda \, d\lambda \qquad (14)$$

the probability that the electron transition results in a photon of wavelength λ. *See* LINEWIDTH; PERTURBATION (QUANTUM MECHANICS).

Complicated geometries. If the scattering phase function is not simple, or if the radiation and gas properties are coupled, it is impractical to try to solve the radiative transfer equation in geometrically complicated domains. However, in nuclear reactors it can normally be assumed that the scattering is either isotropic or nearly so, and that the absorption and scattering properties of the medium are independent of the neutron distribution. These assumptions make it possible to model reactors with hundreds or thousands of elements. The equation of radiative transfer can be written in the quite general form of Eq. (15), where c is the neutron velocity, $\hat{n}$ is a unit

$$\frac{1}{c} \frac{dI}{dt} + \hat{n} \cdot \nabla I = I - S \qquad (15)$$

vector, and ∇ denotes the gradient operator. Since reactor designers are interested in how the neutron flux changes with time, they include the time derivative as done in Eq. (15). *See* CALCULUS OF VECTORS.

One reason nuclear reactors generate energy is that each collision between an atom and a neutron can result in the emission of additional neutrons. In terms of the radiative transfer equation, this means that the single-scatter albedo ω is greater than unity. A part of the reactor is said to be critical when $\omega \geq 1$. Other regions can be subcritical, that is, $\omega < 1$. The goal of reactor design is to balance the critical and subcritical domains to meet the desired energy production. *See* REACTOR PHYSICS.

Difficulties. The modeling of radiative transfer is difficult. Problems with complex geometries that have complicated scattering phase functions and in which the absorption and scattering properties depend on the radiation are too hard to solve even with the largest computers available.

One simple problem that has not yet been solved accurately is the calculation of the radiation emerging from a single cloud in an otherwise clear sky. Nor has an accurate solution been presented for the transfer of radiation through a clear sky that includes the effects of horizontal variations in the reflectivity of the ground.

Even if such problems could be solved, more difficult problems remain. For example, radiation can be polarized, that is, the waves making up the light are lined up. The intensity of polarized radiation has four distinct components, so that Eq. (10) must be replaced by four equations. Furthermore, calculating the scattering phase function is considerably more difficult than if polarization is ignored. Only in the 1980s was substantial progress made on this problem. *See* POLARIZED LIGHT. Alan H. Karp

Bibliography. M. Q. Brewster, *Thermal Radiative Transfer and Properties*, 1992; S. Chandrasekhar, *Radiative Transfer*, 1960; J. Lenoble, *Atmospheric Radiative Transfer*, 1993; D. Mihalas, *Stellar Atmospheres*, 2d ed., 1978; M. F. Modest, *Radiative Heat Transfer*, 1993; H. C. van de Hulst, *Light Scattering by Small Particles*, 1957, reprint 1982; H. C. van de Hulst, *Multiple Light Scattering*, 2 vols., 1980.

Radiator

Any of numerous devices, units, or surfaces that emit heat, mainly by radiation, to objects in the space in which they are installed. Because their heating is usually radiant, radiators are of necessity exposed to view. They often also heat by conduction to the adjacent thermally circulated air. *See* HEAT RADIATION.

Radiators are usually classified as cast-iron (or steel) or nonferrous. They may be directly fired by wood, coal, charcoal, oil, or gas (such as stoves, ranges, and unit space heaters). The heating medium may be steam, derived from a steam boiler, or hot water, derived from a water heater, circulated through the heat-emitting units.

Cast-iron radiators are made in sections of varying widths and heights and are assembled in the required number by top and bottom nipples. Some are made in a flat panel. They are set on legs or similar supports or affixed to walls by adjustable hangers. The preferred location is under windows.

When windows are low, cast-iron baseboard radiators, less than a foot high, are available for use under the windows. Because of the limited heat from such a radiator, it is frequently extended along the nonwindow baseboard. This radiator is assembled with valve and end sections, extension blocks, and inverted and

projecting corner covers. Cast-iron convectors may be used also.

Finned-tube radiators consist of a pipe or tube with affixed fins of steel, copper, or aluminum. They are available with or without enclosures, and usually are wall-mounted. They are more compact than cast-iron radiators, having greater heating area per volume.

Mass production was responsible for the introduction of convectors, which, like finned-tube radiators, consist of steel, copper, or aluminum tubes with extended fin surfaces affixed to them. They have almost completely displaced cast-iron radiators, for the fact that they are lightweight causes freight costs to be lower. They also require less labor to install. They cannot, however, be used as replacements in an existing hot-water heating system with cast-iron radiators for heating surfaces, because the heat-dissipating characteristics of convectors are completely different from those of cast-iron radiators with changes in hot-water temperatures.

Convectors are customarily made of nonferrous finned tubes in a wide variety of enclosures for freestanding, recessed, or wall-hung installation. Air is circulated over the elements by natural convection or by fans, in which case they are known as unit ventilators. Convectors and unit ventilators are not true radiators because they emit most of the heat by conduction to the circulated air.

A limited amount of cooling may be produced by passing chilled water through convectors or radiators, but care must be exercised to dispose of moisture condensation in humid weather.

Electric heating elements may be substituted for fluid heating elements in all types of radiators, convectors, and unit ventilators. *See* COMFORT HEATING; ELECTRIC HEATING; HOT-WATER HEATING SYSTEM; RADIANT HEATING; STEAM HEATING.

Erwin L. Weber; Richard Koral

Bibliography. American Society of Heating, Refrigerating and Air Conditioning Engineers, *ASHRAE Handbook*: *Equipment*, 1988.

Radio

Communication between two or more points, employing electromagnetic waves ("carriers") as the transmission medium. *See* ELECTROMAGNETIC RADIATION.

At the start of the twenty-first century, radio transmission services are in the midst of a fundamental transition from traditional (and increasingly antiquated) analog technology to highly sophisticated and advanced digital technology. The basic methods of analog transmission and reception of audio signals were developed principally in the twentieth century (**Fig. 1**). For transmission, an audio pulse and a carrier pulse are blended into a modulated carrier wave, which is then amplified and fed into the antenna. The process of blending audio and carrier waves is called modulation. For reception, the receiving antenna and tuner catch

the weak signal, amplify it, sort the audio pulse from the carrier, and play a now reamplified audio pulse through the speaker at home. Digital radio techniques differ from analog in the nature of the information being transmitted (audio represented digitally, as a sequence of 1's and 0's) and the methods of modulation employed. Digital techniques are significantly more flexible, robust (that is, less susceptible to interference and other deleterious effects), and more efficient than the analog techniques they replace. *See* ANTENNA (ELECTROMAGNETISM); ELECTRICAL COMMUNICATIONS; MODULATION; RADIO RECEIVER; RADIO TRANSMITTER; RADIO-WAVE PROPAGATION.

Methods of information transmission. Radio waves transmitted continuously, with each cycle an exact duplicate of all others, indicate that only a carrier is present (also called an unmodulated carrier). The two principal characteristics of a carrier are its frequency (the number of cycles per second) and amplitude (which determines how much power is contained in the signal). The information desired to be transmitted must cause changes in the carrier which can be detected at a distant receiver. The method used for the transmission of the information is determined by the nature of the information which is to be transmitted as well as by the purpose of the communication system (**Fig. 2**).

Code transmission. This method, the first ever used, dates from the dawn of radio, and was patented by Gugliemo Marconi in 1896. The carrier is keyed on and off to form dots and dashes (Morse code). The technique, traditionally used in ship-to-shore and amateur communications, has been almost entirely superseded by more efficient methods. This primitive method is actually digital (binary) in nature, since only one of two possible symbols is transmitted at any given moment. *See* TELEGRAPHY.

Frequency-shift transmission. The frequency of the carrier is shifted a fixed amount to correspond with telegraphic dots and dashes or with combinations of pulse signals identified with the characters on a typewriter. This technique was widely used in handling the large volume of public message traffic on long circuits, principally by the use of teletypewriters. *See* TELETYPEWRITER.
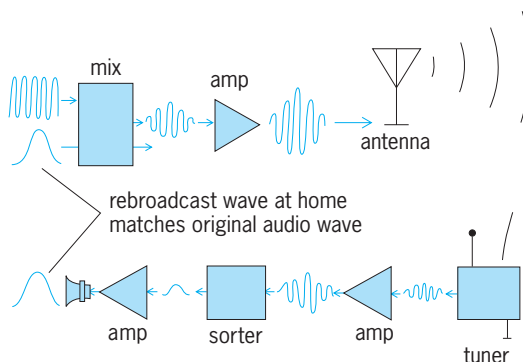

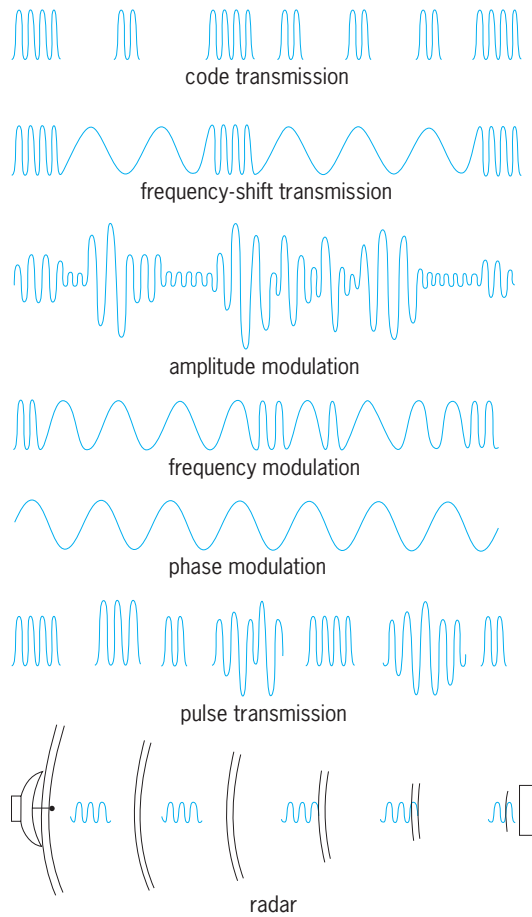Fig. 1. Transmission of audio information by radio.

Fig. 2. Methods for transmission of information.

Amplitude modulation. The amplitude of the carrier is varied in a way that conforms to the fluctuations of a sound wave, or, in the case of digital transmission, the specific sequence of 1's and 0's being transmitted. This technique is used in amplitude-modulation (AM) terrestrial radio broadcasting, analog television picture transmission, and many other services. One widely used digital method, called quadrature AM (QAM), is employed in some terrestrial digital radio systems. *See* AMPLITUDE MODULATION; AMPLITUDE-MODULATION DETECTOR; AMPLITUDE-MODULATION RADIO; AMPLITUDE MODULATOR; TELEVISION.

Frequency modulation. The frequency of the carrier is varied in a way that conforms to the fluctuations of the modulating wave. This technique is used in frequency-modulation (FM) broadcasting, analog television sound transmission, and analog microwave relaying. *See* FREQUENCY MODULATION; FREQUENCY-MODULATION DETECTOR; FREQUENCY-MODULATION RADIO; FREQUENCY MODULATOR; MICROWAVE.

Phase modulation. The phase of the carrier is varied in a way that conforms to the fluctuations of the modulating wave. A family of popular digital methods, known as phase-shift keying (PSK), are used in digital satellite broadcasting. Specific forms include binary PSK (BPSK), quadrature PSK (QPSK), and octal PSK (8PSK), differentiated by the number

of digital bits used to create each transmitted symbol. *See* PHASE MODULATION; PHASE-MODULATION-DETECTOR; PHASE MODULATOR.

Pulse transmission. The carrier is transmitted in short pulses, which change in repetition rate, width, or amplitude, or in complex groups of pulses which vary from group to succeeding group in accordance with the message information. These forms of pulse transmission are identified as pulse-code, pulse-time, pulse-position, pulse-amplitude, pulse-width, or pulse-frequency modulation. These complex techniques are employed principally in microwave relay and telemetry systems. *See* PULSE DEMODULATOR; PULSE MODULATION; PULSE MODULATOR.

Radar. The carrier is normally transmitted as short pulses in a narrow beam, similar to that of a searchlight. When a wave pulse strikes an object, such as an aircraft, energy is reflected to the radar station, which measures the round-trip time and converts it to distance. A radar receiver can display varying reflections in a maplike presentation on a video display. *See* RADAR.

Uses of radio. The first practical application of radio, in the 1900s between ships and shore stations, was followed quickly by overseas communication and communication between other widely separated points. Subsequently, the applications have become widely diversified. Some of the important uses are listed below.

1. Public safety: marine and aviation communications, police and fire protection, forestry conservation, and highway traffic control.

2. Industrial: power utilities, pipelines, relay services, news systems, financial institutions, agriculture, and petroleum processing.

3. Land transportation: railroads, motor carriers, urban public transit and taxicabs, and automobile emergency needs.

4. Broadcasting: television (terrestrial and satellite), terrestrial radio broadcasting (FM, AM, digital), satellite digital audio radio services (SDARS), and shortwave broadcasting. *See* RADIO BROADCASTING; RADIO BROADCASTING NETWORK; SATELLITE RADIO; SATELLITE TELEVISION BROADCASTING.

5. Military and space: radar, communications, navigation, telemetering, missile tracking, satellite photographic surveys, and missile guidance. *See* ELECTRONIC NAVIGATION SYSTEMS; REMOTE SENSING; SPACE COMMUNICATIONS.

6. Fixed point-to-point: long-distance message, data, and picture transmission.

7. Relaying: television, sound, data, picture, and public message relaying over long distances. Communications satellites handle both domestic and overseas communications. These satellites supply not one but many services, including thousands of narrow channels for voice communications with wider channels for television and data transmission. *See* COMMUNICATIONS SATELLITE.

8. Telemetering: remote indication of water levels in reservoirs and rivers, performance of

experimental aircraft, missiles, and satellites. *See* TELEMETERING.

9. Weather reporting: early warning and location of hurricanes and other storms, trends of weather for industrial and public information. *See* WEATHER FORECASTING AND PREDICTION.

10. Interpersonal: cellular telephones, personal communications services, amateur radio services, radio paging systems, interconnection (for example, Bluetooth®, Wi-Fi®, WiMax). *See* AMATEUR RADIO; MOBILE RADIO; RADIO PAGING SYSTEMS.

Frequency separation. Hundreds of thousands of radio transmitters exist, each requiring a carrier at some radio frequency. To prevent interference, different carrier frequencies are used for stations whose service areas overlap, and receivers are built which can select the carrier signal of the desired station. Resonant electronic circuits in the receiver are adjusted, or tuned, to accept the desired frequency and reject others. Digital technology, in particular the availability of high-speed, low-cost microprocessors and analog-to-digital converters, has made possible the development of software radios which select and demodulate radio signals according to a computer program which runs on a microprocessor. Software radios are more flexible and offer higher performance than do traditional radios.

Each station operates within a specific radio-frequency (RF) channel. All other stations within a geographical area are excluded from using this channel. However, an advanced digital technology called code-division multiple access (CDMA) allows multiple stations to operate on the same frequency without interfering with one another. CDMA technology has found its principal application in some cellular telephone systems. Another digital technology, time-division multiple access (TDMA), allows stations to share a frequency by precisely coordinating the time at which each station transmits, making sure that only a single station is transmitting at any one time. A new technology called ultra-wideband (UWB) spreads the information over many thousands of megahertz, allowing it to operate at such a low power level per megahertz that it does not interfere significantly with other overlapped services. *See* MULTIPLEXING AND MULTIPLE ACCESS; UL-TRAWIDEBAND (UWB) SYSTEMS.

Each channel must be wide enough to accommodate the message information, provide tolerance for small carrier frequency drift, perhaps provide a guard band, and allow for imperfect receiver selectivity capabilities. The minimum usable channel widths (or bandwidths) vary from service to service, depending upon the amount of information a channel must accommodate. In analog television it is 6000 kHz because of the large amount of essential picture information. In FM radio broadcasting it is 200 kHz and in AM radio broadcasting 10 kHz. Digital systems can accommodate corresponding services with smaller bandwidths. For example, digital television services can accommodate multiple video programs in the same 6000-kHz bandwidth required for a single video program using analog services. The great demand for new frequency authorizations requires efficient channel utilization and has been a major factor in the drive to develop and deploy digital services. *See* BANDWIDTH REQUIREMENTS (COMMUNICATIONS); RADIO SPECTRUM ALLOCATIONS; TELEVISION.

Regulation of spectrum usage. All nations have a sovereign right to use freely any or all parts of the radio spectrum. But numerous international agreements and treaties divide the spectrum and specify sharing among nations for their mutual benefit and protection.

Each nation designates its own regulatory agency. Functioning within the international agreements, it issues authorizations; assigns frequencies; polices operations; creates technical standards, rules, and practices; and safeguards and protects the public interest.

In the United States all nongovernmental radio communications are regulated by the Federal Communications Commission (FCC), according to the provision of the Communications Act of 1934, as amended. Creation of a radio station or service requires authorization by the FCC. Upon construction of an authorized facility, a license to operate is issued. Radio stations may be inspected by engineers attached to FCC field offices. Stations must comply with the terms of their authorization regarding carrier-frequency tolerance, power limitations, permissible communications, calls signals, and control by properly licensed personnel.

Governmental use of spectrum in the United States, which is extensive, is regulated by the National Telecommunications and Information Administration (NTIA), an agency which also serves as the principal advisor to the President on telecommunications and information policy issues. In managing spectrum the NTIA's stated goals include serving the national security and defense, supporting crime prevention and law enforcement, and promoting scientific research, development, and exploration.

The International Telecommunication Union (ITU) is a specialized agency of the United Nations with headquarters in Geneva, Switzerland. It is responsible for the international agreements necessary to assure provision of radio and telecommunications services among the nations of the world.

The ITU is divided into three sectors. The radio communication sector (ITU-R) is responsible for the allocation of radio spectrum, allocation of orbital slots used by international satellites, and establishment of various technical characteristics of operating radio systems. The telecommunication standardization sector (ITU-T) is responsible for developing internationally agreed technical and operating standards and defining tariff and accounting principles for international telecommunications services. The telecommunication development sector (ITU-D) promotes investment and fostering the expansion of telecommunication infrastructure in developing nations throughout the world. Additional ITU functions include the maintenance of an

international registry of radio transmitters, and the availability to facilitate the resolution of international disputes on uses of the radio spectrum among member countries.

David H. Layer; John D. Singleton; Michael C. Rau

Bibliography. American Radio Relay League, *The ARRL Handbook for Radio Amateurs*, rev. ed., annually; G. Jones, *A Broadcast Engineering Tutorial for Non-Engineers*, NAB and Focal Press, 2005; J. G. Proakis and M. Salehi, *Communication Systems Engineering*, 2d ed., Prentice Hall, 2001; T. S. Rappaport, *Wireless Communications: Principle and Practice*, 2d ed., Prentice Hall PTR, 2001; C. H. Sterling and J. M. Kittross, *Stay Tuned—A History of American Broadcasting*, 3d ed., Lawrence Erlbaum Associates, 2002; M. E. Van Valkenburg and W. M. Middleton (eds.), *Reference Data for Engineers: Radio, Electronics, Computer, and Communications*, Newnes, 9th ed., 2002; J. Whitaker (ed.), *NAB Engineering Handbook*, 9th ed., National Association of Broadcasters, 1999.

Radio astronomy

The study of celestial bodies by examination of the energy they emit at radio frequencies. Celestial radio noise originates by one of several processes. These include both broadband continuum radiation owing to (1) thermal radiation from solid bodies such as the planets, (2) thermal or bremsstrahlung radiation from hot gas in the interstellar medium, (3) synchrotron radiation from ultrarelativistic electrons moving in weak magnetic fields, (4) coherent processes as found in the Sun and on Jupiter, and (5) pulsed radiation from the rapidly rotating neutron stars, as well as (6) narrow "spectral line" radiation from atomic or molecular transitions that occur in the interstellar medium or in the gaseous envelopes around stars.

Radio astronomy observations cover the entire radio spectrum from less than 1 mm, where the cosmic signals become heavily attenuated by the atmosphere, to wavelengths of tens of meters, beyond which the incoming signals are attenuated by the ionosphere. However, radio waves between about 1-cm and 1-m wavelength not only get through the Earth's atmosphere and ionosphere with little distortion, they also penetrate much of the gas and dust in space as well as the clouds of planetary atmospheres. Radio astronomy can therefore sometimes give a much clearer picture of stars and galaxies than is possible by means of conventional observation using visible light. Sophisticated antennas equipped with very sensitive radio receivers that are used to detect cosmic radio emission are referred to as radio telescopes. *See* RADIO TELESCOPE.

Spectral line emission occurs at specific wavelengths characteristic of the atomic and molecular species. However, due to the motion of gas clouds, the wavelength is shifted toward longer wavelengths if the source is moving away from the observer and toward shorter wavelengths if it is moving toward the observer. Precise measurements of this so-called Doppler shift are used to determine the radial velocities due to the random motion of gas clouds, the rotation of the Galaxy, and even the expansion of the universe. *See* DOPPLER EFFECT; REDSHIFT.

Historical background. In 1932, the physicist Karl Jansky, working at the Bell Telephone Laboratories, first detected cosmic radio noise from the center of the Milky Way Galaxy while investigating radio disturbances interfering with transoceanic telephone service. Grote Reber, a young engineer and amateur radio operator, later built the first radio telescope at his home in Wheaton, Illinois, and found that the radio radiation came from all along the plane of the Milky Way and from the Sun. Solar radio emission was also detected during the World War II by military researchers in the United States and the United Kingdom but was not announced until after the conclusion of hostilities due to wartime secrecy.

During the late 1940s and 1950s, Australian and British radio scientists were able to use techniques and instrumentation developed for wartime radar to locate a number of discrete sources of celestial radio emission which they associated with old supernovae and active galaxies later known as radio galaxies. The construction of ever larger antenna systems and radio interferometers, improved radio receivers, and data processing methods have allowed radio astronomers to study fainter radio sources with increased angular resolution and improved image quality. Today more than 1 million sources of radio emission are known. The precise measurement of their position in the sky has usually allowed their identification with an optical, infrared, x-ray, or gamma-ray counterpart; however, some cosmic radio sources remain unidentified at other wavelengths even with the most sensitive instrumentation.

Solar system radio astronomy. Solar flares and sunspots are strong sources of radio emission. Their study has led to increased understanding of the complex phenomena near the surface of the Sun and provides advanced warning of dangerous solar flares which can interrupt radio communications on the Earth and endanger sensitive equipment in satellites and the health of astronauts. Even in the absence of activity, the quiet Sun is the strongest radio source in the sky, but only because it is so close to the Earth. *See* SUN.

Radio telescopes are used to measure the surface temperatures of all the planets as well as some of the moons of Jupiter and Saturn. Radio studies of the planets have revealed the existence of a greenhouse effect on Venus which results in a surface temperature of about 600°F (315°C), intense Van Allen belts surrounding Jupiter, powerful radio storms in the Jovian atmosphere, and an internal heating source deep within the interiors of Jupiter, Saturn, Uranus, and Neptune. *See* JUPITER; NEPTUNE; SATURN; URANUS; VAN ALLEN RADIATION; VENUS.

Astronomers also use radar observations to image features on the surface of Venus, which is completely obscured from visual scrutiny by the heavy cloud cover that permanently enshrouds the planet.

Terrestrial-based radar measurements also have re- vealed the rotation of Mercury, which was previ- ously thought to always keep the same side toward the Sun. Accurate measurements of the travel time of radar signals reflected from Venus near superior conjunction (that is, nearly opposite the Earth on the other side of the Sun) show that radio waves pass- ing close to the Sun slow down owing to gravity and thereby provide a new independent test of Einstein's general theory of relativity. *See* MERCURY; RADAR AS- TRONOMY; RELATIVITY; SOLAR SYSTEM.

Radio emission from the Milky Way. Broadband con- tinuum radio emission is observed throughout the radio frequency spectrum from a variety of stars including flare stars, binary stars, x-ray binaries, novae, and supernovae; from supernova remnants; and from magnetic fields and relativistic electrons in the interstellar medium thought to be produced in supernova explosions (for example, **Fig. 1**). Bursts of radio emission are frequently observed from some binary star systems thought to be the result of mass transfer between the two components of the bi- nary system. The remnants of powerful supernovae explosions are often observed as bright sources of radio emission due to synchrotron radiation from rel- ativistic electrons accelerated by supernova events. One of the brightest observed radio sources in the Milky Way is the Crab Nebula, which is the remnant of a supernova observed by Chinese astronomers in the year 1054. Of particular interest to astronomers is the complex radio source at the Galactic Center known as Sagittarius A, which contains both ther- mal and nonthermal continuum radiation as well as a rich mixture of molecular and ionized hydrogen gas. High-resolution radio observations also show a weak, embedded, very compact radio source called Sgr A*, which is thought to lie at the precise center of the Galaxy (**Fig. 2**). Sgr A* is believed to be associ-
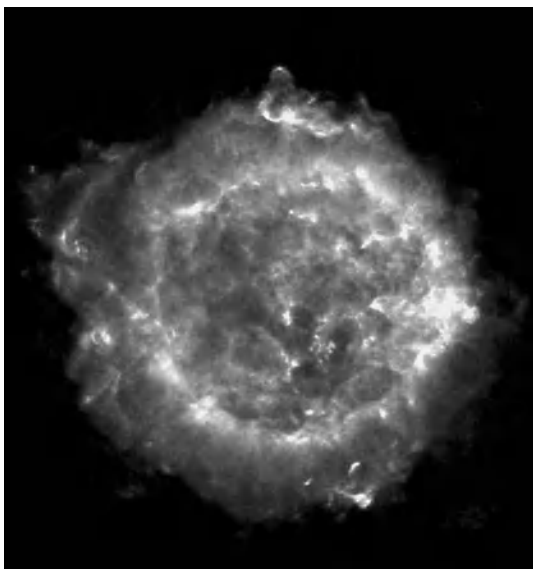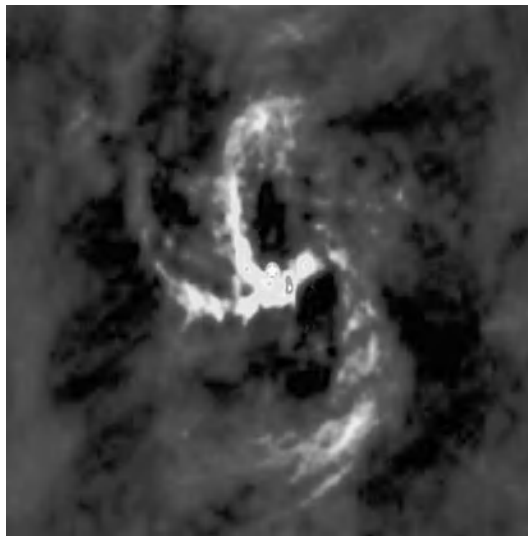

Fig. 2. Radio image of the center of our Milky Way Galaxy known as Sagittarius A. The bright spot is the very compact radio source known as SgrA*, which is associated with the million-mass black hole at the very center of the Galaxy.

ated with a black hole containing the mass equivalent of 2 million suns within a region only 5 solar diam- eters across. *See* BINARY STAR; CRAB NEBULA; MILKY WAY GALAXY; NOVA; SUPERNOVA; SYNCHROTRON RA- DIATION; X-RAY ASTRONOMY.

Atomic and molecular gas. The important spectral line of neutral atomic hydrogen (H I), which occurs at 1421.405 MHz (21 cm), was first observed in 1951. H I is observed throughout the Milky Way Galaxy as well as in other spiral galaxies, although the galactic hydrogen is heavily concentrated toward the plane of the Milky Way. The radial (along the line of sight) ve- locity of interstellar H I clouds is determined by mea- suring the shift in the wavelength of the observed lines arising from the Doppler effect. Measurements of the position and velocity of H I are used to trace out the spiral arms in the Milky Way system to study the rotation of the Galaxy.

H I is also observed in nearby galaxies. The ob- served dependence of rotation velocity on distance from the center of the galaxy is referred to as a ro- tation curve and is used to study the distribution of mass in galaxies. Observations of rotation curves in a number of spiral galaxies show the presence of dark matter located far beyond the extent of visible stars and nebula. The nature of this dark matter is unknown. It could be composed of ordinary matter in the form of planetlike objects, brown dwarfs, or neutron stars; or neutrinos; or some unknown form of matter. *See* DARK MATTER.

Molecular gas found in dense interstellar clouds radiates at radio wavelengths characteristic of each molecule. The wavelength of emission corresponds to transitions between different rotational and some- times vibrational energy levels of the molecular species. More than 100 different molecular species have been observed, only a few of which can be observed at optical wavelengths. These include car- bon monoxide, ammonia, water, ethyl and methyl

Fig. 1. Radio image, known as Cassiopeia A, of the remnants of a supernova which exploded about 300 years ago. (*NRAO/AUI/NSF*)

alcohol, formaldehyde, and hydrogen cyanide as well as some heavy organic molecules. Most of the observed molecules are found only in dense giant molecular clouds, especially near the center of the Galaxy. Carbon monoxide has been detected in some of the most distant known galaxies and quasars.

In the case of hydroxyl (OH), water (H_2O), and silicon oxide (SiO), collisions or ultraviolet radiation from nearby stars can "pump" the molecules to an excited state resulting in maser (microwave amplification by stimulated emission of radiation; a maser is the radio equivalent of a laser) emission. Maser emission is characterized by very narrow line widths and angular dimensions as small as one thousandth of a second of arc. Masers are found around both newly forming and dying stars. Very powerful H_2O masers found in other galaxies are known as megamasers. *See* ATOMIC STRUCTURE AND SPECTRA; GALAXY, EXTERNAL; INTERSTELLAR MATTER; MASER; MOLECULAR CLOUD; MOLECULAR STRUCTURE AND SPECTRA.

Pulsars. Pulsars (a term derived from pulsating radio stars) were discovered by Jocelyn Bell and Anthony Hewish in Cambridge, England, in 1967. Pulsars are neutron stars. Their electrons have all combined with protons in nuclei to form neutrons, and they have shrunk to a diameter of only a few kilometers following the explosion of the parent star in a supernova. Because they have retained the angular momentum of the much larger original star, neutron stars spin very rapidly, up to hundreds of times per second, and contain magnetic fields as strong as 10^8 tesla (10^{12} gauss) or more. The radio emission from pulsars is concentrated along a thin cone which produces a series of pulses corresponding to the rotation of the neutron star, much like a beacon from a rotating lighthouse lamp. Due to dispersion in the ionized interstellar gas, the arrival time of pulses observed at long wavelengths is delayed compared with short wavelengths, and observations of this dispersion measure are used to estimate the interstellar electron density and magnetic field strength. The rotation rates of pulsars are extremely stable. They are the most precise clocks known in nature with pulse repetition rates accurate to better than 1 second in a million years. Pulsars are found mostly close to the plane of the Milky Way, especially toward the direction of the center of the Galaxy.

Studies by Russell Hulse and Joseph Taylor of the slowing down of the pulsar PSR 1913+16 in close orbit about another neutron star led to the first observations of the effect of gravitational radiation. Precise measurements of the arrival times from the PSR 1257+12 in 1992 detected small perturbations due to the motion of planets orbiting the pulsar. This was the first detection of an extrasolar planetary system. *See* GRAVITATIONAL RADIATION; NEUTRON STAR; PULSAR.

Galaxies, quasars, and active galactic nuclei (AGN). Normal galaxies, particularly spiral galaxies, are sources of weak radio emission with a luminosity of about 10^{37} watts/Hz, comparable to the luminosity of our own Milky Way Galaxy. In some galaxies, regions of intense star formation release cosmic rays from supernovae causing intense radio emission which can be observed even at very great distances. Even more powerful are the radio galaxies, which can emit more than 100 million times more radio radiation than normal spiral galaxies do. Radio galaxies are surrounded by huge clouds of relativistic plasma located up to millions of light-years from the parent galaxy (**Fig. 3**) and containing electrons moving with nearly the speed of light in weak magnetic fields to produce the synchrotron radiation observed throughout the radio spectrum. By contrast, in the normal and star-forming galaxies, the radio emission comes from within the optically visible region of the galaxy.

In 1963, the study of radio galaxies led astronomer Maarten Schmidt to discover quasars. Quasars are
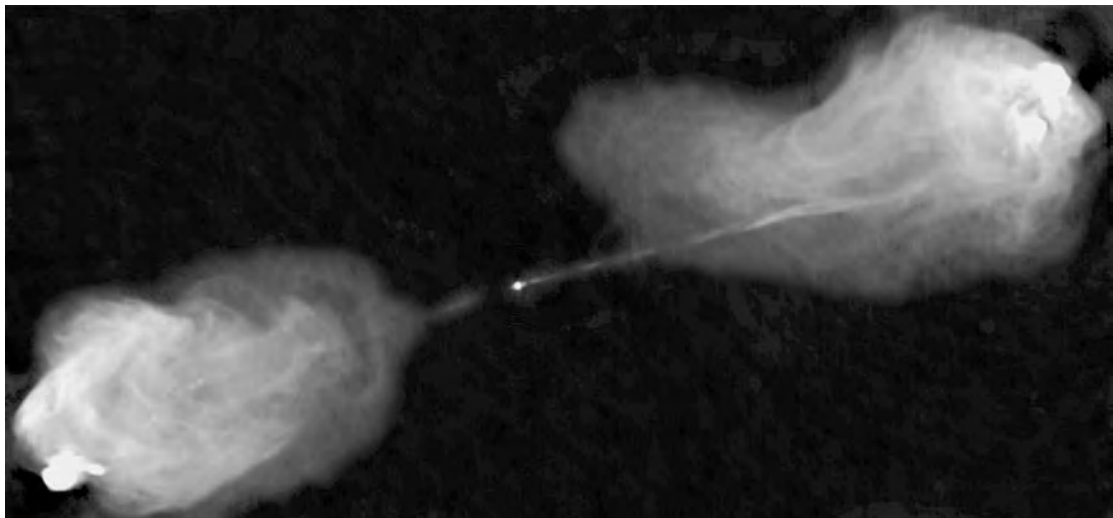

Fig. 3. Radio image of the radio galaxy Cygnus A showing the thin radio jet extending from the bright active galactic nucleus (AGN) toward the distant radio lobes. This 5-GHz (6-cm) image was made using the NRAO Very Large Array located near Socorro, New Mexico. (*NRAO/AUI/NSF*)

found in the central regions of rare galaxies and may shine with the luminosity of up to 100 ordinary galaxies. While some quasars also contain distant radio lobes, more often the radio emission from quasars is confined to a very small region coincident with the optically visible quasar. Very high resolution radio images may show relativistic plasma appearing to flow along a narrow jet away from the quasar with an apparent velocity greater than the speed of light. This so-called superluminal motion is an illusion due to the finite signal travel time from a relativistically moving source oriented close to the line of sight so that the radio source nearly keeps up with its own radiation. Most quasars are only weak radio sources, and the reason why only some 10 to 15% of all quasars become strong radio sources is unclear. Quasars which are not powerful radio sources are referred to as radio-quiet quasars. *See* QUASAR; SUPERLUMINAL MOTION.

The energy contained in the form of relativistic particles and magnetic fields surrounding radio galaxies and quasars is enormous, reaching in some cases as much as 10^{61} ergs. It is widely believed that the source of this energy is massive black holes at the center of the galaxy or quasar containing up to 10^9 times the mass of the Sun (see **colorplate**). Matter is accreted in an accretion disk surrounding the black hole. Electrons are heated to very high energy in the accretion disk and are then propelled out along a thin jet to form the distant radio-emitting plasma clouds (Fig. 3). The best direct evidence for the existence of a massive black hole comes from precise measurements of the angular velocity and acceleration of H_2O masers in the galaxy NGC 4258, leading to the conclusion that the masers are in orbit about a massive black hole containing 100 million times the mass of the Sun. The observed motions appear to accurately follow Kepler's laws of planetary motion, which allowed radio astronomers to more accurately determine the distance to the galaxy than previously possible. *See* BLACK HOLE; KEPLER'S LAWS.

The discovery of relativistic motion in radio jets was the foundation of the so-called unified models which postulate that the appearance of a radio source depends on its geometry as well as on its intrinsic properties. Special relativity shows that the radiation from a rapidly moving object is focused into a narrow cone along the direction of motion. For an object moving at 99% of the speed of light, the radiation can appear to be boosted by up to a factor of 1000 over an object at rest. The strong compact radio sources associated with quasars and active galactic nuclei (AGN) are thought to be the result of relativistic boosting of synchrotron radiation from plasma clouds ejected toward the observer from galactic nuclei. But when the plasma clouds are ejected in other directions, astronomers see only the radiation from the large, stationary, more distant clouds characteristic of radio galaxies.

Sometimes the observed radio emission from a distant AGN, quasar, or galaxy passes very close to a massive galaxy or cluster of galaxies, which bends both the radio and optical waves so that astronomers may see more than one image arriving from different directions. This phenomenon is known as gravitational lensing. Depending on the nature of the lensing galaxy and its angular separation from the remote source, there may be multiple images or just a single ring image, known as an Einstein ring, surrounding the lens. *See* GRAVITATIONAL LENS.

Since the first observations of large numbers of radio galaxies in the early 1960s, it has been clear that the number or luminosity of radio galaxies increases with increasing distance from our Galaxy. Due to the finite signal travel time of radio waves, this means that in the early universe, conditions were very different from the local (current epoch) universe, and such studies provided the earliest evidence that we are living in an evolving or changing universe.

Cosmic microwave background radiation. In 1965, Robert Wilson and Arno Penzias detected the faint comic microwave background (CMB) signal left over from the original big bang, which occurred some 14×10^9 years ago. As the universe expanded, it cooled to its present value of $2.7°$ above absolute zero. Subsequent observations with both ground- and space-based radio telescopes have detected the fine details in the CMB of less than one thousandth of a percent corresponding to the initial formation of structure in the early universe. These studies give new insight into the formation of galaxies and the rate of expansion of the universe, and show a surprising apparent acceleration of the expansion corresponding to an unknown form of dark energy.

When the faint CBR passes through a cluster of galaxies, some of the radiation may be scattered by relativistic electrons contained in the intergalactic gas, causing a slight diminution of the CMB observed in the direction of the cluster. This is known as the Sunyaev-Zeldovich effect. At the same time, the intergalactic electrons are then boosted in energy and radiate x-rays which are observed by astronomers using orbiting x-ray telescopes. Comparison of the x-ray emission, which depends on distance with the radio signal decrement, which is independent of distance, gives a direct measure of the distance to remote galaxies, independent of the conventional complex and less direct methods used by optical astronomers. *See* BIG BANG THEORY; COSMIC BACKGROUND RADIATION; COSMOLOGY.

Kenneth Kellermann

Bibliography. B. F. Burke and F. Graham Smith, *An Introduction to Radio Astronomy*, Cambridge University Press, 2d ed., 2002; J. D. Kraus, *Radio Astronomy*, 2d ed., Cygnus-Quasar Books, 1986; K. Rohlfs and T. Wilson, *Tools of Radio Astronomy*, 4th ed., Springer, 2003; G. L. Verschuur, *The Invisible Universe Revealed: The Story of Radio Astronomy*, Springer, 1986; G. L. Verschuur and K. I. Kellermann, *Galactic and Extragalactic Radio Astronomy*, Springer, 1991.

Radio broadcasting

The transmission, via radio-frequency electromagnetic waves, of audible program material for direct reception by the general public. Electromagnetic waves can be made to travel or propagate from a transmitting antenna to a receiving antenna. By modifying the amplitude, frequency, or relative phase of the wave in response to some message signal (a process known as modulation), it is possible to convey information from the transmitter to the receiver. In radio broadcasting, this information usually takes the form of voice or music. *See* ELECTRICAL COMMUNICATIONS; ELECTROMAGNETIC WAVE TRANSMISSION; MODULATION; RADIO.

Radio broadcasting occurs in seven frequency bands (**Fig. 1**). So-called longwave broadcasting is permitted by international agreement in a portion of the low-frequency band from 150 to 290 kHz in Europe. The most widely used broadcast band is in the medium-frequency (mf) range between 525 and 1700 kHz. It is commonly known as AM after amplitude modulation, the technique employed. So-called shortwave broadcasting is permitted worldwide in eight frequency bands between 5950 and 26,100 kHz. The very high frequency (VHF) band of 88 to 108 MHz is used for what is commonly called FM broadcasting, after frequency modulation that is used for transmissions. During the 1990s, a digital audio broadcasting (DAB) service in the 1452–1492- and 174–240-MHz frequency bands was put in place in Europe, Canada, and other countries. This band is unavailable in the United States for this service, so an alternate DAB system is being devised for use there. Radio broadcasting from satellites to listeners has been authorized in the 2310–2360-MHz frequency band. *See* AMPLITUDE MODULATION; AMPLITUDE-MODULATION RADIO; FREQUENCY MODULATION; FREQUENCY-MODULATION RADIO.

The frequencies set aside for radio broadcasting have been assigned by the International Telecommunications Union (ITU), an organization similar to the United Nations in which all countries may participate. The ITU is charged with allocating use of the entire radio spectrum (not just radio broadcasting), and developing technical standards through the International Telecommunications Union—Radio communications (ITU-R). The recommendations and technical standards for spectrum use are formally adopted as treaties among the member countries. *See* RADIO SPECTRUM ALLOCATIONS.

AM Medium-Frequency Band

This section discusses the more important technical aspects of AM broadcasting as it is currently used.

AM transmission standards. Broadcast stations in the medium-frequency band use amplitude modulation of a carrier wave to transmit information. The amplitude of the wave is modified in response to the changing amplitude of an audible voice or music signal. The AM receiver detects these amplitude changes and converts them back into audible signals, which can then be amplified and reproduced on acoustical transducers or speakers. *See* RADIO RECEIVER; RADIO TRANSMITTER.

The audible frequency range is generally considered to extend from 20 to 20,000 Hz. As a practical matter, AM broadcasting transmissions are limited to
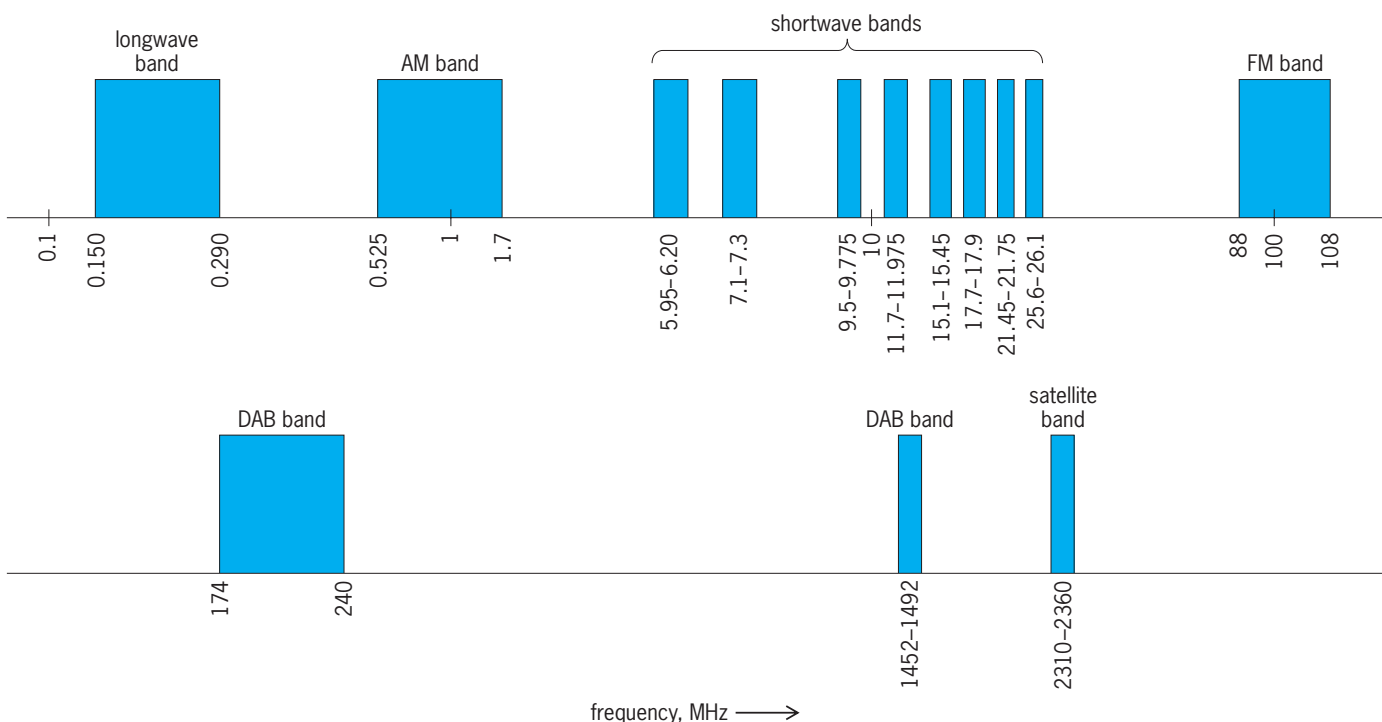

Fig. 1. Seven frequency bands used in radio broadcasting. Bands for digital audio broadcasting and satellite broadcasting are at higher frequencies.

a range of 50 to 10,000 Hz. Because of transmission components and directional antennas, the fidelity of many stations is more severely restricted, resulting in voice transmission that is still acceptable but music transmission that is of relatively low fidelity.

In North and South America and the Caribbean (ITU Region 2), AM stations are assigned to frequencies spaced 10 kHz apart. In most of the rest of the world, stations are spaced at 9-kHz intervals.

In the United States, the maximum permitted power is 50 kW. The vast majority of AM stations in the United States operate with 5 kW or less. Stations are often licensed with higher power levels for daytime than nighttime hours to allow for the increased potential for interference at night. In other parts of the world, AM transmitters operate at much greater power levels, with 1-MW power levels rather common in Europe, the Mediterranean, and the Middle East.

AM station allocation criteria. The fundamental concept in allocating and licensing broadcast stations is to limit interference from one station to the service area of another. For AM stations, the service area is defined as those locations receiving a signal level greater than 0.5 millivolt per meter. Because of atmospheric noise and unregulated artificial noise sources such as appliances, automobiles, and electrical systems, as a practical matter the 0.5-mV/m signal level may provide useful service only in electrically quiet rural areas. In urban areas, a signal level of 2 mV/m or higher may be necessary to overcome noise and provide useful service. *See* ELECTRICAL INTERFERENCE; ELECTRICAL NOISE.

To protect this service area, other stations on the same frequency (cochannel stations) are required to locate far enough away, or limit their transmitting power, so that their interfering signal is less than 5% of the desired station's signal strength at the boundary of the desired station's 0.5-mV/m service area, a 20:1 ratio of desired signal to undesired signal. For stations on adjacent channels (±10 kHz difference in carrier frequency), the required ratio is 1:1. This latter ratio is used primarily in the Western Hemisphere as an allocation standard; different, more restrictive, adjacent-channel ratios are used in other countries. Additionally, less restrictive desired-to-undesired signal ratios, or prohibitions on overlapping field-strength contours, are also employed when the frequency separation between stations is ±20 and ±30 kHz.

During nighttime hours a cochannel interference protection ratio of 20:1 is also employed. In addition, during nighttime hours the 0.5-mV/m, 50%-time sky-wave signal service area of stations in the highest class (class A) may also be protected from interference.

Transmitting antennas. An AM station may use a single tower for an antenna, resulting in an omnidirectional radiation pattern; or two or more transmitting towers to augment the radiation in certain directions while suppressing it in others, in order to comply with station allocation criteria. Such a directional antenna is developed by making use of the geometric relationship of the towers, and the relative amplitudes and phases of the currents in each tower, to create controlled constructive and destructive phase interference relationships in the desired directions. Directional antennas that use three to six towers are common, and stations with up to 12 towers have been built.

Since the allocation restrictions may be different during the daytime and nighttime hours, many stations employ two different directional antennas. **Figure 2** shows the daytime and nighttime radiation patterns for a three-tower directional antenna used by an AM station. The radiation from the antenna is expressed in millivolts per meter at 1 km (0.62 mi) from the antenna. *See* ANTENNA (ELECTRO-MAGNETISM).

Medium-frequency signal propagation. AM broadcast signals propagate from the transmitter by three mechanisms: ground-wave, space-wave, and sky-wave.

Ground waves travel along the ground surface (the boundary between the Earth and the atmosphere). Because they are surface waves, they penetrate into the ground, resulting in the energy being diminished because of losses in the ground. The degree of energy lost or signal attenuation is a function of the conductivity and permittivity of the near-surface ground, the frequency of operation, and the presence of any major surface discontinuities such as mountain ranges. The conductivity along the transmission path is the quantity primarily used to determine the extent of ground-wave propagation. The conductivity along salt-water paths can be as high as 5000 millisiemens per meter, while rocky terrain can have conductivities as low as 0.1 mS/m. Fertile farmland, alluvial plains, and other flat open areas can have conductivities ranging from 4 to 30 mS/m. Depending on the frequency, power level, and conductivity, useful ground-wave signals can propagate from a few tens of miles out to several hundred miles from the station.

The current flowing in the antenna also produces space waves, which travel through the atmosphere from transmitter to receiver. Space-wave propagation is usually limited by intervening terrain obstacles or the curvature of the Earth. Consequently, space-wave propagation is not as important for AM broadcasting as ground-wave and sky-wave propagation.

Sky-wave propagation occurs when space waves directed toward the ionosphere are reflected toward the Earth. This phenomenon can result in substantial signal strengths at distances of several hundred miles from the antenna. AM sky-wave propagation occurs primarily during nighttime hours by reflections from the E and F layers of the ionosphere at about 60 and 130 mi (100 and 220 km) altitude above the Earth's surface, respectively. During daytime hours the Sun's radiation ionizes the D layer of the ionosphere at about 40 mi (60 km) altitude above the Earth's surface. When the D layer is ionized, it tends to absorb and scatter the skyward-directed space waves from the AM antenna, thus preventing them from reaching the reflective E and F layers. As nighttime
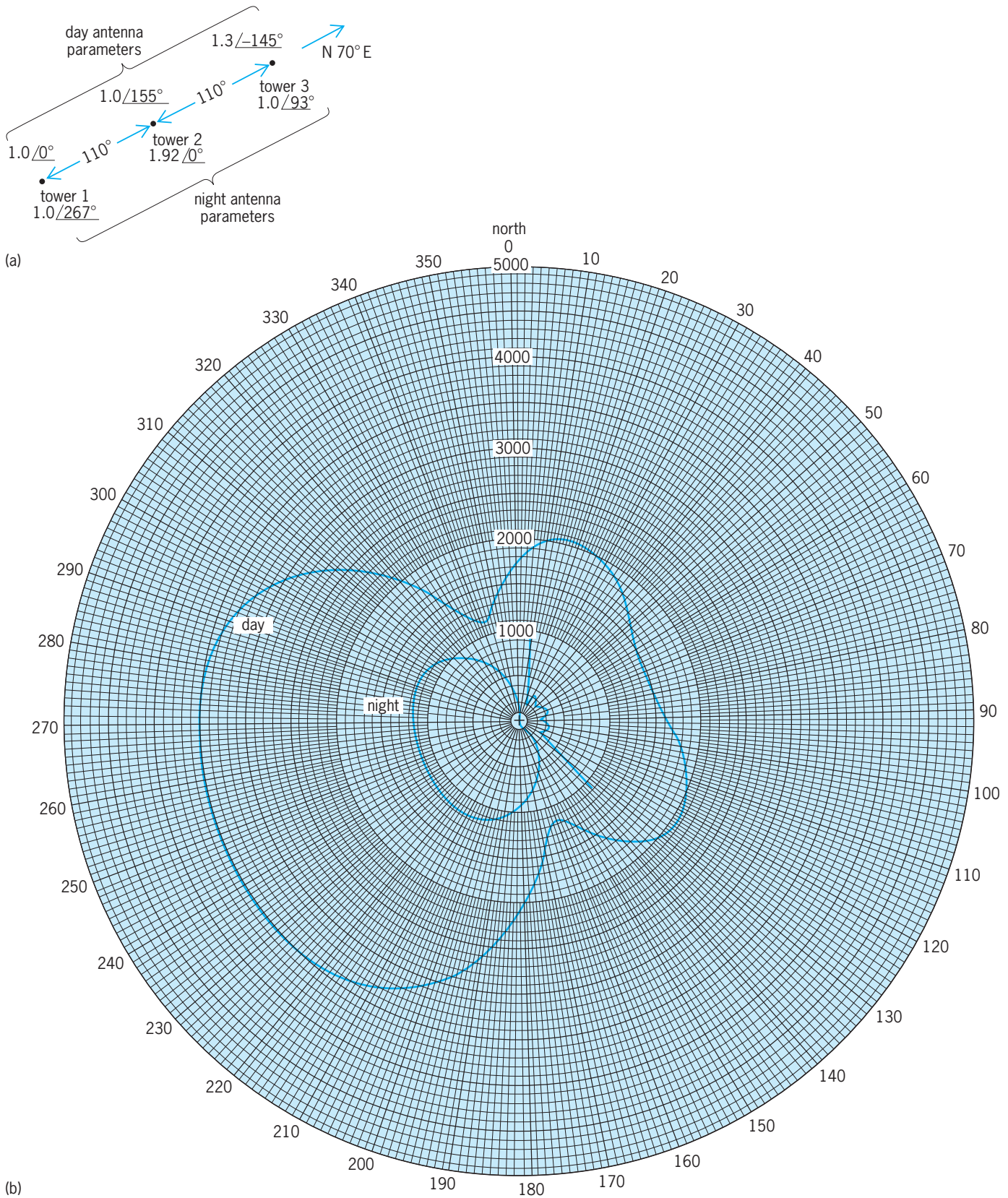
Fig. 2. Directional antenna array of the AM station KLOK in San Jose, California, broadcasting at 1170 kHz. (*a*) Configuration of array. Spacing between towers is given in electrical degrees. For the frequency of 1170 kHz, 1 electrical degree = 2.3 ft = 0.71 m and 110° = 257 ft = 78 m. The antenna parameters shown are field ratios and relative phases for each tower. (*b*) Resulting daytime (power = 50 kW) and nighttime (power = 5 kW) radiation patterns. The radiation from the antenna is expressed in millivolts per meter at 1 km (0.62 mi) from the antenna.

approaches, the ionization of the D layer diminishes, and so the AM signals pass through it to the E and F layers. The reflected sky waves increase in amplitude to levels at which they can provide useful service or substantial interference to other stations. *See* IONOSPHERE; RADIO-WAVE PROPAGATION.

Broadcast service area. The daytime ground-wave signal level protected by allocation criteria is usually 0.5 mV/m, although much higher signal strengths may be necessary to overcome noise from atmospheric and artificial sources, especially in highly urbanized areas. The extent of ground-wave service may be determined by using theoretical calculations or by measurements of the field strengths. The accuracy of the theoretical prediction method is limited primarily by the limited accuracy of the available conductivity data taken from maps, and secondarily by the approximate nature of the simple mathematical formulas used to calculate radiation from the antenna. Field-strength measurements provide an inefficient but definitive method of assessing station coverage.

During nighttime hours, the service area for most AM stations (other than class A stations) is usually limited by interference from other cochannel stations. In such cases, the service area is confined to the locations where the desired ground-wave field strength is 20 times greater than the RSS (root of the sum of the squares) of all the strongest interference signals on the channel at a given location. Because of interference during the night, the useful service may be limited to areas inside the 2.5 to 25 mV/m ground-wave contours. For low-power stations, this may be only 10 mi (16 km) or less from the transmitter.

For the highest class of stations (class A), nighttime service is protected by the allocation criteria to the 0.5-mV/m, 50%-time sky-wave contour. Because such service is subject to the time variations and fading of sky-wave propagation, the 0.5-mV/m, 50%-time contour is considered to be a secondary service area.

Figure 3 shows a plot of the theoretically predicted daytime and nighttime service areas for the station whose radiation patterns were shown in Fig. 2. The daytime service area is taken to be the 0.5-mV/m ground-wave contour, while the nighttime service area is limited by cochannel interference as described above to 8.37 mV/m. The distances to the contours vary because of radiation variations in the directional antenna and because of the different ground conductivity values in the region around the station, including that of salt water in San Francisco Bay.

AM stereo. In the United States, the Federal Communications Commission (FCC) has declined to designate a standard AM stereo transmission system. By 1990, approximately 600 stations, about 12% of the AM stations in the United States, operated in stereo. Of these, approximately 500 stations used a phase-modulation-type system, while another 100 used an independent-sideband system with the left and right channels transmitted on the lower and upper sidebands to form a complete composite AM stereo signal. *See* PHASE MODULATION; STEREOPHONIC RADIO TRANSMISSION.
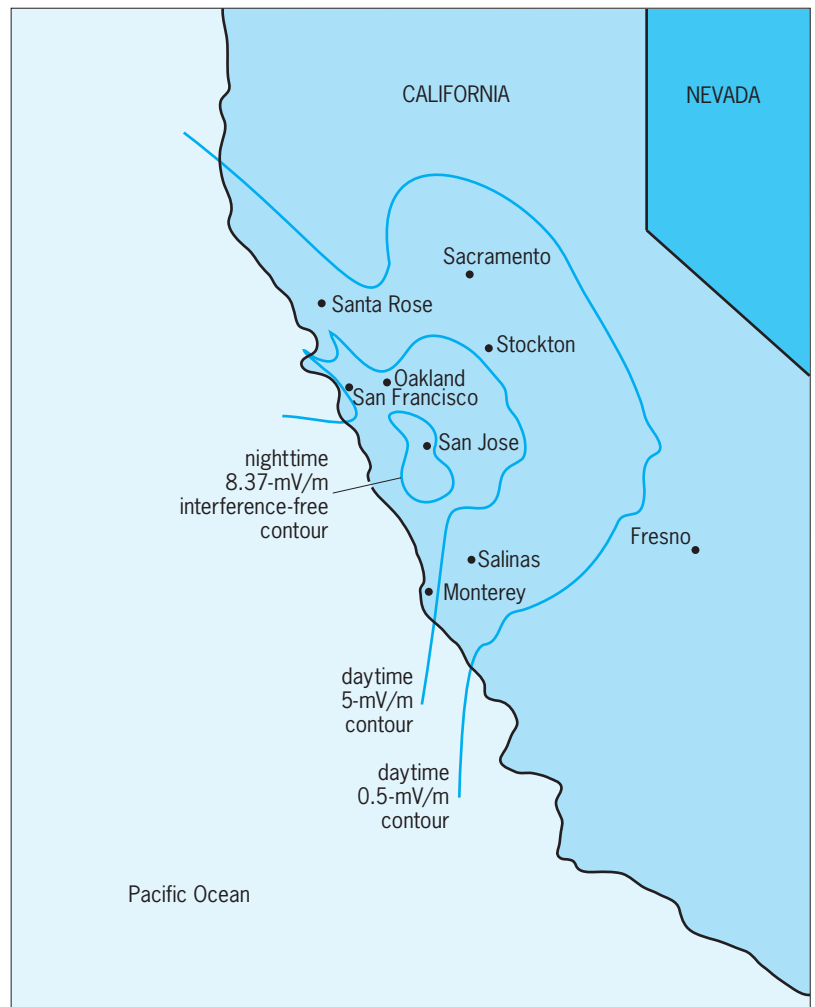

Fig. 3. Predicted daytime and nighttime interference-free field strength coverage contours of the AM station KLOK in San Jose, California.

FM VHF Band

FM broadcasting has become the dominant broadcast service in the United States primarily because of its better fidelity and its superior reception, which is less subject to noise and interference than that of AM.

FM transmission standards. Information is conveyed by frequency modulation or deviation of a carrier wave. In the United States, the carrier frequency may be deviated ±75 kHz around the assigned carrier frequency. The carrier frequencies, or channels, are spaced at 200-kHz intervals in the United States; a few other countries use slightly different channel spacings.

Nearly all FM stations transmit in stereo. A stereo audio signal consists of left and right channels. For broadcasting, the stereo signal is constructed at audio or baseband frequencies before being sent to the FM transmitter. The baseband frequency portion of the spectrum from 0 to 15 kHz (**Fig. 4**) is used for a signal that is the sum of the left and right channels $(L + R)$. The frequency band from 23 to 53 kHz is used for an amplitude-modulated signal that is the difference between the left and right channels $(L - R)$. A pilot tone at 19 kHz is also transmitted so that the receiver can recognize when a stereo signal is pres-
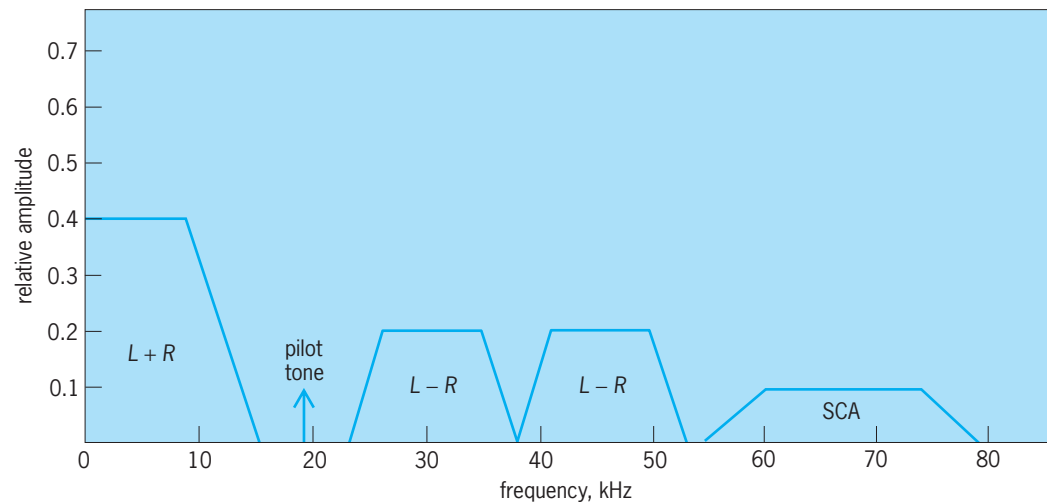
Fig. 4. FM broadcast baseband frequency spectrum.

ent. The pilot tone is also used to synchronize the detection of the $L - R$ signal. In the receiver, the $(L + R)$ signal is added to and subtracted from the $(L - R)$ signal to yield separate $2L$ and $2R$ audio signals at the receiver output. In this way the two separate left and right channels are reconstructed from the composite signal. This stereo transmission method was chosen so that it would be compatible with monaural FM broadcasts, which use only the 0–15-kHz $(L + R)$ signal.

The power of FM stations is limited by regulatory agencies so that the degree of coverage and potential interference to other stations is limited. A wide range of station classes has evolved using power levels from 3 to 100 kW. For FM, the allowed power rating is the effective radiated power (ERP) from the antenna, not the transmitter power. The effective radiated power is approximately equal to the transmitter power multiplied by the antenna gain minus any system losses.

FM station allocation criteria. FM stations are assigned to channels in the United States by using a fixed-distance separation table between transmitter sites. The amount of required separation depends on the station class, which in turn depends on the allowed effective radiating power and antenna height above average terrain (HAAT). For the highest class of station (class C) with 100 kW effective radiating power and 1969 ft (600 m) antenna height above average terrain, the required separation for stations on the same channel is 180 mi (290 km). Required separations are less for stations of lower class and for stations that are separated in frequency by ±200, ±400, or ±600 kHz.

Rather than use a fixed-distance separation scheme, some other countries use desired-to-undesired signal ratios to determine station locations to avoid interference. In still other countries, FM frequencies are freely assigned, with no plan for protecting service areas or limiting interference.

VHF signal propagation. The service area of an FM station depends on the propagation of space waves from the transmitter to the receiver. Ground waves

and sky waves, which are the dominant propagation mechanisms for AM broadcasting, are unimportant in the VHF band. Space waves propagate through the atmosphere and are diffracted around, and reflected off, mountains, buildings, and other objects. Propagation within areas that have an unobstructed line-of-sight from transmitter to receiver is most reliable and predictable. When obstructions lie along the path, the FM signal strength is attenuated below the level that would exist for a line-of-sight path of the same length. The amount of attenuation can be calculated by using simple physical models based on diffraction theory. The accuracy of the calculation is reduced by the inability of simple models to account for real hills and mountains. *See* DIFFRACTION.

Because of reflections that occur with VHF FM transmission, a condition known as multipath can exist. Multipath occurs when signals from more than one direct or reflected signal source arrive at the receiver. The combination of these signals can be such that they destructively interfere with each other, resulting in diminished signal strength and distorted program reproduction.

The degree of signal diffraction and reflection that occurs depends on the reception location, and can vary with time because of changing atmospheric or other propagation path conditions. For this reason, VHF FM broadcast signal strengths are usually described statistically. For example, the field strength level exceeded 50% of the time at 50% of the locations is denoted F(50,50). A higher-reliability signal level is one exceeded 90% of the time at 90% of the locations, denoted F(90,90).

FM VHF service area. The primary service area of FM stations is generally taken as that area where the F(50,50) signal strength exceeds 1.0 mV/m. In the United States, the FCC requires each station to have a city of primary service, where the F(50,50) signal level must exceed 3.16 mV/m.

For FCC purposes, the service area of an FM station can be found by using a set of curves showing field strength as a function of distance, effective radiated power, and antenna height above average

Fig. 5. Predicted 3.16-mV/m coverage contour for the 100-kW station KUGN in Eugene, Oregon. The broken line represents coverage based on FCC field strength curves. The solid lines and "islands" represent a more realistic depiction of coverage that explicitly takes into account terrain obstructions.

terrain. For the highest class of station (class C), the 1.0 mV/m service extends to about 60 mi (96 km) according to FCC curves. In flat terrain, with a high antenna, the effective service may be rendered with field strengths much less than 1.0 mV/m, resulting in service areas that extend to 130 mi (210 km) or more from the transmitter.

Because of mountains and other obstructions, FM signal strength does not decrease monotonically with distance, but may fluctuate with distance (as well as with time). Consequently, the FCC propagation curves provide only an approximation of FM service. By explicitly taking into account signal attenuation over terrain obstacles (mountains), a more realistic estimation of FM service may be obtained. **Figure 5** shows the 3.16-mV/m service area

for a 100-kW FM station. The broken line is the predicted FCC service area, while the solid lines show the service area prediction with terrain obstructions considered. The "islands" within the broken line are locations where the signal strength falls below 3.16 mV/m because of terrain obstacles.

Subcarrier transmission. To achieve greater utility from their transmitting facilities, FM broadcast stations have taken advantage of baseband spectrum space above 53 kHz for so-called subsidiary communication authorization (SCA) transmissions. For stereo stations, any frequency between 53 and 99 kHz may be used for an SCA, although the frequencies of 67 and 75 kHz are typically used. An SCA centered at 67 kHz is shown in Fig. 4. The amount of power that can be devoted to the SCA is limited

to 10% of the total station power, but within these loose technical constraints an FM broadcast station may transmit a wide variety of audio and digital signals on the SCA. Because the SCA signal is not detected by a normal FM stereo receiver, but requires a special receiver that is distributed to a limited audience (who usually pay for the service), SCA transmission services are sometimes referred to as narrowcasting rather than broadcasting. SCAs carry a wide range of services including background music, stock market data, and sophisticated nationwide paging services with watchband receivers that can automatically scan the FM frequency band and lock in on the FM broadcast station transmitting the appropriate digital coded message on the SCA. *See* RADIO PAGING SYSTEMS.

Shortwave Broadcasting

For reaching audiences in foreign countries or other distant places, shortwave broadcasting is most often used. Nearly 600 million shortwave radio receivers are in use worldwide.

Technical standards. Shortwave broadcasting is permitted worldwide in eight frequency bands from 5950 to 26,100 kHz. The assigned transmitting frequencies are spaced at 5-kHz intervals, resulting in a limited usable audio bandwidth. Voice transmissions are most effective, while music transmissions have limited fidelity.

Double-sideband (DSB) amplitude modulation is usually employed, although to make better use of the spectrum space some single-sideband (SSB) transmission is used. In single-sideband transmission, one sideband of a normal double-sideband signal is suppressed, thus reducing the occupied frequency bandwidth by about one-half without sacrificing audio bandwidth. The penalty with single-sideband transmission is that a more complex and stable receiver is required for suitable reception.

Transmitter powers for shortwave stations range from 1 to 500 kW, with the higher powers commonly used. Transmitters are normally coupled to high-gain directional transmitting antennas, which can result in effective radiated powers of several megawatts.

Frequency assignments. Four times each year, a shortwave station desiring to transmit to particular areas conducts a theoretical computer study of the propagation conditions from its transmitter site to the reception area for a coming season. The result of the study is a list of proposed frequency assignments and times when those frequencies are put to best use. This list is then submitted through the national government to the ITU. The ITU analyzes the frequency and time requests from all countries and then publishes a list of so-called incompatibilities or potential interference problems 3 months before the season begins. During this 3-month period, stations with conflicts have the opportunity to resolve them in bilateral agreements. In most cases, the interference conflicts can be resolved so that 85–90% of the transmissions will not suffer interference when operation for the season actually begins (based on the theoretical predictions).

To take best advantage of the varying propagation conditions throughout the day, a shortwave station may shift frequencies every 4 h to different bands, or redirect particular frequencies to different reception areas. The operating frequencies and schedules of a station may become quite complex.

Shortwave signal propagation. Shortwave signals propagate via sky waves that are reflected one or more times from the E and F layers of the ionosphere. Multiple reflections are possible because a signal can also bounce off the Earth's surface after reflecting off the ionosphere in a "Ping-Pong" effect. With multiple reflections, signals can propagate around the Earth, reaching countries distant from the broadcasting station. Because the degree of ionization of the E and F layers varies with the Sun's radiation, shortwave propagation is highly dependent on the solar illumination along the transmission path between the transmitter and the receiver.

Shortwave service areas. Several factors limit the reception of a shortwave signal, chief among them being atmospheric noise, interference, and signal fading. Through the use of high-gain antennas, shortwave stations can target a desired reception area by concentrating the energy in that direction. Usually a signal strength of 0.25 to 1.0 mV/m, 50% of the time, is needed in the targeted service area to overcome atmospheric noise and provide good reception. It may be impossible to overcome a strong cochannel interference source, and sometimes such interference has been used to deliberately jam reception of another signal. Signal fading is a result of the varying conditions of the ionosphere; however, usable signal strengths can sometimes exist for several hours even with changing atmospheric conditions.

Digital Audio Broadcasting

To enhance audio quality, a digital audio broadcasting service has been put in place in Europe and elsewhere, operating primarily in the 1452–1492-MHz band. As in VHF FM broadcasting, DAB uses transmitters located at elevated locations (mountaintops, building roofs) that provide the best line-of-sight paths to the intended service area. Digital broadcasting differs from VHF FM broadcasting in that the audio signal (voice or music) is first converted to a stream of binary digits (data bits) that represent the audio signal. These data bits are then used to modulate the radio-frequency carrier signal using one of several techniques. After transmission via the radio waves, the radio-frequency carrier is demodulated at the receiver to recover the stream of data bits, and the bits are then converted back to the audio signal. DAB offers improved reception quality and fidelity because error-correcting codes in the digital signal can be used to eliminate many flaws that may occur during transmission. It is not possible to use such codes to correct similar flaws in AM and FM analog transmissions. *See* INFORMATION THEORY; MODULATION.

The DAB system in use in Europe and elsewhere is known as Eureka 147, and employs a digital modulation technique known as orthogonal

frequency-division multiplexing (OFDM). This technique provides improved resistance to transmission problems resulting from fading and multipath that limit the coverage and quality of traditional VHF FM broadcasting. *See* MULTIPLEXING AND MULTIPLE ACCESS.

Due to the shortage of radio spectrum in the United States, a different DAB system that shares the same radio channel as medium-wave AM or VHF FM is being developed. Using sophisticated digital techniques, the aim of these systems is to allow simultaneous transmission of traditional analog AM and FM signals in the same channel as the digital broadcasting service.

Satellite Broadcasting

Satellite broadcasting also uses digital modulation techniques. However, in this case the transmitters are located on satellites high above the Earth. From this position, satellite broadcasting can achieve essentially universal coverage of an entire nation or most of a continent from one or two transmitters. Both geostationary and low-Earth-orbit (LEO) satellites are used for these systems. Specialized receiving antennas placed in locations visible to the sky are usually needed, along with a special radio receiver specifically designed for satellite service. To extend coverage into areas that are not visible to the sky, some satellite broadcasting networks employ a network of ground-based transmitters (repeaters) to supplement the coverage of the satellite signal. Satellite broadcasting is capable of providing 100 or more channels of audio programming, usually on a subscription or fee basis. *See* DIRECT BROADCASTING SATELLITE SYSTEMS.   Harry R. Anderson

Bibliography. D. G. Bobbett (ed.), *World Radio-TV Handbook 2001*, Watson-Guptill Publications, 2001; Federal Communications Commission, *Rules and Regulations, Part 73*, 1999; Institute of Electrical and Electronic Engineers, *IEEE Transactions on Broadcasting* (journal), quarterly; International Telecommunications Union, *Recommendations and Reports of the ITU-R*, vol. 10: *Broadcasting Service* (*Sound*), 1998; National Bureau of Standards, Transmission Loss Prediction for Tropospheric Communication Circuits, *NBS Tech. Note*, no. 101, 1967.

Radio broadcasting network

A group of radio broadcast stations interconnected by leased channels on wire, fiber, microwave, or satellite to one or more central feed points for the purpose of receiving and rebroadcasting program material of a timely nature to a geographically diverse audience; or alternatively, an organization which provides programming for, but does not actually own or operate any, broadcast stations. The terms "radio broadcasting network" and "radio network" are often used interchangeably.

The need for radio networks arises from the local nature of terrestrial radio broadcasting, which is due both to the physical nature of how radio waves propagate and to the way in which terrestrial radio services are regulated. At best, each AM and FM radio station will be receivable only on the order of 100 mi (160 km) from the transmitter site, with the exception of high-powered AM stations at night (so-called clear channel stations) which, because of a phenomenon known as skywave propagation, can be received over much greater distances (albeit with substantial interference). This distance is often further reduced by interference from what otherwise might be possible due to the manner in which broadcasting licenses are allocated. *See* RADIO BROADCASTING; RADIO-WAVE PROPAGATION.

In the United States, nongovernmental use of radio-frequency (RF) transmissions are regulated by the Federal Communications Commission (FCC), including the allocation of "channels" in the AM and FM broadcast bands which utilize RF frequencies 550–1705 kHz and 88–108 MHz, respectively. These channels, which listeners typically refer to by their broadcast frequency or call sign (a four-letter combination beginning with either W or K), are allocated by the FCC on a community-by-community basis. Before channels are allocated to a particular community, engineering analyses are performed to determine the impact on other nearby radio signals already allocated on either the same frequency or on the frequencies adjacent to the desired allocation channel. Based on these studies, the FCC determines the allowable distance a station can transmit without causing interference to its neighbors, which dictates the power level at which the broadcast station can operate. *See* RADIO SPECTRUM ALLOCATIONS.

The use of radio networks makes it possible to overcome these distance limitations and broadcast live and prerecorded programs to many communities, or even a national or global audience, through affiliated terrestrial radio stations (affiliates); the stations make national and regional markets available to advertisers and offer stations entertainment and public service programs.

Types of radio networks. Radio networks exist in many forms and are distinguished by whether they are government-owned (or sponsored) or are operated as a commercial enterprise (being either publicly or privately owned), whether they are free or subscription services, and the transmission medium used for reception. Many countries around the world have broadcasting systems which consist solely of government-run or -affiliated networks, although over time the trend is that many of these countries are developing commercial broadcast segments as well.

Commercial networks. Radio networks in the United States began as commercial enterprises and today remain a model for non-government-centric broadcasting systems which other countries (having a history of government-only networks) emulate. Commercial radio networks obtain revenue primarily

through the sale of advertisements which are broadcast along with network-produced or -purchased programming (such as news, entertainment, and sports).

Public networks. Public networks are nonprofit in nature and can be owned by either the government or not-for-profit corporations. They are funded either exclusively by the government or through a combination of government funding and private donations. In the United States, National Public Radio (NPR), a not-for-profit corporation, is the principal public radio network with over 780 affiliate radio stations in the AM and FM bands. NPR's government funding is provided principally by the Corporation for Public Broadcasting (CPB).

International networks. Many countries operate radio networks to deliver programming outside of their own, often to provide free and uncensored information to citizens of countries where government censorship is practiced. An example is the Voice of America (VOA), which operates a global network of transmitting stations that transmit over 1000 hours of news, informational, educational, and cultural programs every week to an audience of more than 100 million people worldwide in 44 languages. VOA broadcasts primarily in the AM and shortwave (approximately 6–18-MHz) bands.

Satellite Digital Audio Radio Service (SDARS). Radio services broadcast directly from a satellite to the listener are a new form of radio system made possible by advances in perceptual audio coding, mobile satellite antenna, and digital signal processing technologies. An SDARS system utilizes one or more satellites (in most cases complemented by terrestrial repeaters) to provide multiple channels of programming (on the order of 100) to listeners over a broad area, resulting in a coverage area commensurate with that of a radio network but with far fewer transmitters. So far, all SDARS systems are subscription services and have been built and operated by individual companies, which distinguishes them from terrestrial radio networks in that there are no affiliate stations. SDARS systems make use of RF signals which are at much higher frequencies than those of terrestrial radio stations, typically in the L-band (1452–1492 MHz) and S-band (2310–2360 MHz). The first SDARS system, WorldSpace, began broadcasting to Africa in 1999 and to Asia in 2000. The United States has two SDARS providers, XM Radio and Sirius, which began operation in November 2001 and July 2002, respectively. *See* SATELLITE RADIO.

Internet radio. The proliferation of the Internet and audio streaming technology has led to the development of so-called Internet radio stations, which have the ability to reach a global audience since their transmission medium is the Internet. As wireless Internet installations increase in number, utilizing wireless fidelity (Wi-Fi) technology, Internet radio will be available to mobile listeners and may over time offer competition to terrestrial and satellite radio services. Currently the principal limitations of Internet radio are that listening is primarily tied to the use of a personal computer, and the limited capacity of the computer servers from which each Internet radio station originates, since the amount of bandwidth required to broadcast the station increases incrementally with each listener. As multicast server technology is implemented throughout the Internet, this bandwidth problem will diminish. *See* INTERNET; WIRELESS FIDELITY TECHNOLOGY.

Early history. The development of radio broadcasting networks (originally called chain broadcasting) in the first part of the twentieth century is analogous to the development of the first printing press in that it revolutionized the manner in which information was disseminated. Radio networks made possible, for the first time, the simultaneous, instantaneous delivery of news and information on a continental (later global) scale. It is generally agreed that the first example of a radio network was demonstrated in 1923 involving two stations—WEAF(AM) in New York and WNAC(AM) in Boston—connected using a telephone line. Other methods of connection were experimented with, including telegraph lines and shortwave radio links (able to be received over greater distances than AM broadcast signals), but these proved to be unreliable and did not provide for audio of sufficient quality.

The first bona-fide radio network was the National Broadcasting Company (NBC), created in 1926 as a joint venture of the Radio Corporation of America (RCA), Westinghouse, and General Electric. This early network (and others too) owned the broadcast facilities and was responsible for the content broadcast over them. As networks proliferated, the government became concerned about the level of control being exercised by the networks on their affiliates, and in 1940 the FCC issued its "Report on Chain Broadcasting" which, in addition to reducing this level of control, placed limits on the amount of time the networks may broadcast through their affiliates.

Distribution technologies. Distribution of network programming to affiliates was accomplished using wired telephone circuits or wireless terrestrial microwave circuits until the 1980s, when the availability of affordable satellite uplink and downlink equipment and the ability for radio networks to lease satellite capacity made possible widespread use of geosynchronous satellite circuits as well. This distribution technology is similar to terrestrial microwave, using uplink and downlink frequencies in both the C-band (4–6 GHz) and Ku-band (12–14 GHz). Each satellite contains banks of transponders, each of which receives the uplink signal for amplification and retransmission by the downlink transmitter, which returns the signal to Earth station receivers. *See* COMMUNICATIONS SATELLITE.

These geosynchronous satellites, operated and maintained in their orbiting positions by the satellite service providers, repeat the signal on assigned transponder frequencies, beaming a broad pattern of signal which can cover an entire continent, or alternatively, beaming a narrower pattern, using a "spot" beam, for more localized coverage. The signal may be received in any location within the footprint of the

satellite beam by the affiliated radio station utilizing low-noise receiving amplifiers centered in the focus of microwave receiving antennas aimed toward the satellite.

NPR is generally acknowledged to have first successfully demonstrated stereo frequency-modulation (FM) distribution via satellite facilities, and today operates the Public Radio Satellite System (PRSS), a satellite-based distribution network which includes more than 400 downlinks, and is the principal means by which content is delivered to NPR affiliate stations. NPR continues to be a leader in radio network distribution technology. Its Content Depot system is based on Internet Protocol (IP) technology and will support distribution over a variety of media including satellite, terrestrial (point-to-point), and the Internet. *See* RADIO.   David H. Layer

Bibliography. *Broadcasting and Cable Yearbook*, annually; W. B. Emery, *National and International Systems of Broadcasting: Their History, Operation, and Control*, Michigan State University Press, East Lansing, 1969; Federal Communications Commission, *Report on Chain Broadcasting*, Commission Order no. 37, Docket no. 5060,1941; Federal Communications Commission, *Rules and Regulations*, Parts 73 and 74, and *Amendments*, 2004; W. L. Hetrich, *An Improved Audio Pipeline: NPR's Satellite System*, National Public Radio, 1982; H. Sterling and J. M. Kittross, *Stay Tuned—A History of American Broadcasting*, 3d ed., Lawrence Erlbaum Associates, 2002.

Radio compass

A popular term for an automatic radio direction finder used for navigation purposes on ships and aircraft. It is not strictly a compass, because it indicates direction with respect to the radio station to which it is tuned, rather than to the north magnetic pole. The modern radio compass uses a nondirectional antenna in combination with a bidirectional loop antenna to provide a unidirectional bearing indication. The navigator thus knows at all times whether the craft is traveling toward or away from the radio station used as a reference. The antennas are used with a special radio receiver that provides a visual indication of direction on a meter or cathode-ray indicator. *See* DIRECTION-FINDING EQUIPMENT.   John Markus

Radio-frequency amplifier

A tuned amplifier that amplifies the signals commonly used in radio communications. Amplifier designs in the radio-frequency (RF) range differ significantly from conventional low-frequency circuit approaches; they consequently require special, distributed circuit considerations. The need for a distributed treatment is warranted whenever the size of the physical circuit is greater than about one-tenth of the signal wavelength.

The wavelength of a radio-frequency signal in free space is obtained from Eq. (1), where c is the speed

$$\lambda = c/f \qquad (1)$$

of light ($3 \cdot 10^8$ m/s) and f is the frequency of the signal in hertz. The wavelength of a 3-gigahertz signal in free space, for example, is $\lambda = 3 \cdot 10^8/3 \cdot 10^9 = 0.1$ m, or 10 cm (4 in.). Thus, lumped circuit theory will not accurately model the behavior of 3-GHz circuits whose dimensions are in the 1-cm range, and Kirchhoff's voltage and current laws must be abandoned in favor of voltage and current wave propagation phenomena. An immediate implication is that appropriate distributed matching networks are needed to reduce standing-wave effects and avoid undesirable oscillations. The designs of suitable matching networks in conjunction with stability and noise figure analyses are usually the first steps in the amplifier design process. Additional considerations for a radio-frequency amplifier include gain, gain flatness, gain compression, output power, bandwidth, harmonic generation, and bias conditions.

Transition from lumped to distributed design. The starting point for a conventional amplifier circuit is the selection of an active device, either a bipolar junction transistor (BJT) or a field-effect transistor (FET). For optimal power transfer and amplifier stabilization, the device typically requires that input and output impedance matching networks be interposed between its input source and its load. In their simplest forms, each of these networks consists of two reactive elements that transform the impedances of the radio-frequency source and load, typically 50 ohms, to new values at the active device plane. The level of the input source must be matched to the input impedance of the transistor, which is typically very low. For the transistor amplifier to deliver maximum output power to its load, its output impedance—typically higher than 50 ohms—must be decreased to the 50-ohm level of its load. *See* ELECTRICAL IMPEDANCE; IMPEDANCE MATCHING; TRANSISTOR.

Figure 1 depicts an example of a lumped circuit model for a narrow-band amplifier where a capacitor C_1 followed by an inductor L_1 is used to create the input network, while an inductor L_2 followed by a capacitor C_2 is used for the output network. The inductors, RFC (for radio-frequency choke), and capacitors, C_B, isolate the radio-frequency signal paths from the dc bias sources. *See* CAPACITOR; INDUCTOR.
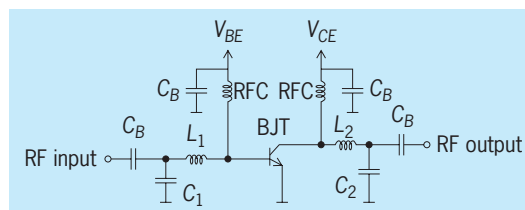

Fig. 1. Generic lumped element amplifier with input (L_1, C_1) and output (L_2, C_2) matching networks. Capacitors C_B and inductors RFC separate the radio-frequency paths from the required bipolar junction transistor (BJT) bias conditions.

Unfortunately, this design implementation suffers serious drawbacks at high operational frequencies. Specifically, the lumped elements in the radio-frequency paths no longer act as ideal inductors and capacitors; resistive losses and parasitics cause undesired departures from the ideal electric circuit behaviors at high frequency. Fortunately, transmission lines—pairs or conductors of carefully chosen dimensions—can be made to function as capacitors or inductors at frequencies where discrete inductors and capacitors perform poorly. Whether a length of transmission line acts as a capacitor or an inductor depends on its length at its operating frequency and on whether its far end is open or shorted. *See* TRANSMISSION LINES.

The signal propagation speed in transmission lines is usually considerably less than that in free space, so the electrical wavelength of a signal on a transmission line is noticeably shorter than its free-space wavelength. The electrical wavelength is given by Eq. (2), where c is the speed of light, f is the cir-

$$\lambda = c/(f\sqrt{\varepsilon_r}) \qquad (2)$$

cuit frequency of operation, and ε_r is the relative dielectric constant of the circuit-board material. For a relative dielectric constant of 4, the electrical wavelength of a signal on a circuit board will be one-half its wavelength in free space. Therefore the wavelength of the 3-GHz signal on the circuit board would be 5 cm (2 in.). Most lines used in circuit design are about 1/4 wavelength long, around 1.25 cm (0.5 in.) in the 3-GHz case.

Figure 2 depicts an amplifier circuit using transmission-line segments whose response is equivalent to the circuit of Fig. 1. Because of its parasitic properties, the circuit of Fig. 1 will not perform properly at high frequencies, while the one of Fig. 2 will. Open-circuited transmission lines of less than 1/4 wavelength at the frequency of operation of the amplifier appear as capacitors, while those greater than 1/4 wavelength act as inductors. The value of capacitive or inductive reactance of an open-circuited line depends on the length of the line and on its charac-
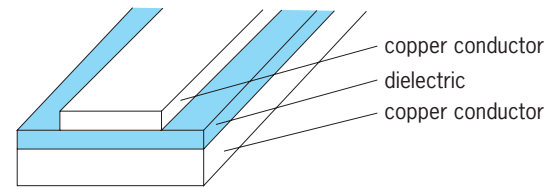

Fig. 3. Microstrip line.

teristic impedance. The special-case line segments denoted $\lambda/4$ are exactly 1/4 wavelength long. When a transmission line is exactly 1/4 wavelength long, its input terminals will look like a short circuit if its far end is open-circuited, and an open circuit if its far end is short-circuited. This quarter-wave property is used to emulate the action of a radio-frequency choke. In Fig. 2, for example, the far end of each quarter-wave line effectively is shorted by a capacitor that has a very low reactance at the desired operating frequency. At the transistor terminals, 1/4 wavelength away, this short appears as an open circuit to radio-frequency energy, and these connections do not disrupt the operation of the matching networks. At dc the lines are direct low-resistance connections between the power sources and the amplifier collector and base bias circuits. *See* CHOKE (ELECTRICITY).

Microstrip, a particularly useful form of transmission line for these applications, is implemented by laying down copper strips on the insulating surface of a circuit board whose opposite side is completely covered by a thin layer of copper (**Fig. 3**). The critical parameters of the line are its copper thickness and width, the thickness and relative permittivity of the insulating (dielectric) layer between the strip and the copper substrate, and, to a lesser extent, the length of the line.

Microstrip design approach. A microstrip line implementation of the input and output matching networks of Fig. 2 for a narrow-band amplifier at 3 GHz takes the following form: For a 0.76-mm-thick (30-mil) circuit board material with relative dielectric constant $\varepsilon_r = 4.4$, a trace width of 1.45 mm will achieve a characteristic impedance of $Z_0 = 50$ ohms. The effective dielectric constant, however, is only about 3.33 since the microstrip line resides at an air-dielectric interface. Therefore the electrical wavelength of a 3-GHz signal on this line is given by Eq. (3): about 55 mm. The two-element input match-

$$\lambda = c/(f\sqrt{\varepsilon_{eff}})$$
$$= 3 \times 10^8/(3 \times 10^9\sqrt{3.33}) = 0.0548 \text{ m} \quad (3)$$

ing network, L_1 and S_1, needs to transform the transistor input impedance to 50 ohms. A transmission line segment, L_1, of length 9.6 mm followed by a capacitive stub, S_1, of length 5.5 mm does the job.

Similarly, the output matching network, L_2 and S_2, transforms the conjugate complex impedance of the transistor to the external load impedance of 50 ohms. The network capacitive stub, S_2, is of length 9.1 mm and its transmission line segment, L_2, is 12.7 mm
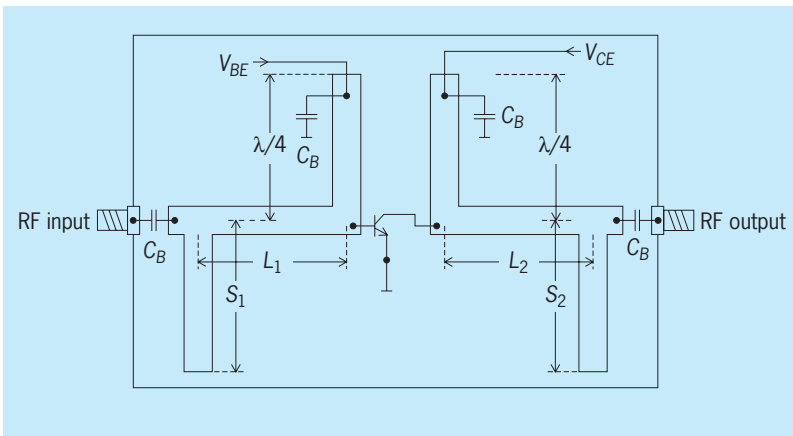

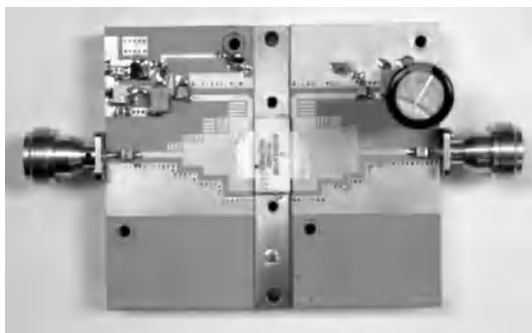Fig. 2. Basic layout of input and output matching networks with microstrip lines.

Fig. 4. Board layout for field-effect transistor (FET) power amplifier with microstrip networks for input and output matching. Total board dimensions are 76 mm (3 in.) in length by 40 mm (1.575 in.) in width. (*From J. E. Davis et al., The Characterization of a Power Amplifier, Senior Project Report, Worcester Polytechnic Institute, Worcester, MA, 2005*)

long. The amplifier gain with this matching arrangement is about 16.57 dB.

Finally, the $\lambda/4$ line segments, each 13.7 mm in length, transform radio-frequency ground (at 3 GHz, capacitor C_B is a short circuit) to very large impedances at the transistor terminals.

Power amplifier matching. A practical matching network for a power amplifier is shown in **Fig.** 4. A high-performance metal-oxide-semiconductor (MOS) field-effect transistor is employed in this case. The amplifier is designed for cellular base station applications in the frequency range 1.9–2.4 GHz. The MOS device is rated for 24 W average output power and 150 W peak power. It operates at a nominal drain-source voltage of 28 V and delivers up to 15 dB of gain.

Figure 4 depicts the device installed on a circuit board with a relative dielectric constant of 3.5, a height of 0.76 mm (30 mil), and a copper conducting surface of thickness 35 μm (1.4 mil). The conductor routing forms the biasing interconnections and the transmission line matching circuits. The bias voltages are supplied through 17.866-mm-long (0.7034-in.) $\lambda/4$ microstrip lines terminated by bypass capacitors on the dc side. The matching networks are constructed from microstrip segments of varying width and length ratios. The widest section of microstrip adjacent to the device has very low characteristic impedance; this is necessary to transform the very low device input impedance to a reasonable level. Several free-floating copper patches are supplied around the matching circuits. These can be connected selectively to the microstrip segments to adjust the matching. Although not visible in Fig. 4, the copper base plate is attached to a water cooling system. *See* AMPLIFIER.   Reinhold Ludwig

Bibliography. C. Coleman, *An Introduction to Radio Frequency Engineering*, Cambridge, 2004; R. Ludwig and P. Bretchko, *RF Circuit Design: Theory and Applications*, Prentice Hall, 2000; F. H. Raab et al., RF and microwave power amplifier and transmitter technologies—Part 1, *High Freq. Electr.*, 2(3):22–36, May 2003.

Radio-frequency impedance measurements

Measurements of electrical impedance at frequencies ranging from a few tens of kilohertz to about 1 gigahertz. This broad radio-frequency (rf) range encompasses frequencies from the extralow-frequency band, through the long-, medium-, and short-wave ranges, and then through the very high frequency (VHF) and the ultrahigh-frequency (UHF) bands. In the electrical context, impedance is defined as the ratio of voltage to current (or electrical field strength to magnetic field strength), and it is measured in units of ohms (Ω). *See* ELECTRICAL IMPEDANCE; RADIO SPECTRUM ALLOCATIONS.

At zero frequency, that is, when the current involved is a direct current, both voltage and current are expressible as real numbers. Their ratio, the resistance, is a scalar (real) number. However, at nonzero frequencies, the voltage is not necessarily in phase with the current, and both are represented by vectors, and therefore are conveniently described by using complex numbers. To distinguish between the scalar quantity of resistance at zero frequency and the vectorial quantity at nonzero frequencies, the word impedance is used for the complex ratio of voltage to current. The real part R of the complex impedance, $Z = R + jX$, is the (series-equivalent) resistance, and the imaginary part, the reactance X, is the reactive impedance of the (series-equivalent) inductance or capacitance (for positive or negative signs, respectively). Admittance Y is defined as the inverse of impedance, $Y = 1/Z$, and some instruments measure admittance directly, in units of siemens (S). Network theory states that at a given frequency the impedance of any two-terminal linear (that is, level independent) circuit may be represented either as a series or a parallel connection of a resistance and a reactance. *See* ALTERNATING-CURRENT CIRCUIT THEORY; DIRECT CURRENT; ELECTRICAL RESISTANCE; NETWORK THEORY.

The measurement of impedance at radio frequencies cannot always be performed directly by measuring an rf voltage and dividing it by the corresponding rf current, for the following reasons: (1) it may be difficult to measure rf voltages and currents without loading the circuit by the sensing probes; (2) the distributed parasitic reactances (stray capacitances to neighboring objects, and lead inductances) may be altered by the sensing probes; and (3) the spatial voltage and current distributions may prevent unambiguous measurements (in waveguides, for instance).

Impedance measurements may be classified according to the principle used, such as resonance methods, bridge methods, or reflectometers, or may be classified according to the physical nature of the conductors, such as two-terminal, three-terminal, coaxial line, stripline, or waveguide.
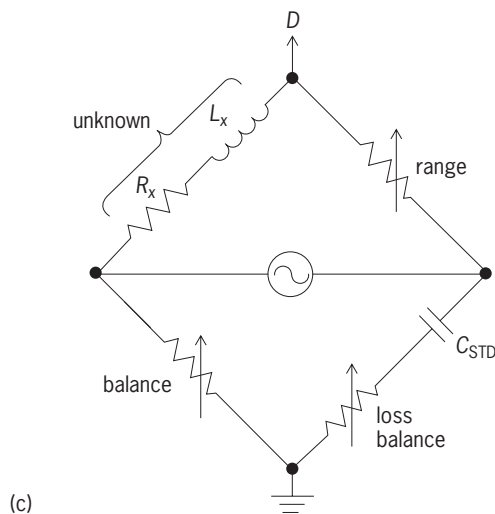
Bridge methods. At low frequencies, impedance measurements are often carried out by measuring separately the resistive and reactive parts, using either Q-meter instruments (for resonance methods), or reconfigurable bridges, which are sometimes
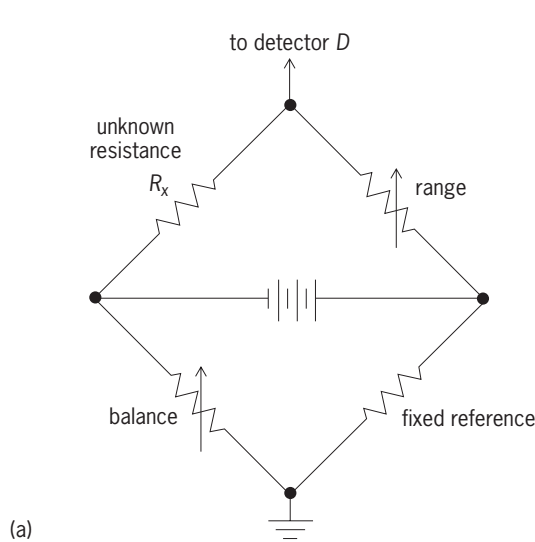
(a)
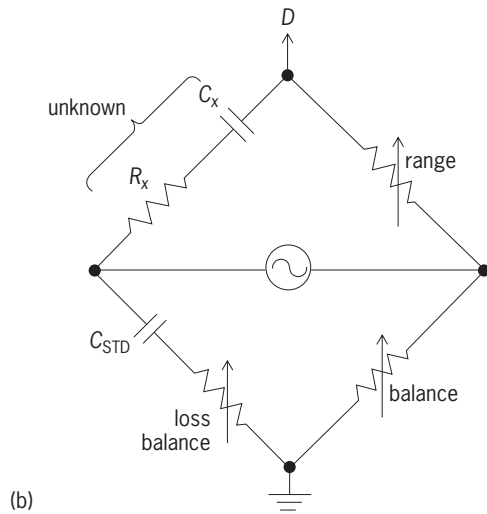
(b)

(c)

Fig. 1. Three internal bridge circuits used in an *LCR* bridge. (*a*) Wheatstone bridge. (*b*) Series-resistance-capacitance bridge for capacitance measurements; C_{STD} is standard capacitance. (*c*) Maxwell bridge for the measurement of inductances.
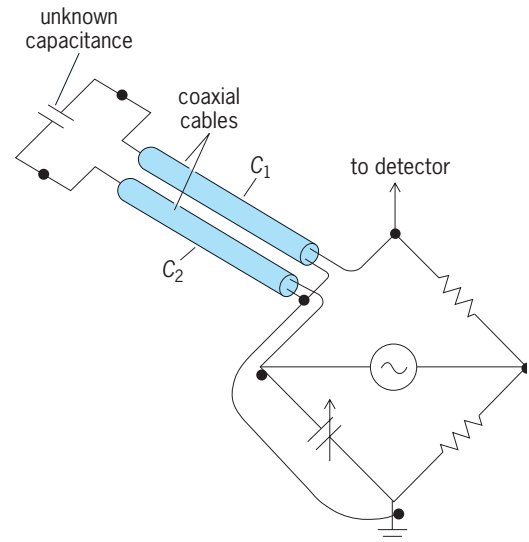

Fig. 2. Three-terminal measurement of impedance. This connection permits the remote measurement of impedance, since the measurement is between the inner conductors of the connecting coaxial cables, and lead capacitances are shielded from each other. Cable capacitance C_1 is across the detector. For accurate measurements, C_2 may be subtracted from the standard capacitance.

called universal *LCR* (inductance-capacitance-resistance) bridges. In one such bridge the resistive part of the impedance is measured at dc with a Wheatstone bridge (**Fig. 1*a***). Capacitive reactance is measured with a series-resistance-capacitance bridge (Fig. 1*b*), and inductive reactance is measured with a Maxwell bridge (Fig. 1*c*), using alternating-current (ac) excitation and a standard capacitance. If the impedances of open structures or remote components have to be measured, lead capacitances may significantly alter the impedance sought. In such cases, a quasi-three-terminal measurement may be performed by making connections to the measurand via shielded (coaxial) cables, in which case the bridge senses primarily the impedance between the inner conductors. The outer conductors of the cables are grounded (**Fig. 2**). The capacitance of one cable is across the detector, and therefore has no effect, and for accurate measurements the capacitance of the other cable may be measured and subtracted from the value of the standard capacitance. A modern, microprocessor-controlled *LCR* bridge operates in the frequency range of 100 Hz–40 MHz and has an accuracy of 0.17%. *See* WHEATSTONE BRIDGE.

Transformer bridge. Transformer bridges (**Fig. 3*a***) are capable of operating up to 100 MHz. The use of transformers (Fig. 3*b*) offers the following advantages: (1) Only two bridge arms are needed: the standard and the unknown arms. Phase opposition between them (necessary for a bridge null) is provided by one of the transformers, driven at a tap on the winding. (2) Both the detector and the source may be grounded at one of their terminals, minimizing ground-loop problems and leakage. The neutral point (Fig. 3) may be regarded as an rf virtual ground

point, and for three-terminal measurements is used for the connection of the coaxial cable outer conductors, because capacitances effectively shunt the transformers and have negligible effect on the balance. *See* TRANSFORMER.

Coaxial line admittance bridge. A coaxial line admittance bridge (**Fig. 4**) is usable from 20 MHz to 1.5 GHz. The currents flowing in three coaxial branch lines are driven from a common junction, and are sampled by three independently rotatable, electrostatically shielded loops, whose outputs are connected in parallel. Two of the branch lines are terminated by conductance and susceptance (inverse of reactance) standards, and the third by the unknown admittance. The currents in the conductance and susceptance arms are in quadrature, and by rotating the loops the correct proportions of these two currents are found to balance the current in the unknown arm. Scales beneath the arms which rotate the loops can be read directly in units of admittance. (The third rotatable loop in the unknown arm provides a common scaling factor to the real and imaginary parts of the admittance.) The attainable accuracy of this instrument is between 3 and 5%, depending on frequency.

Reflection coefficient. A quantity related to impedance is the complex (voltage) reflection coefficient, defined as the ratio of the reflected voltage to the incident voltage, when waves propagate along a uniform transmission line in both directions. Usually, uppercase gamma (Γ) or lowercase rho (ρ) is

$$\Gamma = \frac{Z_T - Z_0}{Z_T + Z_0} \qquad (1)$$

$$\text{VSWR} = \frac{1 + |\Gamma|}{1 - |\Gamma|} \qquad (2)$$
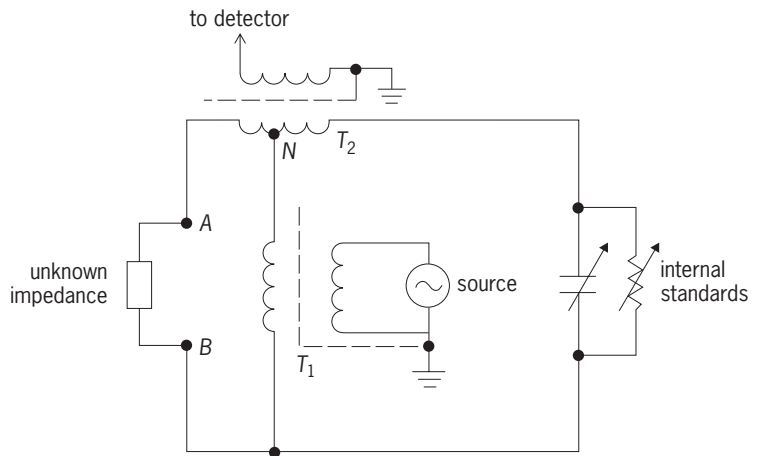
used to represent the reflection coefficient. When a

Fig. 3. Basic circuit diagram of transformer bridge usable up to 100 MHz, with transformers T_1 and T_2. Transformer T_2 provides phase opposition between standard and unknown arms. Point *N* (neutral) is for the connection of the ground terminal of three-terminal measurements. Capacitance between *A-N* and *B-N* effectively shunt the transformers.

transmission line of characteristic impedance Z_0 is terminated in impedance Z_T, the reflection coefficient at the load is given by Eq. (1), and the voltage standing-wave ratio (VSWR) is related to the magnitude of Γ by Eq. (2). *See* REFLECTION AND TRANSMISSION COEFFICIENTS; TRANSMISSION LINES.

Resistive bridge. Resistive bridges employed as reflectometers use a matched source and detector, and therefore differ from the Wheatstone bridge (Fig. 1*a*), which aims to use a zero-impedance voltage source and an infinite-impedance detector. It can be shown that these kinds of bridges, from a network-analysis point of view, are equivalent to a directional coupler having a 6-dB coupling coefficient and having a 6-dB attenuator between the source and its input port. When the bridge is driven by a matched generator and uses a matched rf detector (that is,
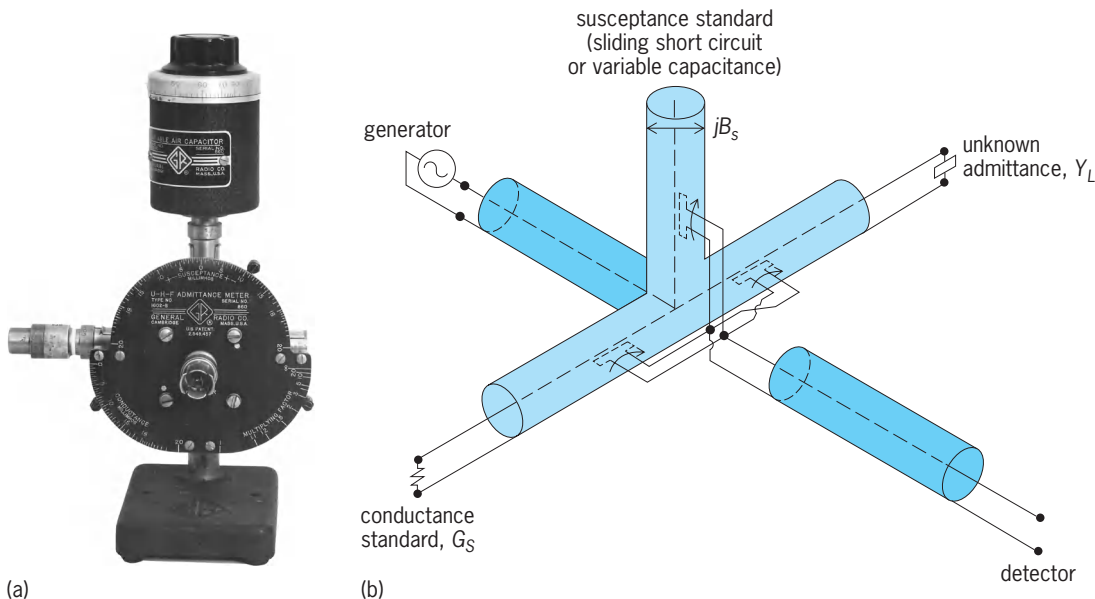

(a)   (b)

Fig. 4. Admittance bridge. (*a*) Instrument (*GenRad, Inc.*). (*b*) Circuit.
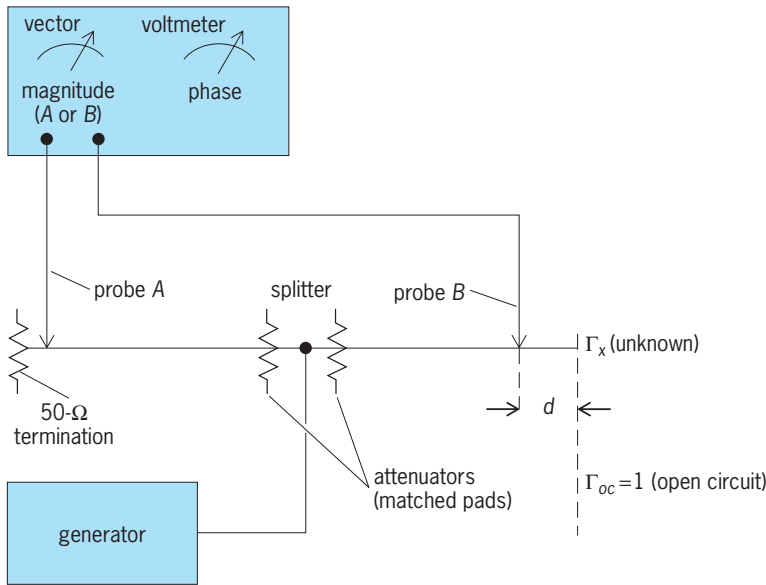
Fig. 5. Possible connection of a vector voltmeter for the measurement of impedance. (Hewlett-Packard)

with the impedances of the generator and detector equal to the characteristic impedance Z_0), the output is proportional to Γ. If this output is measured with a voltmeter and with a suitably referenced phase meter (or a vector voltmeter, described below), the complex reflection coefficient, and from it the complex impedance Z_T, of Eq. (1) is found. When it is sufficient to measure only the voltage standing-wave ratio (VSWR), only an amplitude detector is needed. If a square-law detector (power meter or diode detector at low levels) is used, the output is proportional to $|\Gamma|^2$, from which the VSWR may be calculated by using Eq. (2). Such a bridge is first calibrated by connecting a known termination, such as an open or a short circuit (with reflection coefficients $+1$ or -1, respectively). The advantages of the VSWR bridge are low cost and potentially wideband operation. Without adopting special techniques to reduce measurement uncertainty, $|\Gamma|$ may be measured with an uncertainty of ± 0.01. *See* PHASE-ANGLE MEASUREMENT; VOLTMETER.

Voltage-current method. Some specialized electronic instruments make use of the basic definition of impedance, and effectively measure voltage and current. One such instrument is called an rf vector impedance meter. Instead of measuring both the voltage and the current, it drives a constant current into the unknown impedance, and the resultant voltage is measured. With a constant driving current, the rf voltage across the unknown is proportional to the magnitude of its impedance, $|Z|$. The phase difference between the current and the voltage is measured with a discrete phase-measuring circuit. The unknown impedance is thus obtained in magnitude and phase. The instrument linearly downconverts the measuring frequency (0.5 to 108 MHz) to a 5-kHz intermediate frequency, where the amplitude and phase measurements are performed.

Vector voltmeter method. Vector voltmeters (VVM) are instruments with two (high-impedance) volt-

meter probes, which display the voltages at either probe (relative to ground) as well as the phase difference between them. One type operates from 1 MHz to 1 GHz, and linearly converts to a 20-kHz intermediate frequency by sampling. This instrument can be connected to measure impedance (**Fig. 5**). The signal is split and sent toward a matched termination (say, 50 Ω) and toward the unknown impedance. The reference probe A senses the voltage in the matched transmission line while, in the unknown arm, the other probe B measures the voltage at a position equidistant from the junction. Attenuators may be inserted in both arms to minimize interaction. The measurement is normalized by connecting a (shielded) open circuit in the plane of the unknown, and the complex voltage at probe B is noted as V_{oc}. The unknown is connected and the indication of probe B is noted as V_x. Then the unknown reflection coefficient is calculated from Eq. (3), where ϕ

$$\Gamma_x = e^{j2\phi}\left[\frac{V_x}{V_{oc}}(1 + e^{-j2\phi}) - 1\right] \qquad (3)$$

is the electrical angle between probe B and the unknown (equal to $2\pi d/\lambda$, where d is the distance between probe B and the unknown and λ is the wavelength). Because the substitution is on the same arm, the lack of symmetry of the splitter, the lack of equality of the attenuators, and the lack of equality of the two voltmeter channels A and B do not cause errors.

Since the probe responds to the sum of the incident and reflected waves, the resolution becomes poorer as the reflection coefficient magnitude approaches zero. This method is therefore not suitable for measuring impedances near the characteristic impedance of the transmission line (50 Ω in this case). For the accurate measurement of such impedances, directional couplers are used. *See* DIRECTIONAL COUPLER.

At radio frequencies the realization of directional couplers is quite different from microwave realizations because the component is much shorter than the wavelengths encountered. For this reason, lumped-element directional couplers may be constructed. Equal electric and magnetic coupling results in directional behavior so that the coupled output responds (ideally) only to signals traveling in one direction in the main line. The electric and magnetic couplings may be provided by voltage and current transformers, whose signals are summed. Distributed-circuit directional couplers may also be constructed by twisting a pair of insulated copper wires together to provide mainly capacitive coupling between them. Then the twisted pair is formed into a coil to provide additional inductive coupling. By stretching the coil, the magnetic coupling may be varied, and adjusted to equal the capacitive coupling. Such directional couplers have been constructed from 1 to 300 MHz and, if equipped with small trimmer capacitors near the terminals, can achieve directivities in excess of 50 dB at selected frequencies.

A coaxial-line, rf impedance-measuring circuit can be built using a vector voltmeter and directional couplers. One probe serves as a phase reference only. The other probe samples the reflected signal, and therefore the vector voltmeter indication is proportional to the reflection coefficient. The readings are normalized by first connecting a reflection standard, which may be a short or an open circuit, or a mismatched resistor. In measuring small reflection coefficients (impedances near 50 Ω), the sensitivity of the vector voltmeter may be increased in 10-dB steps, offering high resolution. If the analog magnitude and phase outputs are converted into their rectangular form, these voltages may be connected to the X and Y deflection inputs of an oscilloscope to give a display of the impedance on the circle diagram (Smith chart). By controlling the frequency and the calibration sequence with a computer, the VVM-coupler method may be automated, and this forms the basis of automatic network analyzers (ANA). *See* MICROWAVE IMPEDANCE MEASUREMENT.

Resonance method. When the magnitude of the reactive part of the impedance is much greater than the resistive part at a given frequency f, resonance methods may be employed to measure impedance. The most commonly used instrument for this purpose is the Q meter. The Q (quality-factor) of an impedance is defined as 2π times the ratio of energy stored to energy lost per cycle. Numerically, Q is equal to $\omega L/R$ or $1/R\omega C$ for the series combination of a resistor R and inductor L or a resistor and capacitor C respectively, where $\omega = 2\pi f$.

A Q-meter-type measurement requires a radio-frequency source of frequency f, which drives a current through a current-limiting injection resistor R_i into the Q-meter circuit (**Fig. 6**). The ratio of a magnitude of the voltage E developed across the calibrated capacitor C to a voltage e across the combination of capacitor, inductor, and resistor is indicated with a high-impedance rf voltmeter, and has a maximum value of Q_{ind} when C is set to maximize this reading.

To measure low-loss capacitors, the circuit is first resonated (with a suitably chosen inductor of high Q) to find Q_{ind}. The capacitor setting at resonance is noted. Then the unknown capacitor is paralleled with C, and the reduction in C needed to restore resonance is the value of C_x.

A low-loss inductor is measured by resonating it with C and substituting into Eq. (4). Since the magnitude of E/e is given by Eq. (5), the Q factor of the inductor is given by Eq. (6), which, for high-Q circuits, approximates to $Q = Q_{ind}$. The series resistance is given by Eq. (7), and the inductance is given by

$$L_x \simeq \frac{1}{\omega^2 C} \qquad (4)$$

$$\left|\frac{E}{e}\right| = [(\omega^2 L_x) + (\omega R_x C)^2]^{-1/2} \qquad (5)$$

$$Q = \left(Q_{ind}^2 - 1\right)^{1/2} \qquad (6)$$

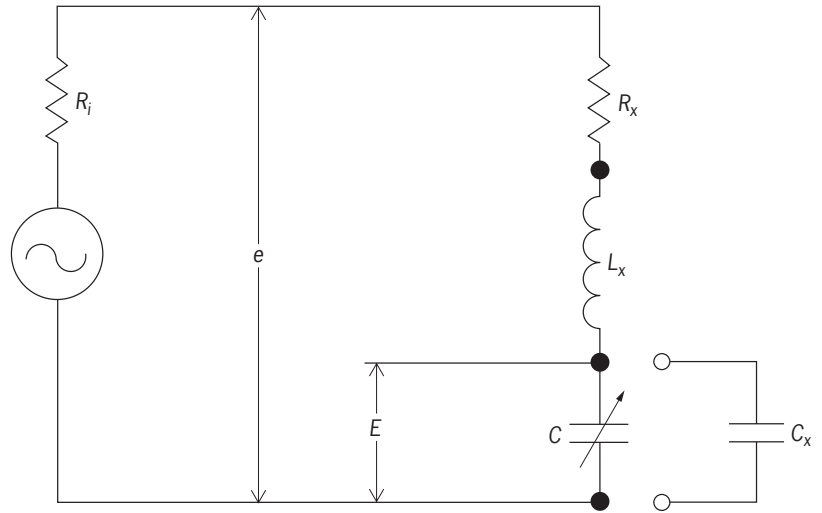cuits, approximates to $Q = Q_{ind}$. The series resistance is given by Eq. (7), and the inductance is given by

Eq. (8), which, for $Q_{ind} \gg 1$ reduces to Eq. (4).

$$R_x = \frac{\left(Q_{ind}^2 - 1\right)^{1/2}}{Q_{ind}^2 \omega C} \qquad (7)$$

$$L_x = \frac{Q_{ind}^2 - 1}{Q_{ind}^2 \omega^2 C} \qquad (8)$$

See Q METER.

Slotted lines and six-port methods. At the upper end of the rf range, microwave methods of impedance measurement may also be used, employing slotted lines and six-port junctions. For completeness, these methods will be briefly discussed.

A slotted line is a transmission line, with an axial slot cut in the outer conductor, permitting a traveling probe to sample the electric field. When an unknown impedance is connected to one end and a generator to the other, the ratio of the maximum to minimum signals as the probe is sliding is the voltage standing-wave ratio, and is related to the reflection coefficient by Eq. (2). The phase of Γ is found from Eq. (9), relating the positions of the minima when the

$$\phi_x = \pi \left[1 - \frac{4(d_{sc} - d_x)}{\lambda} \right] \qquad (9)$$

unknown is connected and when a short circuit is connected, where d_{sc} and d_x are the probe positions for the minima for the short circuit and the unknown connected, and λ is the wavelength.

According to the theory of six-port junctions, an unknown impedance may be uniquely determined from four power-meter readings (if the six-port junction is such that the power meters are sufficiently independent of each other). If diode detectors sample the power along a coaxial line at intervals of $\lambda/8$, then the differences between the outputs of alternate pairs of detectors yield a Smith chart display directly, and the outputs of three of the four detectors together with a sample of the incident power can be used for six-port operations. Six-port junctions

Fig. 6. Circuit of a Q meter.

are calibrated by connecting a number of impedance standards in place of the unknown.

Automatic network analyzers. The most modern impedance measuring units are classed as automatic network analyzers, which may be scalar automatic network analyzers (SANA), measuring insertion loss, gain, return loss, voltage standing-wave ratio, and power; and vector automatic network analyzers (VANA), which, in addition, measure vectorial quantities, like the complex S-parameters of networks.

The S-parameters (scattering parameters) are defined by Eqs. (10), where a_1 and a_2 are the incident

$$b_1 = S_{11}a_1 + S_{12}a_2 \qquad (10a)$$

$$b_2 = S_{21}a_1 + S_{22}a_2 \qquad (10b)$$

signals on ports 1 and 2 of a two-port network, and b_1 and b_2 are the emerging signals of these ports, respectively. The parameters S_{ij} represent signal reflection coefficients, and S_{ij} represent signal transmission coefficients. These quantities are measured by using internal directional couplers which sample the incident and reflected signals in the two measuring ports (between which the network to be analyzed is connected) and, after internal linear frequency down-conversion using mixers, the voltage amplitudes, their ratios, and phase differences are measured with precision low-frequency voltmeters and phase meters. The resulting quantities can be displayed (in the frequency domain, that is, as functions of frequency) on a linear or logarithmic scale versus frequency, or on the Smith chart.

Vector automatic network analyzers are available in numerous frequency bands, from 5 Hz to above 100 GHz. Because of the broad frequency coverage of these instruments, the internal directional couplers used cannot be of sufficiently high directivity, and their use would result in unacceptably large errors in the measurements. This difficulty is overcome by digital error correction. The automatic network analyzer is first calibrated by connecting to it the necessary number of impedance standards, and then the instrument performs the necessary measurements, stores the results, and computes the error-correcting terms at all the measured frequencies. After calibration, the automatic network analyzer is ready for use, and the displayed results are digitally corrected at all the measured frequencies. As an option, voltage automatic network analyzers can be equipped with a time-domain facility which, after digital Fourier transformation of the time-domain results, plots the results as a function of time (that is, distance along the component tested). This permits finding the locations of troublesome reflections along transmission-line components, coaxial lines, or waveguides. *See* FOURIER SERIES AND TRANSFORMS.   Peter I. Somlo

Bibliography. A. C. Lynch and A. E. Bailey (eds.), *Radio Frequency Bridges*, 1988; P. I. Somlo and J. D. Hunter, *Microwave Impedance Measurement*, 1985; M. Sucher and J. Fox, *Handbook of Microwave Measurements*, 1963; M. Valkenburg (ed.), *Reference Data for Engineers: Radio, Electronics, Computers, and Communications*, 8th ed., 1993.

Radio paging systems

Systems consisting of three basic elements—a personal paging receiver, radio transmitter, and an encoding device—whose primary purpose is to alert an individual, or group of individuals, and deliver a short message, typically of a temporary or perishable nature. Characteristics that are used to define a specific paging system include distance covered, radio frequency, modulation type, paging code format, and message type.

Every country has a radio regulatory agency, such as the Federal Communications Commission (FCC) in the United States, which regulates radio spectrum usage and issues licenses for the operation of radio transmitters. Most licenses for radio paging systems limit the number of transmitters and power level that a system is allowed to use. These restrictions limit the distance over which a paging signal may be transmitted. On-site systems cover a single building or a small complex of buildings typically utilizing one low-power transmitter. Wide-area systems can cover an entire city or country and usually use multiple transmitters which simulcast the paging signals.

Nationwide or international paging can be accomplished in several ways. Many countries have allocated specific radio spectrum to provide multicity nationwide paging on a single radio channel. Systems using one of these channels either will simulcast all their transmitters throughout the nation or, at the paging user's direction, will route a paging signal to a specifically requested region or city. This allows users of these systems to travel throughout the country and still receive their messages. Paging in Europe is standardized on a system named ERMES. Users traveling throughout Europe can have their paging messages sent to the country or region to which they are traveling. This protocol allows for the use of up to 16 designated radio channels, and an ERMES paging receiver scans all of these channels, looking for its unique pager address code.

Independent nationwide paging system operators who use the same radio channel in their respective countries have coordinated the assignment of their codes and have connected their paging systems to one another. This has allowed the paging users of these systems to travel to other countries and carry their "hometown" personal pager with them to receive their messages.

Most paging systems now utilize the very high-frequency (vhf) or ultrahigh-frequency (uhf) radio spectrum using frequency modulation (FM). There are systems which use a special subcarrier on an FM broadcast transmitter, and others which use the citizens-band (CB) channels. The latter is often used for "auto-theft" pagers and small on-site systems using amplitude modulation (AM). *See* AMPLITUDE MODULATION; FREQUENCY MODULATION; RADIO SPECTRUM ALLOCATIONS.

Paging formats. Each individual paging receiver is alerted by transmitting its unique address code or codes. Many paging formats are available, and they fall into two categories—analog and binary digital.

Analog codes use audible, or sometimes subaudible, tones. The most common analog formats are two-tone sequential and five/six-tone decimal digital. Two-tone typically provides for several thousand individual codes, while five/six-tone allows 1 million unique codes and is used in many wide-area systems. Binary digital paging formats are composed of binary patterns and can support tens of millions of codes. Moreover, binary digital codes are particularly well suited for encoding numeric and alphanumeric messages.

An additional characteristic of paging formats is the speed at which they transmit the paging code and message. The most common digital format of the 1980s initially was transmitted at 600 bits per second. It was modified to 1200 bps, and there are a few implementations at 2400 bps. The increased number of paging users and the longer messages that they receive have given rise to a family of high-speed, flexible paging codes. One such code enables paging systems to operate at 1600, 3200, or 6400 bps, thus allowing system operators to adapt their system as needs arise. The system uses receivers that can dynamically adjust to any of the speeds.

To complement this code, and to allow a paging user to respond to a message, a two-way paging format has been implemented. This enables the operation of a class of advanced paging services which the FCC has designated narrow-band personal communication services (NBPCS).

Paging receivers. Paging receivers fall into four basic categories: tone alert, tone and voice, numeric, and alphanumeric. Tone pagers emit a beep when they are signaled. Some models silently alert the user with a vibration in place of a beep; other models use differing staccato beeps to provide the user with several alert messages. Tone and voice pagers allow the initiator of the page to transmit a simple voice message which will follow a pager's beep alert. Numeric pagers, sometimes called digital pagers, allow the initiator to convey numerical information. These messages are typically composed by using a tone telephone key pad. Alphanumeric pagers (**Fig. 1**) allow the initiator to send a complete textual message to the pager user. These messages are composed on word processors, personal computers, or dedicated

Fig. 2. Card pager. (*Motorola Inc.*)

terminals which can connect to a paging terminal. *See* MICROCOMPUTER; WORD PROCESSING.

In addition to the basic pagers, there are advanced devices. The pager card (**Fig. 2**) is a pager designed in a standard computer card format. This allows the card to be inserted into a computer or personal digital assistant standard accessory slot, enabling these devices to wirelessly receive updated information or messages. The two-way pager includes an integral transmitter, allowing the user to respond to a received message. This helps eliminate the need for a paging user to find a telephone to respond to a page. A class of alphanumeric pagers allows the receipt and management of selective broadcast information. Services such a news headlines, financial information, and sports scores can be sent to a large number of interested users. These information pagers can be combined with standard paging functions, or can be manufactured as stand-alone information appliances.

Encoders. Every paging system contains an encoding device which provides the interface between the page initiator and the pager user. The encoder's basic function is to receive a pager number and message from the initiator and convert this information into the correct format to be transmitted to the paging receiver. The output from the encoder is connected directly to the system's radio transmitter. Encoders range from simple manual desktop models to large telephone interconnected central offices. Manual encoders are used to initiate pages from a single location such as a dispatcher or switchboard operator. In many large on-site or area-wide systems, it is desirable to initiate pages from the standard telephone network. These systems use an interconnected automatic paging terminal. In addition to generating the

Fig. 1. Alphanumeric personal message pager. (*Comfort Net*)

correct pager code and message format, terminals develop page/message queues, store system usage statistics, and even store short voice messages in solid-state recorders. *See* MOBILE RADIO; TELEPHONE SERVICE.   James A. Wright

Bibliography. D. Baker, *Comprehensive Guide to Paging*, 1992; C. A. Bean, *Overview of the U.S. Radio Paging Industry*, 1988; N. J. Boucher, *The Paging Technology Handbook*, 1998; W. C. Y. Lee, *Mobile Communications Design Fundamentals*, 2d ed., 1993; Telocator, *The Future of Paging II*, 1993.

Radio receiver

The part of the radio communications system that extracts information from radio-frequency (rf) energy intercepted by the antenna. Radio receivers are the most common electronic equipment worldwide and a vital part of all radio, television, and radar systems. Since the 1960s, radio receiver performance has improved greatly, while size, weight, and cost have fallen dramatically. In the past, radio receivers were built from analog circuits, but increasingly they are realized by digital signal processing. *See* RADIO; SIGNAL PROCESSING.

The antenna intercepts a band of energy in the radio frequencies containing many transmissions. These may have different modes of modulation; the two most common are amplitude modulation (AM) and frequency modulation (FM). Signals have a large size range, from a large fraction of a volt down to a small fraction of a microvolt. The receiver must be selective, responding to only one signal, must demodulate the signal, extracting the impressed information from the radio-frequency wave, and must raise it to an acceptable power level by amplification.

Amplitude-modulation and frequency-modulation receivers will be described because they are the most common, but other modes of modulation are also used. Single-sideband (SSB) transmissions are similar to amplitude modulation, but with one of the symmetrical pair of sidebands eliminated and the carrier suppressed. Single-sideband is significant because it conserves electromagnetic spectrum, introducing less spectrum pollution than any other modulation. Because of the growing prevalence of digital signals, digital modulation systems are increasingly important. In the most common of these, known as binary frequency shift keying (FSK), the 0 and 1 states of the binary sequence transmitted are represented by two frequencies. Binary frequency shift keying can thus be regarded as a digital variant of frequency modulation. Systems of digital modulation that combine a form of amplitude modulation (called envelope shaping) with frequency modulation are becoming more common, the purpose again being to reduce spectrum pollution. *See* AMPLITUDE MODULATION; FREQUENCY MODULATION; MODULATION; SINGLE SIDEBAND.

Amplitude-modulation architectures. A simple receiver for amplitude-modulated signals is shown in **Fig. 1**. A band-pass filter selects the required signal, which, after optional amplification, is passed to the demodulator (obsolete term: detector), which in this version consists of a limiting amplifier and a multiplier. The high-gain limiting amplifier has an output of the same sign as the input but constant magnitude (a square wave). When multiplied by the amplitude-modulated waveform, this inverts negative half-cycles. After low-pass filtering to remove residual radio-frequency ripple, the modulating waveform is obtained, shown as an audio waveform, which passes to the loudspeaker. There are other amplitude demodulator circuits, all of which either suppress or invert alternate half-cycles of the amplitude-modulated waveform. *See* AMPLIFIER; AMPLITUDE-MODULATION DETECTOR; ELECTRIC FILTER.
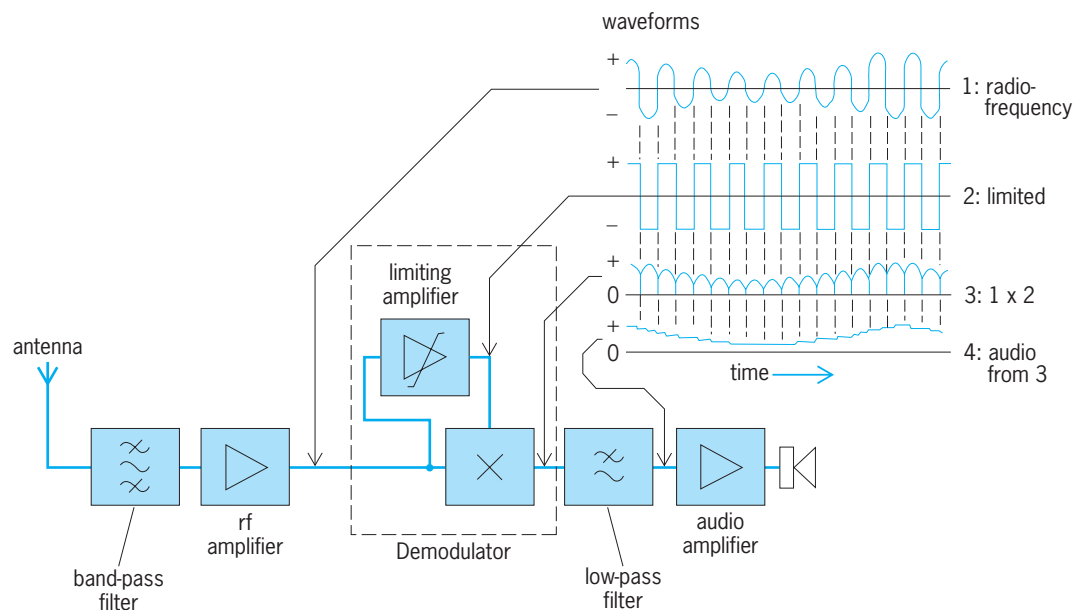

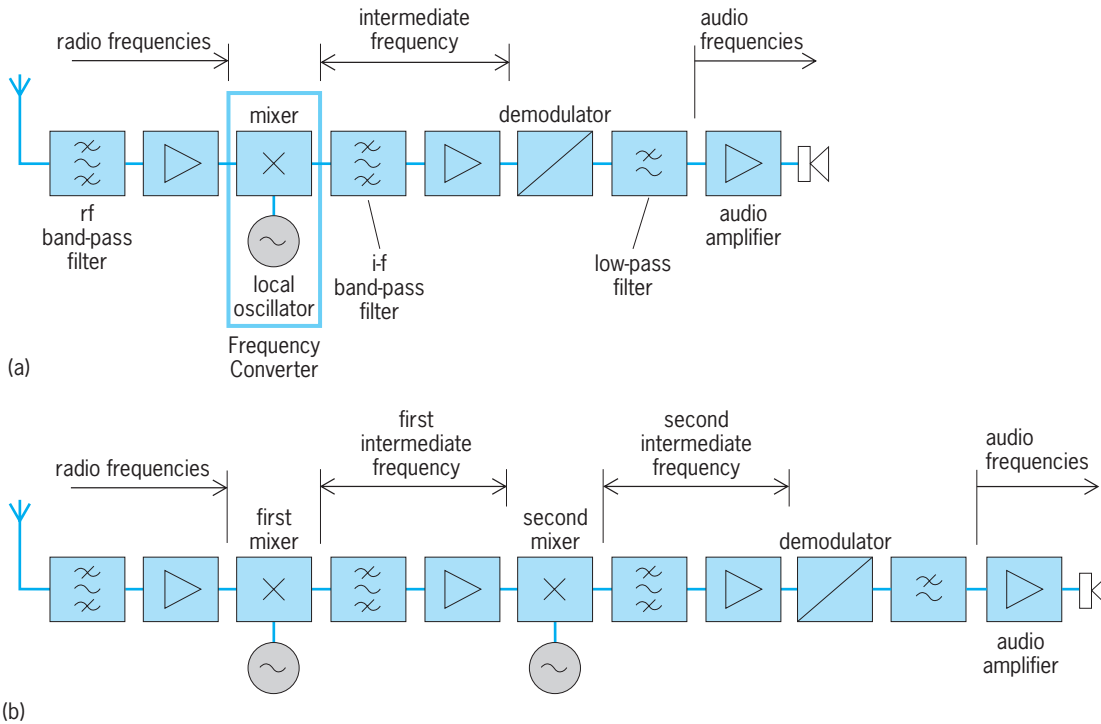Fig. 1. Simple tuned-radio-frequency (TRF) receiver for amplitude-modulated (AM) signals.

Fig. 2. **Superheterodyne (superhet) receivers. (*a*) Basic architecture. (*b*) Double superheterodyne, used to combat image response problems.**

Since the input signal may be on the order of microvolts whereas volts are required to operate a loudspeaker, the total receiver gain may range up to the order of a million or more. In this simple tuned-radio-frequency (TRF) receiver it is difficult to adjust the center frequency of a high-performance band-pass radio-frequency filter in order to receive other stations; also, the radio frequency may be too high for effective amplification. For this reason the superheterodyne (superhet) receiver (**Fig. 2***a*) superseded it.

Superheterodyne receivers. After a radio-frequency filter and amplifier, the signal passes to a mixer, from which the output is formed by the product of the signal and a locally generated wave. If the signal carrier is at f_c and the local oscillator is at f_o, the result is a wave at a frequency $\pm(f_o - f_c) = f_{i\text{-}f}$, known as the intermediate frequency (i–f), and this difference frequency can be kept constant, whatever the value of f_c, by a suitable choice of f_o. Subsequent to the mixer, the superheterodyne receiver becomes similar to the tuned-radio-frequency receiver except that it is now operating at a fixed frequency, so that, provided a low intermediate frequency is chosen, amplification is easy and high-performance filters can be used. Fixed-frequency intermediate-frequency filters, based on mechanical resonance in piezoelectric ceramics and crystals or on surface-acoustic waves, give extreme attenuation outside the passband. The superheterodyne architecture may be used with amplitude modulation, frequency modulation, and other modulation types. *See* HETERODYNE PRINCIPLE; MIXER; OSCILLATOR; PIEZO-ELECTRICITY; SURFACE-ACOUSTIC-WAVE DEVICES.

Assuming that a low value of $f_{i\text{-}f}$ is chosen, a prob-lem arises because of image response. If the carrier is below the oscillator frequency at $f_c = (f_o - f_{i\text{-}f})$, it follows that a frequency $f_s = (f_o + f_{i\text{-}f}) = (f_c + 2f_{i\text{-}f})$ will give the same result, since the sign of the difference is not significant, and the receiver will therefore have a spurious response at f_s, known as the image frequency, separated from the wanted carrier by $2f_{i\text{-}f}$. It is this image response that the first radio-frequency filter must suppress. This is easier the larger the value of $f_{i\text{-}f}$, in conflict with the requirement for a low value explained above. Usually a compromise value of $f_{i\text{-}f}$ is chosen, but in exacting cases the receiver architecture may be changed to a double superheterodyne (Fig. 2*b*), using a relatively high first intermediate frequency which is followed by a second conversion to a lower intermediate frequency, thus gaining the advantages of both at the cost of greater complexity.

Automatic gain control. In amplitude-modulation systems, the amplitude of the signal carries the information, and so to keep the audio output from wide level changes it is desirable to adjust the strength of the signal at the demodulator to be roughly constant, despite variations at the antenna due to transmitter power or range. Thus, automatic gain control (AGC) is used in most amplitude-modulation receivers. The output from the demodulator splits, and, as well as going forward to the audio stages, passes to a low-pass filter which cuts off at a few hertz, well below the audio frequencies. The output from this filter is a voltage proportional to the mean amplitude of the amplitude-modulated wave. After amplification, it controls the gain of the amplifiers. If the amplitude-modulated signal rises, and hence the

control voltage also rises, the gain is cut back, tending to restore the original demodulator output level. *See* AUTOMATIC GAIN CONTROL (AGC).

Single-sideband receivers. Receivers for single-sideband signals are similar to amplitude-modulation receivers, except for the demodulator, a frequency converter in which the single-sideband signal at a nominal carrier frequency converted to $f_{i\text{-}f}$ is mixed with a local oscillator also at $f_{i\text{-}f}$, thus translating the carrier frequency to zero and the sidebands to their correct audio frequencies. Because there is no constant carrier component that can be used as an amplitude reference, automatic gain control presents problems in single-sideband receivers.

Frequency-modulation architectures. For frequency modulation, the signal is of constant amplitude but varies in frequency, and so it is usual for the intermediate-frequency amplifier to be a limiter, giving a constant output. In other respects the receiver may be identical with one of those in Fig. 2. Automatic gain control is often omitted, but may be combined with limiting for the highest performance. The frequency demodulator is quite different, often utilizing a phase-sensitive detector (PSD). *See* LIMITER CIRCUIT.

The phase-sensitive detector (**Fig. 3***a*) consists essentially of a multiplier and a low-pass filter, shown in Fig. 3*b* working with square-wave inputs, from limiters. When the two waves are in phase the output from the filter is positive (Fig. 3*c*), but as they shift in phase it declines, reaching zero when the waves are at 90°, then going to a negative peak when they are at 180°, returning to zero at 270°, a positive peak again at 360°, and so on.
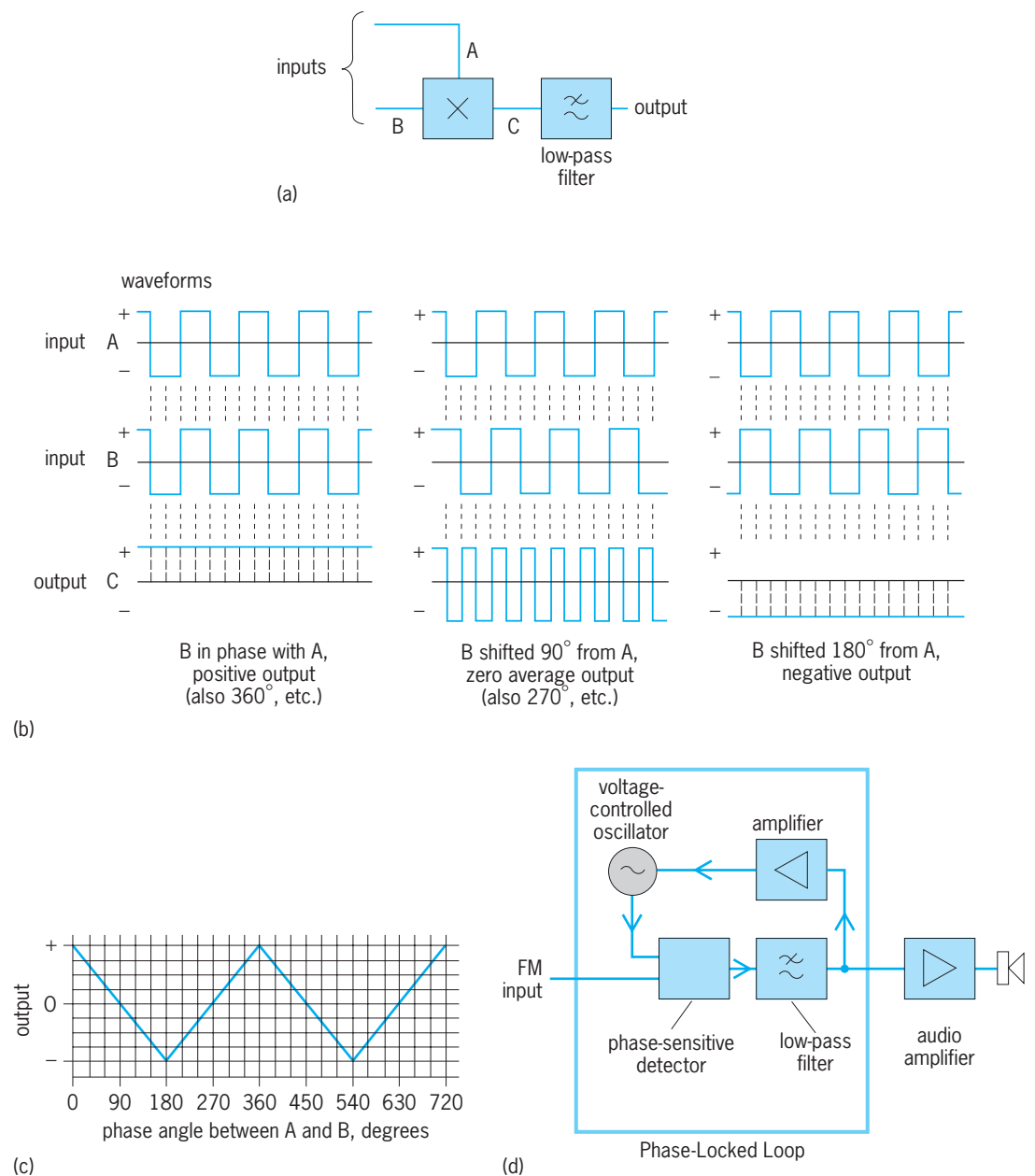

Fig. 3. **Phase-sensitive detector (PSD). (*a*) Diagram. (*b*) Waveforms of inputs and multiplier output. (*c*) Output from low-pass filter versus phase angle from inputs A and B. (*d*) Frequency demodulator using a phase-locked loop (PLL) with a phase-sensitive detector.**

A sophisticated frequency demodulator uses the phase-sensitive detector to create a phase-locked loop (PLL; Fig. 3d). An important component is a voltage-controlled oscillator (VCO). Its frequency can be controlled over a range larger than the excursion of the frequency-modulated wave by an externally applied voltage. This can be achieved on a silicon chip in many ways, such as by charging a frequency-determining capacitor from a voltage-controlled current source.

Since the frequency of a wave is the rate of change of its phase, a small frequency error between the voltage-controlled oscillator and the frequency-modulated waveform may be expressed as a phase error growing larger with time, and thus produces an increasing output from the phase-sensitive detector. This output is amplified and applied to the voltage-controlled oscillator, where it pushes the frequency toward the correct value if it is of the right polarity. (If the polarity is at first wrong, the phase simply slips into the next half-cycle, where the sense of variation of output with phase is reversed.) The amplifier gain is large, and so for an output in the working range its input must be very small. The phase of the voltage-controlled oscillator must lock to a value very close to 90° (or a multiple) from that of the frequency-modulated waveform, since only then is the output of the phase-sensitive detector minimized.

Once the loop has phase-locked, the frequency of the voltage-controlled oscillator faithfully follows that of the frequency-modulated wave, and so its frequency-control voltage represents the frequency deviation of the frequency modulation. This is the basis of the phase-locked-loop frequency demodulator. *See* FREQUENCY-MODULATION DETECTOR; PHASE-LOCKED LOOPS.

A radio for receiving frequency-modulation broadcasts may generate a stereo audio signal pair, which consists of left and right audio signals. In such receivers, the frequency demodulator produces a multiplex signal which contains both sum and difference signals (**Fig. 4**). The sum and difference signals are further processed to create left and right audio signals, which are then supplied to a pair of stereo speakers. To facilitate extraction of the difference signal from the multiplex signal, the receiver uses a low-level pilot signal also present in the multiplex signal. The difference signal is located within the multiplex signal as a double-sideband signal.

The process of extracting the sum and difference signals and the creation of the left and right audio signals is shown in **Fig. 5**. The 19-kHz pilot signal is doubled to create a 38-kHz signal, which is used to frequency-shift the difference signal to its normal audio position at 0 Hz. This operation uses a mixer similar to the mixer shown in Fig. 1. The sum and difference signals are isolated in filters and then added and subtracted to produce left and right audio signals. *See* STEREOPHONIC RADIO BROADCASTING.

The multiplex signal may contain a data signal. In the United States, this signal allows reception of information from the source of the radio broadcast, in accordance with the Radio Data Broadcasting System (RDBS) standard (Fig. 4). The information may
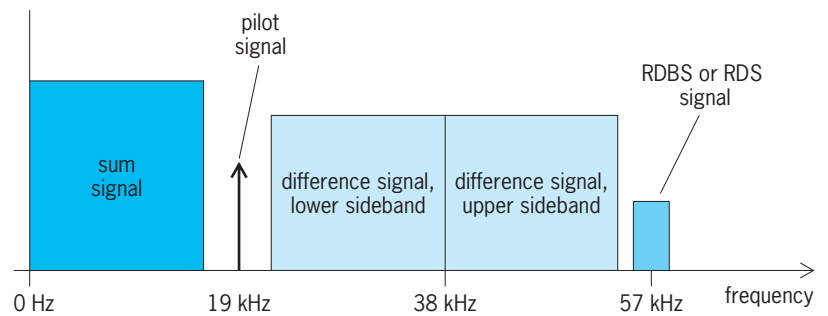

Fig. 4. Frequency spectrum of frequency demodulator output in a radio receiver that generates stereo audio signals, showing content of the multiplex signal.
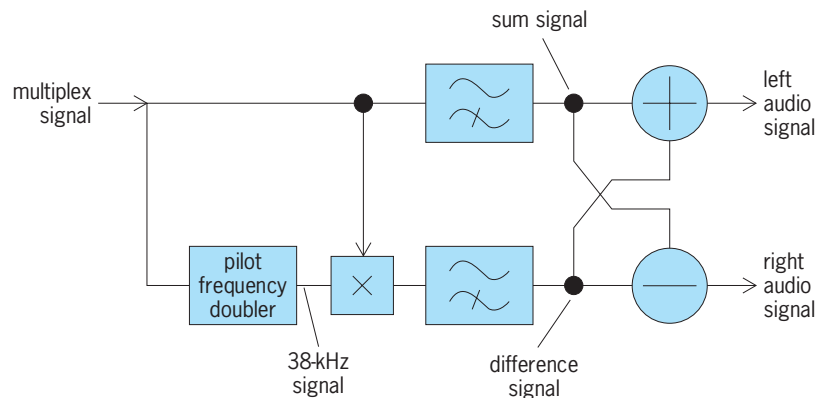

Fig. 5. Process to extract left and right audio signals from a multiplex signal.

include identification of the broadcaster. In Europe, a similar system operating according to the Radio Data System (RDS) standard allows reception of information on a network of broadcasters providing the same audio material, allowing the receiver to automatically select the broadcaster providing the signal with the least reception impairments.

Typical consumer radio receivers provide reception of both amplitude- and frequency-modulated signals. Such receivers employ common amplifier, filter, and mixer circuits that operate for receiving either type of signal, depending on which is selected by the listener. The use of common circuits allows a lower cost to manufacture the radio.

Homodyne receivers. The image frequency response in a superheterodyne receiver occurs at a frequency $2f_{i-f}$ from the wanted signal. If f_{i-f} is reduced to nearly zero, the image and wanted responses coalesce. This is the basis of homodyne, or zero-intermediate-frequency, receivers (**Fig. 6**). The intermediate frequency (corresponding to the signal carrier wave) is made zero by setting the local oscillator to the carrier frequency. Sidebands, whether higher or lower, are converted to low frequencies, and all adjacent channels come out higher. The desired signal may be selected by low-pass filters, which are easier to realize than band-pass filters.

Frequencies both above and below the carrier are converted into the same range (since the sign of the frequency difference is immaterial). With a symmetrical spectrum signal, such as in amplitude modulation, this would not matter if the converted lower sidebands fell exactly on the converted upper
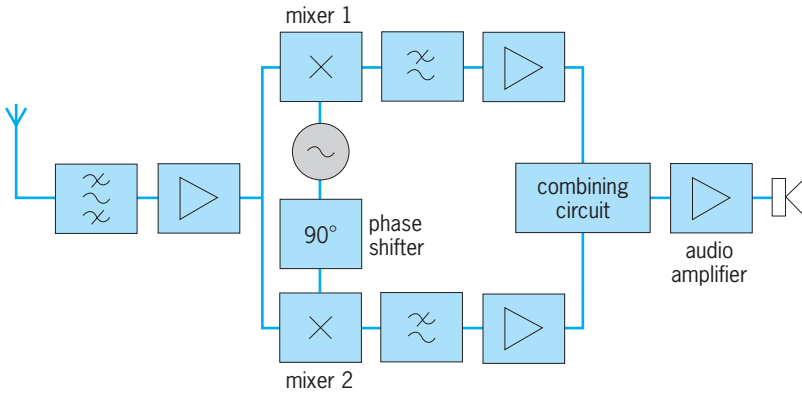
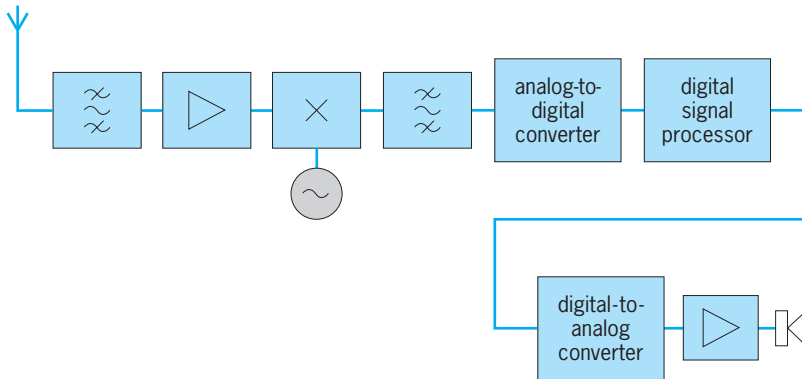Fig. 6. Homodyne (zero-intermediate-frequency) receiver.

Fig. 7. Superheterodyne receiver with digital signal processing.

$$e_1 = E \cos \left[2\pi(f - f_c)t\right] \qquad (1)$$

$$\frac{de_1}{dt} = -2\pi(f - f_c)E \sin \left[2\pi(f - f_c)t\right] \qquad (2)$$

$$e_2 = E \cos \left[2\pi(f - f_c)t + \frac{\pi}{2}\right] \qquad (3a)$$

$$\text{for } (f - f_c) > 0$$

$$e_2 = E \cos \left[2\pi(f - f_c)t - \frac{\pi}{2}\right] \qquad (3b)$$

$$\text{for } (f - f_c) < 0$$

sidebands, but this could happen only if the frequency of the local oscillator were exactly locked to that of the incoming carrier, which is difficult because the incoming carrier is so weak. In any event, the method would not work for asymmetrical spectrum signals such as single sideband.

A better alternative is to use two frequency converters with oscillator inputs of the same frequency $f_o = f_c$ but at 90° phase difference (Fig. 6), which makes it possible to separate the upper and lower sideband products. If the oscillator that feeds one of the mixers is in phase with the signal carrier, then, whether the signal frequency is above or below the carrier, the output will be the difference frequency, without phase shift. However, if the second mixer is 90° ($\pi/2$ radians) phase-advanced relative to the carrier, its output will be phase-advanced by 90° ($\pi/2$) for frequencies above the carrier and phase-retarded by 90° ($-\pi/2$) for frequencies below, since in the latter case the sign of the frequency difference $f - f_c$ is negative, and hence opposite to that of the phase angle. The output phases of the two carriers are summarized in the **table**.

This difference in phase relationships between the two outputs makes it possible to distinguish positive and negative frequency excursions, and demodulation is facilitated. For example, in the case of a frequency-modulated signal, the output from mixer 1 is given by Eq. (1) and its derivative by Eq. (2), where E is the maximum output. The output from mixer 2 is, similarly, given by Eqs. (3). Multiplication of the differentiated output of mixer 1 and the output of mixer 2, after low-pass filtering to remove radio-frequency components, results in a voltage proportional in sign and magnitude to $f - f_c$, forming a frequency demodulator. Different combining circuits can demodulate amplitude, single-sideband, and other modulations.

The homodyne configuration is well suited to digital signal processing in place of analog circuits. By inserting analog-to-digital converters in the two intermediate-frequency channels, the subsequent filtering, processing, and combining can be performed by means of digital calculations, using a digital processing chip. After combining in this way the digital signals are reconverted to analog form for the audio circuits. Receivers of this kind are inexpensive to fabricate and are increasingly commonplace, as in cellular telephones and consumer broadcast receivers. Superheterodyne receivers may also employ digital signal processing, where the output of the analog superhet mixer and filter intermediate-frequency circuitry is converted to digital form and mixed, filtered, and demodulated by digital calculations. *See* ANALOG-TO-DIGITAL CONVERTER; DIGITAL-TO-ANALOG CONVERTER.

A typical arrangement of a superheterodyne receiver with digital signal processing is shown in **Fig. 7**. A significant number of the functions of the receiver is processed in the digital domain, including filtering, amplifier, automatic gain control, mixing, and demodulation, and this processing may be performed in one or more digital processing chips. A key characteristic of most digital signal processing circuits is programmability. For a receiver, the digital signal processor can be programmed to process either amplitude- or frequency-modulated signals. Within the digital signal processor, a homodyne structure may additionally be employed. Processing may also be included to extract the RDBS or RDS signals.

Output phases of mixers in a zero-intermediate-frequency receiver		
	Output phase	
	Mixer 1	Mixer 2
$f > f_c$	0	+90° or $\pi/2$ radians
$f < f_c$	0	−90° or −$\pi/2$ radians

Phase-locked-loop applications. The phase-locked loop has important applications beside the frequency demodulator described above. One is for tracking receivers, which are able to follow a signal drifting in carrier frequency. These are used, for example, in space vehicles, where, because of the high velocity, there is a considerable Doppler shift of received frequencies. *See* DOPPLER EFFECT.

To overcome this problem, the receiver (frequency modulation or amplitude modulation) uses a phase-locked loop feeding back to the frequency converter. In this so-called long-loop receiver, this feedback pulls the frequency of the first oscillator so as to center the wanted signal at $f_{\text{i-f}}$. The intermediate-frequency filter can then be made just wide enough to accommodate the signal without adding extra width to accommodate frequency drift, so that interfering signals or noise is reduced.

A more important application of phase-locked loops is in frequency synthesizers. Since the intermediate frequency $f_{\text{i-f}}$ is fixed in a superheterodyne receiver, the value of f_o determines the signal f_c that is received, so the frequency stability of the local oscillator is critical. Sometimes receivers use quartz crystals as piezoelectric mechanical resonators to determine oscillator frequency directly [they can give better than 20 parts per million (ppm) repeatability without difficulty], but since each crystal is ground for a single frequency, many crystals are needed to receive different transmissions. A better solution is to use one crystal as a high-accuracy clock (as in a quartz watch), deriving all required frequencies from it by using a phase-locked loop. *See* QUARTZ CLOCK.

In such a frequency synthesizer (**Fig. 8**), a crystal oscillator generates a stable frequency f_Q, which passes to a counter circuit (easily realized in digital electronics), where it is counted down to another frequency, say f_Q/n. Similarly, a voltage-controlled oscillator (part of a phase-locked loop) is counted down by m to f_o/m, in this case by a programmable counter, so that m can be set by external command. The phase-locked loop brings the two counted-down waveforms into phase lock by pulling the frequency of the voltage-controlled oscillator. Hence $f_Q/n = f_o/m$, or $f_o = m \cdot (f_Q/n)$. *See* DIGITAL COUNTER; FREQUENCY DIVIDER.

Usually radio channels are equally spaced in the bands, and thus f_Q/n is chosen to give the correct channel spacing; for example, with a 5-MHz crystal, if $n = 200$, the channel spacing will be 25 kHz, a commonly used value in the ultrahigh frequencies. The value of m is then selected to correspond to the channel required and changed whenever the receiver is retuned. With a synthesizer of this kind and a crystal maintained at a constant temperature, it is not difficult to attain a frequency setting accuracy of under 1 ppm.

Other receiver types. Radio receivers are used in many types of broadcast and communication systems, each of which uses some combination of amplifiers, filters, mixers, and demodulators to fulfill the purpose of that system. More sophisticated systems may combine the basic principles of superhetero-
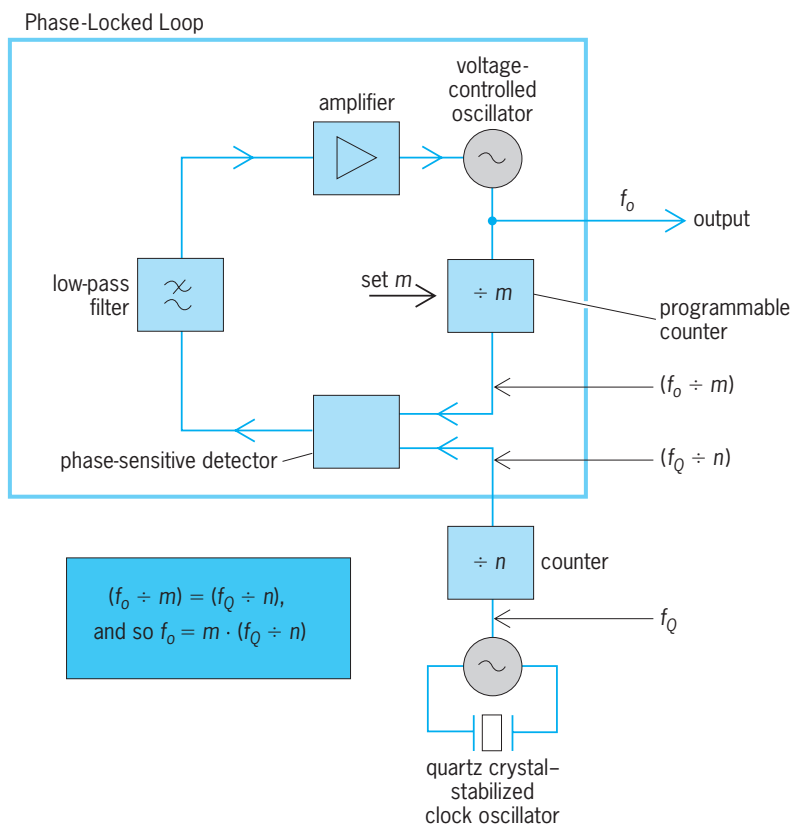

Fig. 8. Frequency synthesizer used to generate accurate local oscillator frequencies.

dyne or homodyne architectures with digital modulation techniques to allow reception of complex signals, such as those containing large amounts of data. Many receiver systems now employ digital signal processing, which in many cases is required for their practical realization.

Television. A television receives radio frequencies modulated with a video signal representing the picture to be displayed. Television receivers typically employ the superheterodyne architecture, with the picture and picture synchronization signals obtained from the demodulator. Analog television receivers use vestigial sideband modulation, a form of amplitude modulation. A digital television receives the picture in digitally modulated form, which provides higher picture quality than that of analog modulation. *See* TELEVISION RECEIVER.

Rake receiver for cellular telephones. In code division multiple access (CDMA) cellular telephones, the rake receiver is used to take advantage of multipath appearing in the received signal. A signal traveling from a transmission source to a receiver typically takes multiple paths due to reflections from buildings and other objects in the environment. Each path will have a different reduction in signal power and a different time delay. The signals traveling on each path combine at the point in space of the receiver antenna input. A rake receiver locks to a reference signal inserted at the transmission source to separate the components in the received signal that have taken the different paths to reach the receiver. Once separated, the signal from each path

is demodulated independently, each in a finger of the rake, and the demodulator outputs are time synchronized and then combined to create a single received signal. The use of the multiple paths in this manner results in a higher-quality received signal than that which could be obtained by only demodulating the incoming combined signal. *See* MOBILE RADIO; SPREAD SPECTRUM COMMUNICATION.

Multiple-carrier systems. Multiple-carrier systems have been developed for transmitting and receiving signals consisting of high-speed digital data streams. The multiple carriers are each digitally modulated to represent data bits, using for example phase-shift keying. The bits from all the carriers taken together are structured to represent some type of digital data signal or signals, such as multiple digital audio data streams. Using a method called orthogonal frequency division multiplexing (OFDM), the carriers are arranged to allow the receiver to easily separate the carriers and extract the bits from each carrier without interference from the other carriers. This approach is used for digital broadcast receivers, in which the multiple-carrier signal is combined with a traditional amplitude- or frequency-modulated carrier.

Impairments. All receivers suffer certain impairments in their performance: the problem of image responses has already been described, together with a solution, the homodyne.

However, the most important impairment is noise. Electrical noise is both received by the antenna and generated within components of the receiver, the first radio-frequency amplifier and mixer being critical, because the signal is weak at this stage. Amplifiers generate less noise than mixers in most radio bands, and so high-performance receivers usually have a radio-frequency amplifier first. Noise can also come from the local oscillator, often run at a high power to improve the ratio of wanted output to noise. Noise is reduced by keeping the receiver bandwidth as narrow as possible, thus reducing the total noise power passed; this is one reason for using tracking receivers, as described above. *See* ELECTRICAL INTERFERENCE; ELECTRICAL NOISE.

A second important impairment is due to nonlinearity, particularly at the front end. The radio-frequency amplifier and converter should give output linearly proportional to their input, but in practice the relationship is better represented by a polynomial. This results in spurious receiver responses—to harmonics and subharmonics—but, more significantly, also in spurious sensitivity very close to the wanted response, due to complex third- and fifth-order interactions. If there are strong unwanted signals near the wanted transmission frequency, these spurious responses cause problems, and are not easily suppressed by the radio-frequency filter. Although better filters can help, there is no solution except to make the early stages of the receiver as linear as possible, and much attention is given to this goal in the design stage. *See* DISTORTION (ELECTRONIC CIRCUITS); LINEARITY.

Impairments to reception performance may also originate from sources outside the receiver.

Amplitude-modulation receivers are susceptible to noise from electrical power lines. Noise pulses from electrical motors and switches can enter receivers and lower performance. To combat degradation due to these pulses, receivers may include special noise blanking circuits to detect and reduce the pulses. Noise may also come from electronic equipment near the receiver.

Multipath can degrade receivers that do not have a means of separating the signal components from the multiple paths, such as in the rake receiver already described. A common type of receiver that suffers from multipath is the frequency-modulated broadcast receiver. The result of the multipath is a drop in the level, or fading, of all or part of the signal entering the receiver, causing bursts of noise in the receiver's audio output. Sophisticated mobile receiver systems may select from two or more antennas coupled into a single receiver or may use multiple antennas coupled to multiple receivers to take advantage of the fact that the multipath condition will usually be less at one of the antennas at different times.

More complex receivers may contain means that allow the receiver to modify its operating characteristics automatically depending on the condition of the received signal. For example, the receiver may detect that the incoming signal is very weak, which will result in a noisy audio signal. To reduce the noise in the audio signal the receiver may respond by automatically adding audio filtering to remove some of the noise.   William Gosling; J. William Whikehart

Bibliography. *The ARRL Handbook for Radio Amateurs*, American Radio Relay League, annually; H. L. Davidson, *Radio Receiver Projects You Can Build*, Tab Books, 1993; S. J. Erst, *Receiving System Design*, Artech House, 1984; K. Feher, *Wireless Digital Communications*, Prentice Hall PTR, 1995; F. M. Gardner, *Phaselock Techniques*, 2d ed., Wiley, 1979; W. Gosling (ed.), *Radio Receivers*, Peter Peregrinus, 1986; H. L. Krauss, C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, Wiley, 1980.

Radio spectrum allocation

The process of identifying specific frequency bands or blocks of the radio spectrum and determining the broad category of radio communication service or services that will be permitted there, and under what specified conditions. The radio spectrum is the part of the natural spectrum of electromagnetic radiation, lying between the frequency limits of 9 kilohertz and 300 gigahertz, that has been extensively used to provide communications services by means of electromagnetic waves. To make effective use of this resource, the radio spectrum is divided, in the frequency dimension, into a series of discrete bands or blocks, arranged in a Table of Frequency Allocations. Because different portions of the radio spectrum are marked by different propagation features and the nature and scope of radio communication services vary greatly, it has become vital that allocation decisions be made in a way that permits radio communication

services to operate in an efficient and effective manner and that minimizes, to the extent feasible, the potential for interference between services. This process, generally, is referred to as frequency management and involves the work of governmental and international bodies. *See* ELECTROMAGNETIC RADIATION.

The radio portion of the electromagnetic spectrum is not subject to a fixed definition, but has gradually expanded as technology has permitted and demand for radio communication services has grown. For example, in 1959 the "upper end" of the usable radio spectrum band was considered to be 40 GHz, which was lifted to 275 GHz in 1971, and is now considered to be 300 GHz. More significantly, there is nearly constant activity within the radio band as allocations are added, deleted, and modified. This is done in order to facilitate the deployment of new radio communication services, to accommodate the expanded needs of existing services, to recognize efficiencies that can be realized through technological improvements or because of diminished use of a particular radio service, and to react to international decisions and treaties between individual countries regarding use of the radio spectrum.

Need for Spectrum Allocation

Two broad principles underlie spectrum allocation. First, the radio spectrum is a finite resource that multiple users can easily render unusable if they cause harmful interference (that is, interference which seriously degrades, obstructs, or repeatedly interrupts a radio communication service) to each other by, for example, trying to simultaneously broadcast in the same band at the same power in the same geographic area. One means of avoiding this potential problem while simultaneously promoting robust radio use is for the government licensing bodies to adopt specific service and technical rules for the use of radio frequencies within a specific band, and to limit entry to specified parties (licensees). Other times, technical protocols developed by standard-setting bodies, such as those associated with Wi-Fi equipment use, can allow multiple users to effectively operate within a frequency band, even though that is lightly regulated by the government and there are minimal barriers to entry. *See* ELECTRICAL INFERFERENCE; ELECTROMAGNETIC COMPATIBILITY.

Regardless of whether a government decides to designate a spectrum band for a narrowly defined radio service or it permits entities to make more flexible use of the band (and regardless of whether it uses licensing as a means of promoting efficient use of a band), such decisions are still made within the framework of the allocation process, and the Table of Frequency Allocations remains the primary means of organizing the radio spectrum. Allocations are sufficiently broad to permit a wide variety of uses, but are designed to avoid conflicts between incompatible uses. For example, a jurisdiction may choose to allow for cellular telephone service by adopting very specific licensing and service rules within a band broadly allocated for mobile use. However, such a mobile allocation would not typically be considered to coexist within a band allocated for broadcasting, due to the fact that it would be difficult to reconcile potential interference between mobile services and broadcasting services operating in the same band.

A second characteristic of radio waves is that they do not respect political boundaries, and radio communication use in one country can negatively affect the ability of citizens of another country to use the same radio spectrum. This is particularly significant in that the propagation characteristics in certain radio bands, such as those associated with shortwave radio and many of the amateur bands, can be heard on the other side of the world. Thus, radio spectrum allocations are necessarily a matter of worldwide concern. The International Telecommunication Union (ITU), an international organization within the United Nations System where governments and the private sector coordinate global telecommunications networks and services, is the primary entity in matters of international radio allocation. The ITU publishes the Radio Regulations, which is a binding international treaty that governs the allocation and use of radio frequencies worldwide. The Radio Regulations includes an International Table of Frequency Allocations (International Table). For purposes of the International Table, the ITU divides the world into three regions and provides specific allocations for different classes of radio services. Within the allocations, there are some differences between regions, and some frequency bands are subject to greater conditions on use than others. At periodic world radio communication conferences, the ITU evaluates spectrum sharing and allocation matters and provides necessary updates and modifications to the International Table.

It is not uncommon to have regional and national variations of the International Table. First, regional bodies exist to develop consensus for and promote common interests in telecommunications-related matters, and may only be able to foster support for changing the International Table in a single region. One such body is CITEL, the Inter-American Telecommunication Commission, that is an entity of the Organization of American States. Also, individual countries retain sovereign control over spectrum use within their borders and may decide to permit a spectrum use that is different from that contained in the International Table. In these cases, countries that are signatories of the Radio Regulations must ensure that such services are provided in a way that does not cause interference with services that are operating in accordance with the International Table.

Within the United States, spectrum management responsibilities are divided by the Federal Communications Commission (FCC) and the National Telecommunications and Information Administration (NTIA), which is a part of the Department of Commerce. The NTIA governs all federal government use of spectrum, while the FCC governs all non–federal government use of spectrum and is responsible for other telecommunications-related issues. Frequency allocations within the United States

are governed by the United States Table of Frequency Allocations (U.S. Table), which consists of the federal government Table of Frequency Allocations and the non–federal government Table of Frequency Allocations. Within the United States, frequency bands are allocated for exclusive federal government use, exclusive non–federal government use, and shared federal government/non-government use. The Federal Communications Commission publishes a Table of Frequency Allocations that includes, for reference, both the U.S Table and the International Table.

Both the ITU and the United States government have incorporated the concept of primary and secondary use into their tables of frequency allocations. Stations of a secondary service are not permitted to cause harmful interference to stations of primary services to which frequencies are already assigned or to which frequencies may be later assigned, and cannot claim protection from harmful interference caused by stations of a primary service but are entitled to protection from interference caused by other secondary stations. This primary–secondary distinction serves to promote more robust use of the radio spectrum and can permit new services to be deployed in already-occupied spectrum bands, while maintaining protections for those services that have been designated as primary.

Making Allocation Decisions

The characteristics of radio waves vary greatly throughout the radio spectrum, and the propagation characteristics of a particular band, along with the type of services expected to be deployed, will greatly influence allocation decisions. Efficient spectrum allocations take into account these characteristics. Because the wavelength of a broadcast signal can range from hundreds of meters to mere millimeters depending on where in the electromagnetic spectrum the operations are located, the distance a radio signal can travel, its ability to penetrate buildings, and its susceptibility to atmospheric conditions can vary greatly. For example, it is advantageous to allocate broadcasting services in bands that are less susceptible to rain fade (the weakening of a radio signal as it passes through raindrops). Other technical considerations include the ability to control interference, both within a given service and with other, adjacent services, as well as apparatus limitations such as antenna size and power requirements. Allocation decisions also involve a number of policy and economic considerations, including the public need and benefits for the service (both standing alone, and in competition with other services that could be provided in the spectrum), as well as the amount of spectrum required to promote efficient use and ensure that the service can be made economically viable.

Allocations to particular services can often be found throughout different frequency bands. Moreover, as suggested above, allocated spectrum is usually further refined by designating it for a particular service, which involves the adoption of specific service and technical rules, and—if the designated service is a licensed service—by assigning it to entities who then hold license rights to develop and deploy systems within the spectrum. As part of this process, policymakers typically subdivide a frequency band into a series of smaller channels, which are sometimes further divided geographically. For example, within the broadband personal communications service (PCS) in the United States, the "A" block consists of two paired 15-MHz blocks licensed in each of 51 Major Trading Areas. The choice of a specific band plan will also take into account a number of technical factors, such as the separation between frequencies that is necessary to avoid harmful interference, the characteristics of the transmitters and receivers that can be expected to be deployed in the band, the likely distance between units being used in the service and receivers operating on an adjacent band, and any constraints on transmitter power and location. Accordingly, the wide variation in the necessary separation between assigned frequencies that are adjacent to each other within the radio spectrum is due to both the unique characteristics associated with each service and the overall propagation characteristics of that portion of the electromagnetic spectrum. *See* BANDWIDTH REQUIREMENTS (COMMUNICATIONS); RADIO-WAVE PROPAGATION.

Allocated Services

Specific allocated services include the broadcasting service, fixed and mobile services, unlicensed services, and amateur and personal radio services.

Broadcasting service. A broadcast allocation permits transmissions that are intended for direct reception by the general public, and includes sound, visual, and other types of transmissions. The commercial AM radio band consists of the frequency range 535 kHz-1.7 MHz, FM radio operates in the band 88-108 MHz, and television stations operate on several frequency bands—54-88 MHz for channels 2-6, 174-220 MHz for channels 7-13 (collectively, the very high frequency or VHF channels), and 470–794 MHz for channels 14-69 (the ultrahigh-frequency or UHF channels). Channels above 70 (794 MHz) are no longer assigned to television. *See* AMPLITUDE-MODULATION RADIO; FREQUENCY-MODULATION RADIO; RADIO BROADCASTING; TELEVISION.

High-frequency (HF) broadcasting, also known as shortwave broadcasting, is an international service that operates between 5900 and 26,100 kHz. Its broadcast signals propagate over very long distances and thus can be received by the general public in foreign countries. However, due to these propagation characteristics, broadcasts in these bands are subject to interference from HF stations around the world, and the bands are extremely congested. Use of these bands is subject to international planning and coordination, and HF broadcasters are authorized frequencies on a seasonal basis that allows broadcasters to account for changes in propagation conditions, altered programming needs, and objectionable interference situations. The 2003 World Administrative Radio Conference in Geneva adopted a transition plan for the HF bands that, among other

things, increases the amount of spectrum allocated to exclusive HF broadcasting use throughout the world to 3720 kHz in ten frequency bands after April 1, 2007, and permits broadcasts to use new digital modulation techniques that result in clearer reception and more efficient use of the spectrum.

For domestic commercial radio bands in the United States, the FCC began exploring rules for digital broadcasting in 2004, and currently permits in-band on-channel (IBOC) broadcasts, which consist of the transmission of a digital radio broadcast signal centered on the same frequency as the AM or FM station's present frequency. Digital radio is expected to provide near–compact disk (CD) quality reception for stations operating in the FM broadcast band. For AM stations, IBOC digital radio is expected to provide reception approximately equal to present-day analog FM reception.

Within the commercial television bands, jurisdictions have been working to establish a transition plan to digital broadcasting. Because the public reception of digital broadcasts will require new or upgraded tuners, the United States established a complex mechanism by which broadcasters were required to commence digital transmissions between 1999 and 2003 (based on the size and classification of the station), but in which they were also permitted to continue analog operations on their existing 6-MHz-wide channels. Analog transmissions will continue until an official cut-off, which is scheduled to take effect at the end of 2006, but which may be extended under particular circumstances set forth by law, including for those markets where 15% or more households do not have access to either digital television equipped receivers or multichannel video. Ultimately, due to the efficiency of the digital standard, television broadcasting will be able to operate in a "core" broadcast spectrum consisting of channels 2–51, but will also be able to greatly enhance the quality of the video, audio, and supplemental data that are being broadcast. The reclamation of television spectrum has been addressed in two proceedings: the "Upper 700 MHz" band (channels 60–69) which comprises 60 MHz, and the "Lower 700 MHz" band (channels 52-59) which comprises 48 MHz. Within the Upper 700 MHz band, public safety entities will receive 24 MHz of the reclaimed spectrum, with the remaining 36 MHz slated for commercial use including fixed, mobile, and broadcasting services. The Lower 700 MHz band, slated to be auctioned in 2005, is designated for flexible use, and can accommodate fixed, mobile, and broadcast services. In both of these bands, new licensees will be unable to deploy fully operations until television operations cease.

Fixed and mobile services. The fixed service is a radio communication service between specified fixed points, whereas a mobile station is intended to be used while in motion or during halts at unspecified points. From an allocation standpoint, it is most common to create a shared fixed and mobile band in order to promote the most flexible use of the radio spectrum. The types of fixed and mobile services are extensive, and these applications serve many vital commercial, private, and public safety radio needs. One example of a fixed service is microwave communications, which are used mostly for point-to-point communication and serve as backhaul and backbone links for commercial wired and wireless networks in remote or rugged terrain. Mobile services include popular commercial radio telephone services—cellular and broadband PCS—as well as private land mobile radio systems that are used by companies, local governments, and other organizations to meet a wide range of communication requirements, including coordination of people and materials, important safety and security needs, and quick response in times of emergency. Because public safety services are provided under fixed and mobile allocations, a consistent theme in making these allocations has been to provide sufficient spectrum for public safety purposes. In 2002, for example, the United States allocated 50 MHz of spectrum in the 4940–4990 MHz band (4.9-GHz band) for fixed and mobile services (except aeronautical mobile service) and designated the band for use in support of public safety.

One development has been the identification of significant blocks of spectrum to support advanced commercial wireless needs. These voice and data services are commonly known as 3G (third-generation mobile) and IMT-2000, for the study that set forth the characteristics of the advanced services that were expected, and concluded that at least 160 MHz of spectrum were needed to be made available in order to support such applications. These services are expected to build on the success of existing cellular mobile phones. The 2000 World Administrative Radio Conference identified three bands that could be used for such purposes: 806–960 MHz, 1710–1885 MHz, and 2500–2690 MHz. While the bulk of this spectrum was intended for terrestrial-based mobile services, a small portion was also identified for a satellite-based component. One difficulty in making use of these bands on a worldwide basis is the fact that many are already used by jurisdictions for fixed, mobile, and broadcasting operations. For example, the 1920–1980 MHz band—identified for mobile unit transmissions to base stations and allocated as such in European jurisdictions—overlaps with the 1930–1990 MHz broadband PCS base station transmit band in the United States. The United States has identified two 45-MHz blocks of spectrum, consisting of the 1710–1755 MHz and 2110–2155 MHz bands, for the deployment of advanced wireless services, and has begun the process to set service rules for and award licenses in the band. Up to 40 additional megahertz within the 1910–2180 MHz range might also be made available for such use.

Spectrum in the 70-, 80- and 90-GHz bands, the so-called millimeter wave bands, has been made available for new fixed applications, including those that could provide increased broadband capacity. These bands had previously been used primarily for radio astronomy and military applications, and until the 2003 decision authorizing operations in the band, the United States had only authorized operation of

radio technology above 50.2 GHz on an experimental basis and prohibited the marketing of devices that could be used for either licensed or unlicensed operations above 77 GHz. *See* MOBILE RADIO.

Unlicensed services. An increasing number of services are provided in bands whose allocations permit unlicensed uses, where there are few regulatory barriers to entry but in which an individual user has no rights to prevent interference caused by other users. Traditionally, such industrial, scientific, and medical equipment (ISM) bands permitted the development and deployment of equipment or appliances that were designed to generate and use energy in the radio-frequency bands but not for communications purposes. Common consumer ISM appliances are microwave ovens and ultrasonic jewelry cleaners. Other ISM categories include industrial heating equipment and medical diathermy equipment. Another type of unlicensed operation involves the use of equipment that is specifically designed to radiate energy for purposes of communication. While this category has typically been dominated by such devices as cordless phones, garage door openers, television remote controls, and toys, recent growth in the use of this spectrum has been in the area of wireless communication devices that use a common platform established by standards-setting bodies but which can be deployed using off-the-shelf consumer-marketed equipment. Examples of these standards include Wi-Fi®, Bluetooth®, and WiMAX.

Wi-Fi. Wi-Fi is short for "wireless fidelity," and is the term used to describe wireless local-area networks operating on a common set of standards. Wi-Fi applications generally are used to connect personal computers and laptops to a network, and can be used both in private settings (such as a home or business) and via public wireless "hotspots." Wi-Fi is a relatively short-range service, with the best reception taking place within several hundred feet (or less) of a base station. In addition, the speed of data transmission decreases the farther one moves from a base station. Wi-Fi was initially deployed in the 2.4-GHz band, which is shared with such other unlicensed devices as cordless phones, and its increasing popularity led to fears that spectrum congestion could stifle its growth. In the United States, Wi-Fi use has expanded to an additional 300 MHz of spectrum in the 5-GHz band. Internationally, a major focus of the 2003 World Administrative Radio Conference was to expand similarly the available bands for Wi-Fi use, both in order to meet anticipated demand and to simplify the manufacturing of equipment though the establishment of worldwide allocations. The Conference provided a 5-GHz allocation that can be used for Wi-Fi, but with the encouragement that regulators restrict its use to the indoors out of concern for possible interference with existing services in the 5-GHz range. *See* WIRELESS FIDELITY TECHNOLOGY.

Bluetooth. Bluetooth is similar to Wi-Fi in that it is based on a wireless standard and is designed to operate in the 2.4-GHz band. However, the Bluetooth standard operates over a much shorter distance, typically in the range of 30 ft (10 m), and was designed primarily to use short-range radio links to replace cables connecting fixed and portable electronic devices. While Bluetooth has been deployed in such devices as the wireless computer mouse and wireless telephone headset, it can be used across technology platforms, and has been incorporated in cellular phones, personal digital assistants (PDAs), laptop computers, and even automotive electronics. Additional wireless standards have been developed and are in the process of finding both commercial support and adequate spectrum for deployment.

WiMAX. WiMAX is a relatively new wireless networking standard that offers greater range and bandwidth than the Wi-Fi standard. For example, WiMAX can transfer around 70 megabits per second over a distance of 30 mi (50 km), and a single base station can serve thousands of users. Although WiMAX is based on standards that allow for the use of any frequency band between 2 and 11 GHz, it has yet to find a spectrum "home" that would allow for widespread deployment and the necessary economies of scale to produce reasonably priced equipment. Moreover, based on the prospective bands identified by proponents of the service, WiMAX could operate on either licensed or unlicensed spectrum, or possibly even both.

Amateur and personal radio services. The Amateur Radio Service is a voluntary noncommercial communication service used by qualified persons of any age who are interested in radio technique with a personal aim and without pecuniary interest. The amateur service dates back to the 1900s, and its users have made numerous contributions to the development of radio. The amateur service is also noteworthy in its allocations and the wide scope of communications employed by its users. Twenty-seven small frequency bands throughout the spectrum are allocated to this service internationally. Some 1300 digital, analog, pulse, and spread-spectrum emission types may be transmitted. Amateurs throughout the world communicate with each other directly or through ad hoc relay systems and amateur satellites, and use voice, teleprinting, telegraphy, facsimile, and television to communicate. Although amateur radio spectrum is often looked upon as a source of additional spectrum by those seeking to deploy or expand commercial and government services, the amateur community has been quite effective in minimizing any encroachment on these bands. The 2003 World Administrative Radio Conference realigned the allocations near 7 MHz in order to expand the worldwide 40-m amateur band by 100 kHz. Once international broadcasters operating in this band relocate to other frequencies in 2009, the band will be allocated exclusively to the amateur service in all three regions. Also, in 2003 the FCC provided access to five channels in or near the 5250–5400 kHz band on a secondary basis for the amateur service, and upgraded the existing secondary amateur service allocation to primary status in the 2400–2402 MHz band. *See* AMATEUR RADIO.

The personal radio services provide short-range, low-power radio for personal communications, radio signaling, and business communications not

provided for in other wireless services. The range of applications is wide, spanning from varied one- and two-way voice communications systems to non–voice data transmission devices used for monitoring patients or operating equipment by radio control. The Citizens Band (CB) Radio is included in the personal radio services, as is the Family Radio Service—a popular Citizen Band service with a 1-mi (1.6-km) range that is well suited for family use in a neighborhood or during group outings. Additions to the personal radio services have supported wireless medical needs. For example, the Wireless Medical Telemetry Service in the 608–614 MHz, 1395–1400 MHz, and 1427–1432 MHz bands is designed for remote monitoring of patients' health through radio technology and transporting the data via a radio link to a remote location.

Other allocations. Many other important radio services have distinct allocations; among the most notable are space- and satellite-based services, the maritime services, and aeronautical services. In addition, important science research takes place within the radio spectrum, and specific bands have been allocated for the radio astronomy and space research services. Radionavigation-satellite services, for example, include the United States' Global Positioning System (GPS) and Russia's *Glonass* satellite system. The scope of allocated services is reflected in the 30 distinct categories of radio service allocations listed in the U.S Table. International radio telecommunication conferences have considered such issues as providing allocations for the mobile satellite service in the 2-GHz band as part of the larger IMT-2000 allocation for advanced services, finding additional spectrum for high-altitude platform stations, and allowing an aeronautical mobile satellite service in the 14-GHz band to permit in-flight access to the Internet. *See* RADIO ASTRONOMY; SATELLITE NAVIGATION SYSTEMS.

Trends and Developments

A consistent theme in spectrum allocation is the quest to meet the demand for more spectrum. Because technological developments are only able to expand the range of usable radio spectrum to a small degree, and the applications that reasonably can be accommodated in the upper frequency bands are limited due to the different propagation characteristics of different radio frequencies, those parties involved in spectrum allocation decisions have had to employ other means to accommodate increased demand for radio spectrum. The primary tool is the promotion of increased sharing. Different services can use the same spectrum if the risk of interference is minimal and the uses are compatible or can be coordinated. However, increased sharing usually comes at the expense of increased complexity in system design and, thus, greater cost.

Perhaps the easiest means of sharing is by way of geographic separation. Ships and railroads are able to share maritime frequencies in the VHF band due to the distance between land-based rail systems and ocean-going vessels. Geographic separation also permits the reuse of frequencies within a particular service, such as a cellular radio telephone network in which the same frequencies are reused across multiple cell sites within the license area.

A second way to make more use of spectrum is through efficient system design. Directional antennas and antenna discrimination permit fixed microwave links to share the same spectrum with satellite Earth station uplinks, for example. The switch from analog to digital modulation techniques, along with improved coding, has served to increase the amount of information that can be transmitted over a fixed quantity of bandwidth, which in turn has permitted "refarming" in some bands. Refarming maintains the same number of existing channels but reduces the bandwidth occupied by each channel, resulting in additional spectrum that can be put to new uses while permitting existing services to continue to be provided on the narrower channels, by way of more efficient broadcasting techniques. In addition, by employing system protocols such as "listen before talk" and moving to trunked systems versus inflexible dedicated frequencies, users can derive more robust use from a limited amount of spectrum. *See* ANTENNA (ELECTROMAGNETISM); DATA COMPRESSION; DIRECTIVITY; MODULATION.

A third mechanism is band clearing, where lightly used spectrum is made available for new services by moving existing operations to other bands, or even other modes of communication. The deployment of broadband PCS involved the relocation of existing fixed microwave users in order to clear sufficient spectrum to permit PCS licensees to deploy high-value mobile phone services in the band. Some of the legacy microwave users were relocated to other frequency bands where they deployed more modern, spectrally efficient equipment. Others were able to meet their communications needs in different ways, such as by use of fiber-optic cables or subscription to commercial communications services. However, growing demand and band congestion makes successful band clearing an increasingly difficult and contentious task.

A notable development in the quest to make more spectrum available is the deployment of complex systems that have been made possible by advances in computer processing. Cellular phone systems—in which phone calls must be seamlessly handed off from one cell site to another and the base and mobile units must constantly monitor and adjust their signals to maintain a robust connection—would not be possible without advanced computer applications. Many of the unlicensed devices are able to make efficient use of spectrum through such techniques as low-power-sharing protocols and adaptive technologies. For example, cognitive radio technologies can enable a radio device and its antenna to adapt its spectrum use in response to its operating environment, allowing the radio to identify available spectrum that is unusable under current conditions and adjust as conditions change. By contrast, older radios are often programmed to transmit and receive a limited number of frequencies, which can be changed only by physically modifying each radio unit. Mesh networks allow individual radios to connect with other radios

to transmit and relay information instead of connecting to a defined base station, offering the potential for a more robust network and reducing the need for expensive fixed base sites. Spectrum underlays permit the beneficial use of frequencies already being used for traditional high-powered services by adding noninterfering low-powered operations onto the same band. Finally, ultrawideband and spread spectrum applications break away from the traditional view of fixed and discrete band use and instead permit the development of applications that use numerous frequency bands. *See* SPREAD SPECTRUM COMMUNICATION; ULTRAWIDEBAND (UWB) SYSTEMS.

The technology that permits these new developments in spectrum use is likely to outpace the political considerations inherent in spectrum use decision making. Incumbent stakeholders have traditionally been reluctant to share bands that they perceive as having an exclusive right to use, and many of the new technologies do not neatly fit the traditional radio spectrum allocation and licensing models. Nevertheless, because the future of radio spectrum allocation is likely to be driven by these types of technological innovations, the traditional view that radio spectrum allocation requires a central policymaker who must identify specific frequency blocks for the exclusive use of discrete and narrowly defined radio services is quickly becoming outmoded. However, because the radio spectrum continues to be in high demand and increasing quantities of information are being transmitted via radio waves, the core purposes of radio spectrum allocation—the efficient use of the radio resource and the prevention of harmful interference—remain vital to the successful and productive use of this resource.   Jamison Prime

Bibliography. Federal Communications Commission, *Rules and Regulations* (annually within vol. 47 of *Code of Federal Regulations*, National Archives and Records Administration, Washington, D.C.); International Telecommunications Union, *Radio Regulations*, 2004 (ITU, Geneva); B. Z. Kobb, *Wireless Spectrum Finder: Telecommunications, Government and Scientific Radio Frequency Allocations in the US, 30 MHz–300 GHz*, McGraw-Hill, 2001; A. K. Maitra, *Wireless Spectrum Management: Making Effective Use of RF Spectrum*, McGraw-Hill, 2004; J. A. Manner, *Spectrum Wars: The Policy and Technology Debate*, Artech House, 2003; P. C. Roosa Jr., *Federal Spectrum Management: A Guide to the NTIA Process*, NTIA Spec. publ. 91–25, 1992.

Radio telescope

An instrument used in astronomical research to study naturally occurring radio emission from stars, galaxies, quasars, pulsars, interstellar clouds, and other astronomical bodies between wavelengths of about 1 mm (300 GHz) and 10 m (30 MHz). At the short wavelength end, the performance is limited by the opacity of the terrestrial atmosphere, and at the long wavelength end by the opacity of the ionosphere.

The performance of a radio telescope is measured by its sensitivity or ability to measure weak sources of radio emission, angular resolution or sharpness of the radio image, and the image quality or ability to image over a wide range of radio brightness. Radio telescopes vary widely in design and appearance, depending on the wavelength of operation and astronomical goals, but they are all composed of one or more antenna elements, each equipped with a radio receiver.

Because cosmic radio sources are extremely weak, radio telescopes are usually very large and equipped with the most sensitive radio receivers available. Moreover, the weak cosmic signals can be easily masked by human-made radio interference, and great effort is taken to protect radio telescopes from locally generated interference by locating them in remote environments, often surrounded by mountains to provide shielding from terrestrial radio transmissions. But with the increasing use of satellite radio transmissions to provide communications, entertainment, and navigation throughout the world, it is increasingly difficult to avoid the effects of radio interference.

The radio signals that arrive at the Earth from cosmic radio sources are as weak as 10^{-19} watt. In order to detect the faint broadband signals generated in interstellar or intergalactic space, radio telescopes are designed to receive over a very large bandwidth of hundreds of megahertz, and the data may be averaged for tens of hours or more to reduce the effect of random noise generated by the radio receiver itself. Also, many atoms and molecules found in interstellar space and around some stars radiate only at discrete frequencies, mostly at short, centimeter and millimeter wavelengths. These spectral lines are studied with radio telescopes able to simultaneously observe up to thousands of different narrow-frequency channels. The most straightforward type of radio spectrometer employs a large number of filters, each tuned to a separate frequency and followed by a separate detector to produce a multichannel, or multifrequency, receiver. However, it is difficult to build a receiver with more than a few hundred frequency channels in this way. Very large multichannel receivers convert a single broad-bandwidth signal into digital form which is then analyzed by the mathematical process of autocorrelation and Fourier transformation to produce the radio spectrum. *See* ANTENNA (ELECTROMAGNETISM); FOURIER SERIES AND TRANSFORMS; RADIO RECEIVER.

Filled-aperture radio telescopes. The simplest type of radio telescope uses a parabolic antenna—the so-called single-dish or filled-aperture radio telescope—which operates in the same manner as a television satellite receiving antenna, to focus the incoming radiation onto a small antenna element called the feed, which is connected to the sensitive radio receiver. The larger the parabolic surface, the more energy is focused on the feed and the more sensitive the telescope. Cryogenically cooled solid-state amplifiers with very low internal noise are used to obtain the best possible receiver sensitivity.

The performance of a radio telescope is limited by various factors: the accuracy of the reflecting surface that may depart from the ideal shape because of manufacturing irregularities; the effect of winds; thermal deformations that cause differential expansion and contraction; and deflections due to changes in gravitational forces as the antenna is pointed to different parts of the sky. Departures from a perfect parabolic surface become important when they are more than a few percent of the wavelength of operation. Since small structures can be built with greater precision than larger ones, radio telescopes designed for operation at millimeter wavelength are typically only a few tens of meters across, whereas those designed for operation at centimeter wavelengths range up to 100 m (328 ft) in diameter.

The largest fully steerable filled-aperture radio telescope currently in operation is the 110 by 100 m (361 ft by 328 ft) antenna (**Fig. 1**) operated by the National Radio Astronomy Observatory in Green Bank, West Virginia. The moving structure of the Green Bank telescope weighs 16 million pounds (7.3 million kilograms) and is able to point to any direction in the sky with an accuracy of a few arcseconds. Each of the 2004 surface panels making up the parabolic surface is held in place by computer-controlled motors which keep the surface accurate to better than half a millimeter, allowing performance at wavelengths less than 1 cm. Green Bank is located in a valley far from high population densities in the National Radio Quiet Zone, which offers protection from local sources of human-made interference. Somewhat less precise 100-m, 76-m, and 64-m diameter (328, 249, 210 ft) antennas are operated by the Max-Planck-Institut für Radioastronomie near Effelsberg, Germany, the Jodrell Bank Observatory near Manchester, UK, and the CSIRO Australia Telescope National Facility near Parkes, NSW, respectively.

The largest filled-aperture radio telescope in the world is the 305-m (1000-ft) fixed spherical reflector operated by Cornell University near Arecibo, Puerto Rico (**Fig. 2**). Although the Arecibo antenna has an enormous collecting area, it can be used only over a limited angle of about $20°$ from the zenith. The Russian RATAN-600 telescope, located near Zelenchukskaya in the Caucasus Mountains, has 895 reflecting panels, each 7.4 m (24 ft) high, arranged in a ring 576 m (1890 ft) in diameter. Using long parabolic cylinders, standing reflectors, or simple dipole elements, researchers in Australia, France, India, Italy, Russia, and Ukraine have also built antennas with very large collecting areas, but operating at relatively long meter wavelengths where the required precision is less demanding. Several smaller, more precise radio telescopes used at millimeter wavelength are located on mountains, where clear skies and high altitudes minimize absorption and distortion of the incoming signals by the terrestrial atmosphere. A 45-m-diameter (148 ft) radio telescope near Nobeyama, Japan, is used for observations at wavelengths as short as a few millimeters. The French-German Institut de Radio Astronomie Millimétrique (IRAM) in Grenoble, France, operates a 30-m (98-ft) antenna
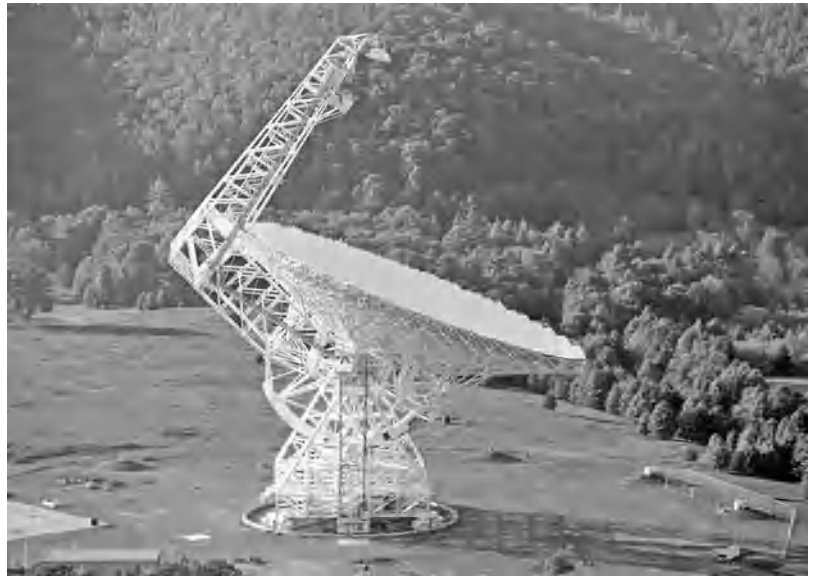

Fig. 1. The 110 by 100 m (361 by 328 ft) fully steerable radio telescope in Green Bank, West Virginia. (*NRAO/AUI/NSF*)

at an altitude of 2850 m (9350 ft) on Pico Veleta in the Spanish Sierra Nevada for observations at wavelengths as short as 1 mm. The 15-m (50-ft) James Clerk Maxwell Telescope located near the summit of Mauna Kea, Hawaii, at elevations above 4000 m (13,100 ft), operates at wavelengths below 1 mm. *See* SUBMILLIMETER ASTRONOMY.

Interferometry and aperture synthesis. The angular resolution, or ability of a radio telescope to distinguish fine detail in the sky, depends on the wavelength of observations divided by the size of the instrument. But even the largest antennas when used at their shortest operating wavelength have an angular

Fig. 2. The 1000-ft (305-m) NAIC fixed spherical dish near Arecibo, Puerto Rico. (*NAIC/NSF-Arecibo Observatory*)

Fig. 3. NRAO 27-element Very Large Array near Socorro, New Mexico, in its most compact configuration. (*NRAO/AUI/NSF*)

resolution only a little better than 1 arcminute, which is comparable to that of the unaided human eye at optical wavelengths. Because radio telescopes operate at much longer wavelengths than optical telescopes, they must be much larger than optical telescopes to achieve the same angular resolution. Fortunately at radio wavelengths, the distortions introduced by the atmosphere are less important than at optical wavelengths, and so the theoretical angular resolution of a radio telescope can be achieved in practice, even for the largest dimensions.

The high angular resolution of radio telescopes is achieved by using the principles of interferometry to synthesize a very large effective aperture from a number of small elements. In a simple two-element radio interferometer, the signals from an unresolved, or point, source alternately arrive in phase and out of phase as the Earth rotates and cause a change in the difference in path from the radio source to the two elements of the interferometer. This produces interference fringes in a manner similar to that in an optical interferometer. If the radio source has finite angular size, the difference in path length to the elements of the interferometer varies across the source. The measured interference fringes from each interferometer pair thus depend on the detailed nature of the radio "brightness" distribution in the sky.

Each interferometer pair measures one Fourier component of the brightness distribution of the radio source. By using multiple antenna elements with a variety of separations and taking advantage of the rotation of the earth, radio arrays can sample a sufficient number of Fourier components to synthesize the effect of a large aperture and thereby reconstruct high-resolution images of the radio sky. Because radio signals can be divided and distributed over large distances without distortion, it is possible to build radio telescope systems of essentially unlimited dimensions. Although the laborious computational task of doing Fourier transforms to obtain images from the large amount of interferometer data

was once formidable, these tasks are now easily accomplished by powerful PCs.

The most powerful radio array of this type is the Very Large Array (VLA) located on the Plains of San Agustin near Socorro, in central New Mexico (**Fig. 3**). The VLA consists of 27 parabolic antennas, each measuring 25 m (82 ft) in diameter. The total collecting area is equivalent to that of a single 130-m (426-ft) antenna. However the angular resolution is equivalent to that of a single antenna 35 km (22 mi) in diameter. Each element of the VLA can be moved along a Y-shaped railroad track; thus, it is possible to change the length of the arms between 600 m (2000 ft) and 21 km (13 mi) to vary the field-of-view and the resolution much like a photographic zoom lens. When used at short, centimeter wavelengths in the largest antenna configuration, the angular resolution of the VLA is better than one-tenth of an arcsecond, or about the same as that of the *Hubble Space Telescope* at optical wavelengths. The VLA is operated by the U.S. National Radio Astronomy Observatory and is used by nearly 1000 astronomers each year for a wide variety of research programs.

In conventional interferometers and arrays, coaxial cable, waveguide, or fiber-optic links are used to distribute a common reference signal to each antenna and also to return the received signal from an individual antenna to a central laboratory where it is correlated with the signals from other antennas. In cases in which antennas are spaced more than a few tens of kilometers apart, however, it becomes prohibitively expensive to employ real physical links to distribute the signals. Very high frequency (VHF) or ultrahigh frequency (UHF) radio links can be used, but the need for a large number of repeater stations makes this impractical for spacings greater than a few hundred kilometers. The Multi-Element Radio-Linked Interferometer Network (MERLIN), operated by the Nuffield Radio Astronomy Laboratories at Jodrell Bank in the UK, uses microwave radio links to connect seven antennas separated by up to 200 km

Fig. 4. Artist's conception of the 10-element Very Long Baseline Array. Signals from each of the 10 antennas are independently recorded on magnetic tape or disks which are then sent to the array headquarters in Socorro, for processing to form radio images with an angular resolution better than 0.001 arcseconds HST. (*NRAO/AUI/NSF*)

(124 mi) in the southern part of England. It is used primarily to study compact radio sources associated with quasars, active galactic nuclei (AGN), and cosmic masers with a resolution of a few hundredths of an arcsecond.

Interferometer systems of essentially unlimited element separation can be formed by using the technique of very long baseline interferometry, or VLBI. In a VLBI system the signals received at each element are recorded by broad-bandwidth tape recorders or large-capacity computer disks. The recorded tapes or disks are then transported to a common location where they are replayed and the signals combined to form interference fringes. A single magnetic tape capable of recording for several hours can contain 10^{12} bits of information, which is roughly equivalent to storing the entire contents of a modest-sized library.

The Very Long Baseline Array (VLBA) consists of ten 25-m (82-ft) dishes spread across the United States from the Virgin Islands to Hawaii **Fig. 4**. The VLBA operates at wavelengths from 3 mm to 1 m and is used to study quasars, galactic nuclei, cosmic masers, pulsars, and radio stars with a resolution as good as 0.0001 arcsecond, or more than 100 times better than that of the *Hubble Space Telescope*. Data tapes or computer disks from the ten individual antenna elements of the VLBA are shipped to a special processing center in New Mexico where they are replayed and the signals analyzed to form images. Precise timing between the elements is maintained by a hydrogen maser atomic clock located at each antenna site.

In 1997, Japanese radio astronomers working at the Institute for Space Science near Tokyo launched an 8-m (26-ft) dish, known as VSOP, in Earth orbit. Working with the VLBA and other ground-based radio telescopes, VSOP gave interferometer baselines up to 20,000 km (12,400 mi). Russian radio astronomers plan a similar space VLBI mission with baselines extending more than 100,000 km (62,000 mi) from the Earth.

Interferometers and arrays are also used at millimeter and submillimeter wavelengths to study the formation of stars and galaxies with resolution better than can be obtained with simple filled-aperture antennas. The operation of arrays at millimeter and submillimeter wavelengths is very difficult and requires siting the instrument at very high and dry locations to minimize the phase distortions of signals as they propagate through the terrestrial atmosphere. Millimeter interferometers and arrays are found at the IRAM Plateau de Bure facility in France, the Japanese Nobeyama Observatory, and near the Owens Valley in California. The Harvard-Smithsonian Center for Astrophysics in collaboration with the Academia Sinica of Taiwan operates the eight-element Submillimeter Array (SMA) near the summit of Mauna Kea at an elevation of 4080 m (13,385 ft). The SMA is designed to work at wavelengths as short as 0.3 mm. A major new international facility under construction in the Atacama Desert in northern Chile at an elevation of more than 5000 m (16,400 ft) is expected to be completed by 2012 and will consist of more than fifty 12-m (39-ft) dishes also operating at wavelengths as short as 0.3 mm. *See* ASTRONOMICAL OBSERVATORY; RADIO ASTRONOMY.   Kenneth Kellermann

Bibliography. B. F. Burke and F. Graham Smith, *An Introduction to Radio Astronomy*, Cambridge University Press, 2d ed., 2002; J. D. Kraus, *Radio Astronomy*, 2d ed., Cygnus-Quasar Books, 1986; K. Rohlfs and T. Wilson, *Tools of Radio Astronomy*, 4th ed., Springer, 2003; G. L. Verschuur, *The Invisible Universe Revealed: The Story of Radio Astronomy*, Springer, 1986; G. L. Verschuur and K. I. Kellermann, *Galactic and Extragalactic Radio Astronomy*, Springer, 1991.

Radio transmitter

A generator of radio-frequency (rf) signals for wireless communication over some distance, which can vary from the short ranges within a building to intercontinental distances. Most applications utilize signals from very low frequencies (VLF) to extremely high frequencies (EHF); some applications require frequencies as low as 45 Hz or as high as 100 GHz. The radio-frequency output power varies from a fraction of a watt in emergency beacons and portable equipment to several megawatts in long-range, low-frequency transmitters. *See* RADIO SPECTRUM ALLOCATIONS.

The architecture (organization) of a radio transmitter is determined by the type of signal it is intended to produce. The four basic architectures are those used for continuous-wave, frequency-modulation, amplitude-modulation, and single-sideband signals. Transmitters for some applications (for example, television) use a combination of these architectures
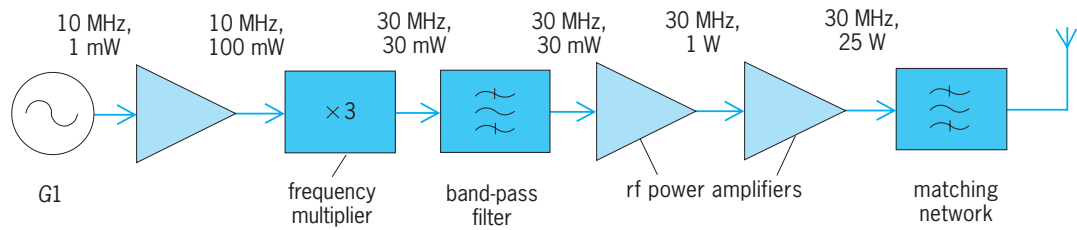
Fig. 1. Basic continuous-wave (CW) transmitter.

(for example, frequency modulation for sound and single sideband for video), while others (for example, Loran C) use unique architectures. Alternative architectures such as envelope elimination and restoration or outphasing can be used to improve efficiency. *See* ELECTRICAL COMMUNICATIONS; MODULATION.

Continuous-wave (CW) transmitter. The most basic type of radio transmitter produces only a continuous-wave signal. Such transmitters are often switched on and off (keyed) to produce telegraph signals. *See* TELEGRAPHY.

CW architecture. The block diagram of a simple continuous-wave transmitter is shown in **Fig. 1**. The oscillator *G*1 produces a low-power signal, which is boosted to the final output power by a series of progressively larger power amplifiers. The optional inclusion of a frequency multiplier improves stability by allowing the frequencies of the oscillator and high-power amplifiers to be different. *See* FREQUENCY MULTIPLIER; OSCILLATOR; POWER AMPLIFIER.

The architecture shown in **Fig. 2** includes both frequency translation and power splitting, which makes it more suitable for generating high-power signals at various frequencies. While at a relatively low level, the signal is translated by a mixer to the desired output frequency. After amplification by a chain of power amplifiers, it is split into two parts to drive two final power amplifiers whose outputs are combined to produce the transmitter output. *See* MIXER.

Typically, each stage of amplification increases the signal amplitude by 10–20 decibels. Since the signal has a constant amplitude, class C power amplifiers are generally used to provide good efficiency in dc-to-radio-frequency power conversion. For higher efficiency, power amplifiers of classes D, E, or F can be used. *See* AMPLIFIER.

Since the applications for continuous-wave transmitters generally require operation over only a small bandwidth, tuned matching networks are generally employed. These networks provide the final power amplifier with the load impedance required to produce the desired power output, given the available power supply voltage. *See* IMPEDANCE MATCHING.

Keying. Keying for telegraphy is generally produced by an analog switch inserted between two low-level stages. Alternatively, the bias points of one or more power amplifiers can be shifted to put their transistors into a cut-off condition. A shaping filter is often employed to produce the desired pulse shape.

Frequency-modulation (FM) transmitter. Analog frequency modulation is widely used for voice communication, high-quality audio broadcasting, and television audio. Frequency-shift keying (FSK) and phase-shift keying (PSK) are widely used for transmission of digital data via radio-frequency signals. *See* FREQUENCY MODULATION; FREQUENCY-MODULATION RADIO; MOBILE RADIO; RADIO BROADCASTING; TELEVISION.

FM architecture. Frequency-modulated and phase-modulated (PM) signals have constant amplitudes and are therefore produced by transmitters with architectures similar to those of the continuous-wave transmitter (Figs. 1 and fig. 2). The principal change is the replacement of oscillator *G*1 by a frequency or phase modulator. In frequency-modulation transmitters, the frequency multiplier increases the frequency deviation as well as the carrier frequency of the frequency-modulated signal. *See* PHASE MODULATION.

Frequency modulators. In communication applications, the frequency modulator is typically a voltage-controlled crystal oscillator (VCXO) in which the
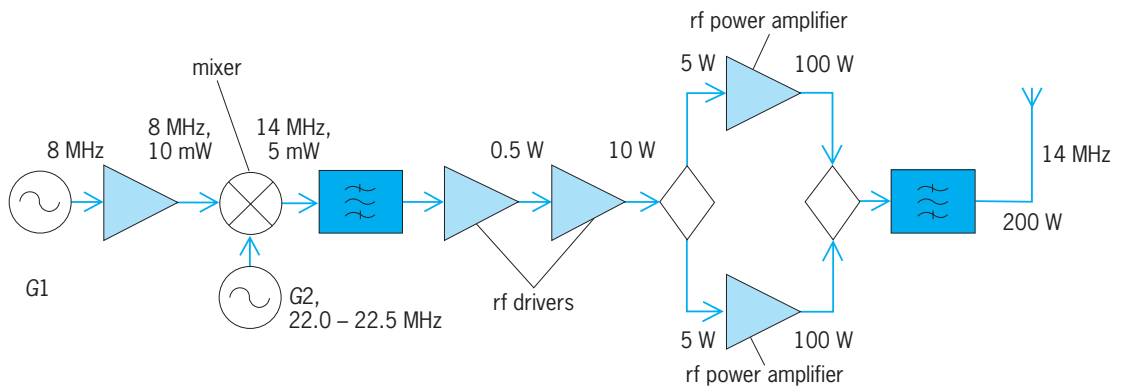

Fig. 2. Continuous-wave (CW) transmitter with frequency conversion and power splitting.

capacitance of a varactor diode is used to vary slightly the frequency of a crystal oscillator. Other applications employ various types of modulators, including phase-shift, phase-locked-loop, comparator, and Armstrong. *See* FREQUENCY MODULATOR.

Frequency-shift-keying signals are generally produced by selecting the output of one of two oscillators according to the current data bit. Binary phase-shift keying is typically produced by using a double-balanced mixer to invert the polarity of the carrier. Other forms of phase-shift keying generally employ digital generation (and selection) of the several phases.

Amplitude-modulation transmitter. Full-carrier amplitude modulation is used in medium-frequency (MF) broadcasting, high-frequency (HF) international broadcasting, citizen-band communication, aircraft communication, and nondirectional navigation beacons. *See* AMPLITUDE MODULATION; ELECTRONIC NAVIGATION SYSTEMS.

AM architecture. Most modern full-carrier amplitude-modulation transmitters produce the output signal by amplitude modulation of the final radio-frequency power amplifier. Generally, the modulation is accomplished by varying the supply voltage of the radio-frequency power amplifier with a high-power radio-frequency amplifier. Since the radio-frequency carrier has constant amplitude until the final power amplifier, the architecture of the radio-frequency chain (**Fig. 3**) is similar to that of a continuous-wave or frequency-modulation transmitter. *See* AUDIO AMPLIFIER.

Amplitude modulators. Class B audio-frequency power amplifiers are sometimes used in low-power transmitters. However, modern high-power transmitters use either a class S audio-frequency power amplifier or pulse-step modulation for high efficiency. The input to the audio-frequency power amplifier is a sum of audio-frequency components (sound) and dc (carrier level).

Grid-bias modulation is sometimes employed in inexpensive, low-power amplitude-modulation transmitters. In this technique, the audio-frequency signal and a bias voltage are fed to the gate of a final class C radio-frequency power amplifier that uses field-effect transistors (FETs). The variation of bias causes variation of the conduction angle, hence variation of the amplitude of the output. While this method is not as linear as series modulation, it eliminates the need for a high-power audio amplifier.

Transformer coupling of high-power audio signals into the supply voltage of the final radio-frequency power amplifier was often employed in older transmitters. However, this technique is now little used because of the size and weight of the transformer. It is, of course, possible to produce full-carrier amplitude-modulated signals with linear radio-frequency power amplifiers, as in the single-sideband transmitter, discussed below. *See* AMPLITUDE MODULATOR.

Single-sideband (SSB) transmitter. Single-sideband amplitude modulation is widely used for high-frequency voice communications, including military, marine, aeronautical, diplomatic, and amateur. It also finds use (as amplitude-compandored single-sideband, or ACSB) at very high frequencies (VHF) and ultrahigh frequencies (UHF).

Although single sideband is technically a form of amplitude modulation, the single-sideband signal itself has variations of both amplitude and phase. Signals such as multitone, independent sideband (ISB), and vestigial sideband (VSB), used for video, also possess such characteristics. Consequently, these signals are traditionally amplified by a chain of linear radio-frequency power amplifiers operating in class B. *See* SINGLE SIDEBAND.

SSB architecture. The architecture of a modern, broad-band single-sideband transmitter is shown in **Fig. 4**. The low-level output of the single-sideband modulator is first shifted by the local oscillator $G2$ to an intermediate frequency (i-f) that is at least twice the highest output frequency. The intermediate-frequency signal is then shifted downward to the desired output frequency by the variable-frequency oscillator (VFO) $G3$. The mixer output is low-pass-filtered and then amplified to the desired power. *See* ELECTRIC FILTER.

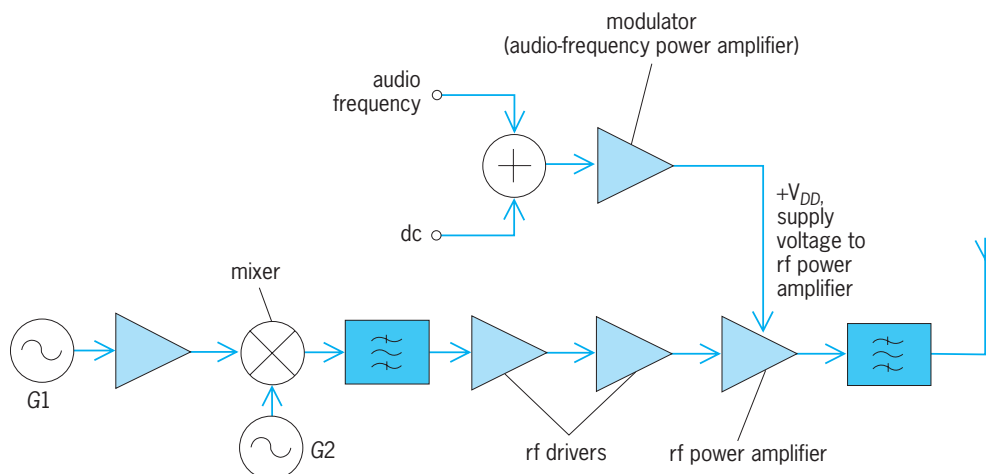This high-side-mixing architecture creates only

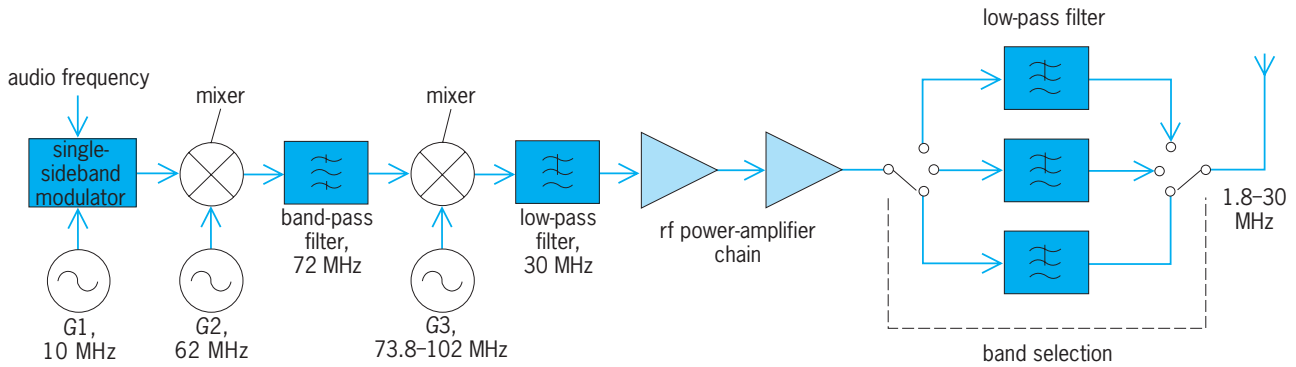**Fig. 3. Amplitude-modulation (AM) transmitter.**

Fig. 4. Single-sideband (SSB) transmitter.

modest demands on the synthesizer used to implement the variable-frequency oscillator $G3$. Further, no first-order products in the final mixer output fall within the output band. Consequently, the entire chain of radio-frequency power amplifiers is untuned and can amplify signals without adjustment for any frequency in the intended range of operation. It is necessary only to switch-select an appropriate output filter to ensure adequate suppression of harmonics. Such filters are generally based upon elliptic designs, and each covers approximately 1.5 octaves of frequency.

Older single-sideband transmitters designed for operation on only a limited number of frequency bands (for example, amateur) typically vary the frequency of oscillator $G1$ for tuning within a given band and select one of several different frequencies for oscillator $G2$ to produce an output in the desired band. In such transmitters, each stage of amplification is generally tuned to the current band of operation.

Direct-conversion architecture is used in some single-sideband transmitters. Such transmitters are analogous to high-power phasing-type single-sideband modulators. While this approach offers some simplicity, a very high degree of amplitude and phase matching is required to achieve good rejection of the unwanted sideband.

SSB modulators. The traditional method for generating high-quality single-sideband signals first provides a double-sideband/suppressed-carrier amplitude-modulation (DSB/SC-AM) signal in a double-balanced mixer. The unwanted sideband is then removed by a mechanical or crystal filter.

Single-sideband signals can also be generated by the phasing technique. The audio signal is applied to an all-pass filter, which produces two outputs that differ in phase by 90°. The two audio-frequency signals then modulate two radio-frequency carriers that differ in phase by 90°. The addition or subtraction of the two double-sideband/suppressed-carrier outputs cancels the unwanted sideband.

The Weaver (third-method) single-sideband modulator is sometimes used, especially when the filtering and modulation processes are accomplished digitally. This technique eliminates the need for audio-frequency all-pass filters by modulating the input audio signal onto two midband carriers (for example, 1.6 kHz) that are in phase quadrature. The outputs of these mixers are then low-pass-filtered (for example, at 1.5 kHz) to remove the carrier and unwanted sideband. The two low-pass-filtered signals modulate in-phase and phase-quadrature carriers. The outputs of these two mixers are then summed to produce the described sideband, as in the phasing method.

Techniques for high efficiency. High-efficiency transmitters are based upon a combination of high-efficiency radio-frequency power amplifiers, modulation techniques, and high-efficiency modulators. These techniques not only increase the efficiency when the output is near the peak output power but also keep it relatively constant at all signal levels. Consequently, the average efficiency can be several times that of a conventional power amplifier for signals with large peak-to-average ratios, such as speech and multiple tones. The increased efficiency reduces the power costs for broadcast transmitters and reduces size, weight, and power supply requirements for portable, remote-relay, and satellite-borne transmitters.

Power amplifiers. Conventional power amplifiers (classes A, B, and C) use their active devices (tubes, bipolar junction transistors, or field-effect transistors) as variable resistors and therefore inherently dissipate a portion of the dc input power as heat. High-efficiency power amplifiers (classes D, E, F, and S) use their active devices as switches or employ wave-shaping circuits, allowing more of the dc input power to be converted into radio-frequency output. *See* WAVE-SHAPING CIRCUITS.

A class D power amplifier employs a pair of transistors driven to act as a switch. An output filter allows only the fundamental-frequency component of the resultant square wave to reach the load. The efficiency is ideally 100% and does not inherently decrease with mismatched loads. Class D power amplifiers are now practical through high frequencies and the lower range of very high frequencies at power levels of 1 kW or more.

A class E power amplifier employs a single transistor driven to act as a switch. The output filter is selected so that it brings the drain voltage to zero at
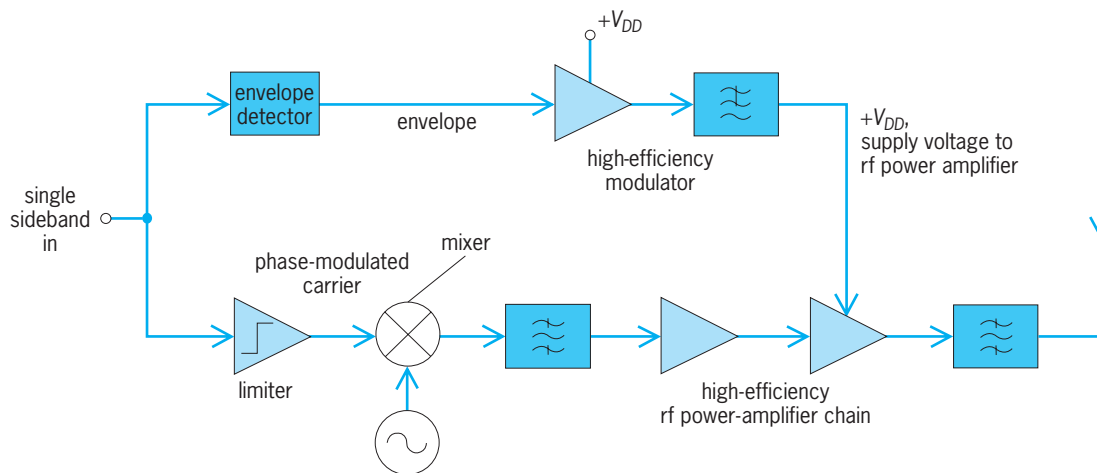
Fig. 5. Envelope-elimination-and-restoration (EER) transmitter.

the instant the transistor is switched on. The power losses due to charging the drain capacitance are eliminated, and the efficiency is ideally 100%. As a result, class E power amplifiers are attractive at very high frequencies and ultrahigh frequencies.

A class F power amplifier employs a single transistor and a multiple-resonator output circuit, preferably implemented from transmission lines. The multiple resonances facilitate the formation of approximate square waves, thus reducing the dissipation of the amplifier. Class F power amplifiers can be implemented at ultrahigh frequencies and microwave frequencies where switching-mode operation of the transistors is not possible.

High-efficiency modulators. Both class S amplification and pulse-step modulation (PSM) are currently in use for high-efficiency modulation. Both ideally achieve 100% efficiency.

Class S modulators are based upon pulse-width modulation (PWM) with a switching frequency several times the highest output frequency. The pulse-width-modulated signal is boosted to the desired power level by switching amplifiers, after which the desired audio output is obtained by a low-pass filter. *See* PULSE MODULATION.

A pulse-step modulator is basically a high-power digital-to-analog converter that selects the one power supply voltage of several (8–12) that is closest to the desired modulating voltage at any given time. *See* DIGITAL-TO-ANALOG CONVERTER.

Envelope elimination and restoration. Any narrow-band signal can be regarded as having simultaneous amplitude (envelope) and phase modulation. The envelope-elimination-and-restoration (EER) technique uses this principle to facilitate high-efficiency linear radio-frequency amplification.

An EER system (**Fig. 5**) first detects the envelope and hard-limits the carrier. The hard-limited carrier is a phase-modulated signal suitable for amplification by efficient but nonlinear radio-frequency power amplifiers (for example, classes C, D, E, and F). The envelope consists of audio-frequency and dc components and is therefore amplified efficiently by a class

S modulator or power supply modulator. Amplitude modulation of the final radio-frequency power amplifier restores the envelope, producing the desired high-power single-sideband signal.

In some applications (for example, AM stereo), the envelope and phase can be produced easily by digital signal processing, thus simplifying the EER system. *See* STEREOPHONIC RADIO TRANSMISSION.

Other techniques. Digital power addition is used to produce amplitude-modulated signals at low frequencies (lf) and medium frequencies. Power amplifiers of various outputs are connected in series and toggled on or off to produce the desired instantaneous output voltage.

In outphasing, the outputs of two power amplifiers are combined. However, the power amplifiers are driven with signals of different phases. The phase difference is varied dynamically so that the vector sum of the two outputs has the desired amplitude.

The Doherty technique combines the outputs of two power amplifiers via a quarter-wavelength line. At intermediate signal levels, one power amplifier operates near its peak output power and is therefore at its maximum efficiency. The second power amplifier is used to produce peak system output power.

Transmitters for the pulsed-continuous-wave signal used in Loran C radio navigation use a half-cycle technique. Sequential discharge of several inductance-capacitance (*L-C*) networks into the antenna efficiently produces the desired high-power signal. *See* LORAN; RADIO.   Frederick H. Raab

Bibliography. *The ARRL Handbook for the Radio Amateur*, annually; H. L. Krauss, C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, 2d ed., 2002; *Proceedings of RF Expo East '87*, Boston, November 11–13, 1987; *Proceedings of RF Technology Expo '88*, Anaheim, California, February 10–12, 1988; W. E. Sabin and E. O.Schoenike (eds.), *Single-Sideband Systems and Circuits*, 2d ed., 1995; V. Shakhgildyan, *Radio Transmitter Design*, 1987; J. R. Smith, *Modern Communication Circuits*, 2d ed., 1997.

Radio-wave propagation

The means by which radio signals are transported through space from a transmitting antenna to a receiving antenna. Radio signals are electromagnetic waves which travel with the velocity of light and can be reflected, refracted, diffracted, scattered, and absorbed. Unlike visible light, radio frequencies cover many octaves, from about 10 kHz to 60,000 MHz (wavelengths from 30,000 m to 0.5 cm). Since frequency is an important parameter, radio propagation characteristics vary over a wide range. At the higher radio frequencies the similarity with visible light is evident. At the lower frequencies the radio waves follow the surface of the Earth by a mechanism that in geometrical optics is unimportant and relatively unknown.

The power radiated from a transmitting antenna is ordinarily spread over a relatively wide area. As a result, the power available at most receiving antennas is only a very small fraction (10^{-8} to 10^{-16} or less) of the radiated power. The principal function of radio-wave propagation analysis is to make possible estimation of the expected received power in order to predict the usefulness of a radio signal at any location remote from the transmitter. For a discussion of antenna radiation *see* ANTENNA (ELECTROMAGNETISM); ELECTROMAGNETIC RADIATION.
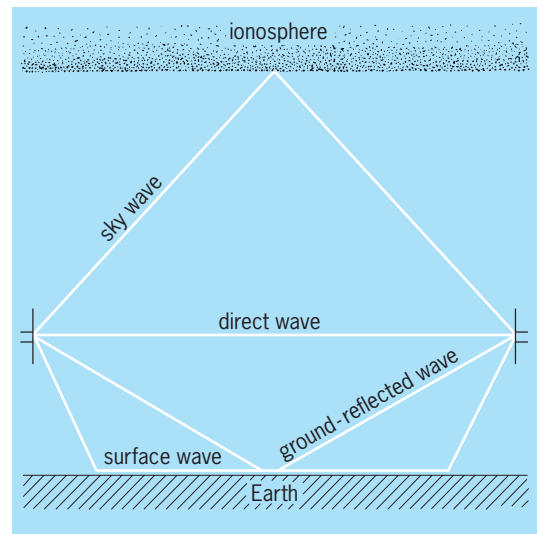
Typical distances that can be achieved with usual

Fig. 2. Possible transmission paths between antennas.

types of equipment are shown in **Fig. 1**. The frequencies around 20 kHz can be received reliably at distances of thousands of miles but are limited to telegraph-type signals and require very large transmitting antennas. Higher frequencies are needed for voice, and still higher frequencies for television transmission. As the frequency increases, the transmission range tends to decrease. Frequencies above 100 MHz can transmit wideband signals, but they are limited to approximately line-of-sight distances with the usual type of equipment. However, distances of 200 mi (320 km) or more are possible by the use of high power and large antennas to provide narrow "searchlight" beams.

Reflections from the ionosphere (ionized layers 50–250 mi or 80–400 km above the Earth's surface) provide a useful but variable long-distance service at frequencies less than about 30 MHz. These reflections account for the long-range broadcast coverage at night and for the shortwave intercontinental communication. *See* IONOSPHERE.

The principal components of the received radio signal are shown symbolically in **Fig. 2**. The vector sum of the direct, reflected, and surface waves has been called the space wave, ground wave, or tropospheric transmission to differentiate it from the ionospheric reflections. The ionospheric and surface waves are the principal components at frequencies below 10–30 MHz. The direct and reflected rays are the principal factors at frequencies above 30–50 MHz. Although the ionospheric, direct, and ground-reflected waves can be easily visualized as rays, the surface wave is more difficult to understand; it originates at the air-Earth boundary because the Earth is not a perfect reflector.

Line-of-sight transmission. This is the first and simplest concept in radio-wave propagation. Radio transmission in free space results in a decrease in energy per unit area in accordance with the inverse-square law. The ratio of the transmitter power to the received power is called the radio transmission loss;
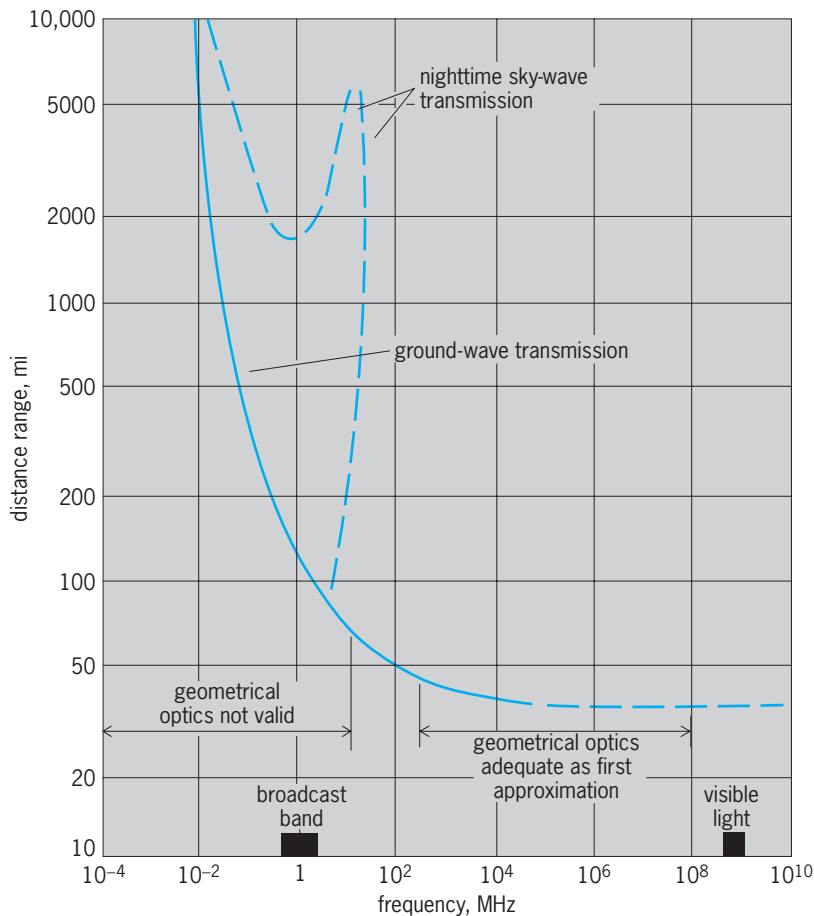

Fig. 1. Transmission range of electromagnetic waves. 1 mi = 1.6 km.

for nondirectional antennas in free space it is $P_T/P_O = (4\pi d/\lambda)^2$, where the distance d and wavelength λ are expressed in the same units. Here P_T is the transmitted power, and P_O is the power received under free-space conditions. The wavelength $\lambda = c/f$, where c is the velocity of light, and f is the frequency in hertz (cycles per second). *See* INVERSE-SQUARE LAW.

Ground effects. The presence of the ground may affect the antenna impedance and efficiency, and reflections from the ground tend to interfere with the free-space wave. The received power near the ground is ordinarily less than would be expected in free space. At frequencies in the amplitude-modulation (AM) broadcast band and lower, transmission to moderate distances is primarily by the surface wave because the direct and reflected waves tend to cancel. For the surface wave the received power P_R is less than the free-space value by the fraction shown as relation (1). The factor K is greater for

$$\frac{P_R}{P_0} \approx \left| K \frac{\lambda}{d} \right|^2 \qquad (1)$$

vertical polarization of the waves than for horizontal polarization and is greater for vertically polarized waves over seawater than for transmission over dry land.

At frequencies in the television band and higher, both the transmitting antenna and the receiving antenna are usually several wavelengths above the ground. In this case the surface wave and its dependence on polarization becomes less important than the direct and reflected waves. Reflections from the ground may either cancel or reinforce the direct wave, and the power P_R received over flat ground is given approximately by Eq. (2). Here h_1 and h_2 are

$$\frac{P_R}{P_0} = \left| 2 \sin \frac{2\pi h_1 h_2}{\lambda d} \right|^2 \qquad (2)$$

the antenna heights in the same units as the wavelength and distance.

At still higher frequencies, that is, in the microwave region, irregularities in the surface of the Earth are frequently large compared to wavelength, and the magnitude of the reflection coefficient becomes substantially less than unity. *See* REFLECTION AND TRANSMISSION COEFFICIENTS.

As the reflected wave is weakened, radio propagation approaches free-space transmission, as long as there is adequate clearance over all intervening obstacles. In free-space transmission between two directive antennas the received power is related to the transmitted power by Eq. (3). Here A_T and A_R are

$$\frac{P_R}{P_T} = \frac{A_T A_R}{(\lambda d)^2} \qquad (3)$$

the effective aperture areas of the transmitting and receiving antennas, respectively, measured in the same units as the wavelength and distance.

Fading. Variations in signal level with time are caused by changing atmospheric conditions. The severity of the fading usually increases as either the frequency or path length increases. The path of a radio wave is not a straight line, except for the ideal case of a uniform atmosphere. The transmission path may be refracted down or up, depending on atmospheric conditions. This bending may either increase or decrease the effective path clearance, and unfavorable (upward) bending may have the effect of transforming a line-of-sight path into an obstructed one. This type of fading may last for several hours at a time. The frequency of its occurrence and its depth can be reduced by increasing the path clearance, particularly in the middle of the path.

Most of the fading that occurs on "rough" paths with adequate clearance is the result of interference between two or more rays traveling slightly different routes in the atmosphere. This fast multipath type of fading is relatively independent of path clearance.

Most fading is a temporary diversion of energy to some direction other than the intended location, but absorption effects are important in the microwave region. At frequencies above 5000–10,000 MHz the presence of rain, snow, or fog introduces an absorption in the atmosphere that depends upon the amount of moisture, the particle size, and the frequency. At frequencies higher than 15,000–20,000 MHz the additional attenuation caused by heavy rain tends to limit the path length to only a few miles, if high reliability is required. In addition to the effect of rain, some selective absorption may result from the oxygen and water vapor in the atmosphere. The first absorption peak due to water vapor occurs at about 23,000 MHz, and the first absorption peak for oxygen occurs at about 60,000 MHz. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION.

Radar transmission. The received radar echo is very weak because only a small part of the transmitted energy reaches the target, and only a small fraction of the energy reflected from the target is returned to the radar receiver. The transmission loss under free-space conditions is given by Eq. (4), where A'_R and A'_T pertain to the reflecting properties

$$\frac{P_R}{P_T} = \frac{A_T A'_R}{(\lambda d)^2} \frac{A'_T A_R}{(\lambda d)^2} \qquad (4)$$

of the target. When accurate information on the target characteristics is unavailable, it is usually assumed that A'_R is the geometrical cross section of the target and that $A'_T = \lambda^2/4\pi$ (uniform radiation in all directions).

Outer-space vehicles. Radio transmission to Earth satellites and other outer-space vehicles is an extension of line-of-sight transmission to long-distance, free-space conditions. Each tenfold increase in distance requires a hundredfold increase in transmitter power, when the antenna size and other parameters remain constant. The assigned frequency needs to be high enough to penetrate the ionosphere and to provide adequate directivity and bandwidth; an upper limit is set by the limitations of rain attenuation and atmospheric absorption. *See* SPACE COMMUNICATIONS.

Passive satellites, such as *Echo 1* or *Echo 2*, are like radar targets or mirrors in that the transmitted signal is reflected but not amplified. The resulting weak echo limits their information capacity to much less than is possible with active satellites.

Active satellites, such as the *Intelsat* and *Molniya* communication satellites or the lunar and interplanetary probes, contain radio receivers and radio transmitters with the necessary power supply on board. Since the transmission loss before amplification pertains to a one-way rather than a round-trip path, the resulting signal is strong enough either to carry more information or to go farther than is possible with a passive satellite. *See* COMMUNICATIONS SATELLITE.

Propagation beyond optical horizon. Radio transmission beyond the line of sight can be achieved by three principal methods: refraction, diffraction, and reflection. When these effects are intermixed and cannot be separated easily, the energy is said to be scattered.

Refraction. The dielectric constant of the atmosphere normally decreases gradually with increasing altitude. The result is that on the average the radio ray is bent or refracted toward the Earth so that the distance to the radio horizon is slightly greater than to the optical horizon. The amount of refraction is variable, and exceptionally long-range transmission may occur occasionally. The corresponding phenomenon in optics makes it possible to see lights or other objects that are normally below the horizon. Conversely, when the radio energy is bent away from the Earth (upward bending), the transmission loss is increased. *See* REFRACTION OF WAVES.

Diffraction. Radio waves are also transmitted around the Earth by the phenomenon of diffraction. When light waves are partially blocked by an obstacle, the transition from light to dark at the edge of a shadow is gradual rather than infinitely sharp. The amount of energy diffracted around an obstruction decreases as the frequency is increased. Typical obstructions include hills, trees, and buildings, as well as the curvature of the Earth. *See* DIFFRACTION.

Reflection and scatter. Most of the experimental data at points far beyond the horizon are intermediate between the values expected for diffraction over a smooth sphere and for diffraction over a ridgelike obstruction. Various theories have been advanced to explain these effects. The explanation most commonly accepted is that energy is reflected or scattered from atmospheric layers and turbulent air masses. At points far beyond the horizon the median received power, relative to free-space transmission, decreases as the first power of the frequency and as the fifth or sixth power of distance. Beyond 300–400 mi (480–640 km) the decrease in signal changes gradually from a power law to an exponential decrease of about $e^{-0.02d}$, where d is measured in miles. Although useful signals can be obtained at distances of 200 mi (320 km) or more at all frequencies up to the 5000–10,000-MHz limit set by rain attenuation, the optimum frequency range is below 1000 MHz for most applications. Beyond-horizon or tropospheric-scatter circuits require very high power and large antennas; they are, however, economically feasible in isolated areas and for overwater paths. Rapid fading occurs nearly all the time, but with ample fading margin and the use of diversity reception, high quality and high reliability can be obtained. It is possible to transmit 100 or more voice channels on a single radio carrier. It is also possible to provide an acceptable grade of monochrome television on a single link of approximately 150–200 mi (240–320 km). *See* TROPOSPHERIC SCATTER.

Ionospheric propagation. Ionospheric reflections return to Earth useful radio energy at frequencies up to 25 MHz, and possibly up to 60 MHz. Information about the nature of the ionosphere has been obtained mainly by transmitting radar-type signals directly overhead and recording the intensity and time delay of the echoes returned from the ionized layers. At night all frequencies below the critical frequency of 2–4 MHz (for vertical incidence) are returned to Earth with a received power that is close to the value that would be expected in free space for the round-trip distance. During the day the critical frequency is two to three times greater than the corresponding night value. This apparent increase in useful frequency range is largely offset by the strong daytime ionospheric absorption, which reaches a maximum in the 1–2-MHz range at about noon. The difference between day and night transmission means that most sky-wave circuits require two or more frequencies for reliable service covering all hours.

For oblique incidence the maximum usable frequency (MUF) is greater; however, even at the longer distances it does not exceed 3–3.5 times the critical frequency for vertical incidence at that time, season, and latitude.

The frequencies most suitable for transmission over distances of 1000 mi (1600 km) or more will ordinarily not be reflected at the high angles needed for much shorter distances. As a result, the range of sky-wave transmission usually does not overlap the range of ground-wave transmission. The intermediate region of very weak or undetectable signals is called the skip zone. At frequencies around 1–2 MHz the ground wave and sky wave may overlap, resulting in severe fading in the region where the two signals are comparable in amplitude.

In addition to the diurnal variations, there are systematic changes with season, latitude, and the 11-year sunspot cycle. For example, on a summer night in middle latitudes during a year of low sunspot activity, the MUF for the longest distance may be limited to less than 7–8 MHz. On the other hand, during a year of high sunspot activity, the corresponding MUF for a winter afternoon may be 40 MHz or higher.

In addition to the normal daytime absorption, there is a second type of absorption which can occur either day or night and is particularly troublesome on transmission paths that travel near or through the polar regions. During periods of magnetic storms the auroral zones expand over an area much larger than normal, thereby disrupting ionospheric communications by introducing unexpected absorption. These conditions of poor transmission can last for hours and sometimes for days.

In addition to the regular ionospheric reflections, strong signals sometimes occur at frequencies as high as 60–70 MHz from the E layer, which is a region of relatively high ionization located about 50–70 mi (80–110 km) above the Earth's surface. These reflections are known as sporadic E signals because they are erratic in both time and space.

All ionospheric circuits are subject to rapid multipath fading with echo delays up to several milliseconds. These delays are about 10,000 times as long for tropospheric transmission. As a result of these relatively long delays, uncorrelated selective fading can occur within an interval of a few hundred hertz, and this effect produces the well-known distortion on voice circuits characteristic of shortwave transmission.

Scatter. As the frequency is increased above that normally reflected from the ionosphere, the signal intensity decreases rapidly, but it does not drop out completely. Although the signal intensity is low, reliable transmission can be obtained at frequencies up to 50 MHz or higher and to distances up to at least 1200–1500 mi (1900–2400 km). Such ionospheric-scatter circuits require much higher power and larger antennas than are ordinarily used in ionospheric transmission. Ionospheric-scatter transmission is suitable for continuous transmission of a few telegraph channels or for one-voice circuit, but the useful bandwidth is limited by the severe selective fading that is characteristic of all ionospheric transmission. Ionospheric scatter is apparently the result of reflections from many patches of ionization which occur in the E region.

Momentary peaks of the ionospheric-scatter signal intensity occur every few seconds and seem to be the result of the relatively strong ionization produced by the passage and disintegration of small meteors. Meteor-burst communication systems transmit information only during the peaks and are inoperative when the signal intensity is significantly less than its maximum value. The required transmitter power is much less than for ionospheric-scatter systems, but the intermittent meteor-burst method is limited to discontinuous telegraph-type service of relatively low information capacity.

Radiation belts. Beyond the E and F ionospheric layers are the Van Allen radiation belts and other regions of ionization and magnetic activity, which are not yet clearly understood. These regions extend to distances of several earth radii and, although they have a negligible effect on terrestrial communications, they may affect radio communication to outer space. The associated ionization and magnetic activity are weaker and more variable with time and space than is the case for the better-understood ionospheric layers.

All these effects are influenced strongly by particles and radiation from the Sun. A continuing exploration and measurement of these variations should improve understanding of the solar system. *See* VAN ALLEN RADIATION.

Underground, underwater, and jungles. Radio waves can be transmitted hundreds of feet underground or through jungles and a few tens of feet to submarines under the sea. When the usual radiation component traveling along a straight line between transmitter and receiver is absorbed in the lossy medium, an alternate "up, over, and down" path becomes dominant. This transmission path consists of three parts: (1) from the transmitter up through the absorbing medium to the open air, (2) along and above the Earth's surface to a point over the receiver, and (3) from the surface down through the absorbing medium to the receiving antenna. Under these conditions unconventional antennas and antenna orientations are needed to match the impedance and the direction of the path that provides the strongest signal.

Noise. The usefulness of a radio signal is limited by noise. The noise may be either unwanted external interference or noise originating in the receiver itself. Atmospheric static is caused by lightning or other natural electrical disturbances and is propagated over the Earth by ionospheric transmission. It is generally higher at night than in the daytime and is higher in the warm tropical areas, where storms are frequent, than in the colder northern regions. Atmospheric static is ordinarily predominant at frequencies below a few megahertz, whereas set noise is the primary limitation at frequencies above 200–500 MHz. In the intermediate region the controlling noise depends on location and time of day and may come from human sources, such as the operation of electric switches or automobile-ignition devices.

The very low frequency components of lightning discharges are propagated along lines of magnetic force to the antipodes, giving rise to a phenomenon known as whistlers. Cosmic and solar noise are of considerable interest in the fields of radio astronomy and satellite transmission but ordinarily are not the controlling factors in terrestrial communication. *See* ELECTRICAL NOISE; MICROWAVE; RADIO ASTRONOMY; RADIO BROADCASTING; SCATTERING OF ELECTROMAGNETIC RADIATION; SFERICS; WAVEGUIDE.   Kenneth Bullington

Bibliography. L. Boithias, *Radiowave Propagation*, 1988; K. G. Budden, *The Propagation of Radio Waves: The Theory of Radio Waves of Low Power in the Ionosphere and Magnetosphere*, 1985, paper 1988; J. Griffiths, *Radiowave Propagation and Antennas: An Introduction*, 1987; T. S. Maclean and Z. Wu, *Radiowave Propagation over Ground*, 1993; K. Rawer, *Wave Propagation in the Ionosphere*, 1993; S. Shibuya, *A Basic Atlas of Radio-Wave Propagation*, 1987.

Radioactive beams

Beams of radioactive (unstable) nuclei. In several nuclear physics laboratories, a capability exists to produce such beams and, before these nuclei spontaneously decay, use them to gain insight into the reactions on and structure of nuclei never before accessible. Radioactive beams are particularly useful to study stellar explosions such as novae, supernovae,

and x-ray bursts. These explosions are some of the most catastrophic events in the universe, generating enormous amounts of energy while synthesizing the elements that make up lifeforms and the world. These spectacular explosions involve, and in some cases are driven by, reactions where the atomic nuclei of hydrogen (protons) and helium (alpha particles) fuse with (are captured by) radioactive isotopes of heavier elements to form new elements. The capability to produce beams of radioactive nuclei allows direct measurements of these reactions, providing crucial information needed to theoretically model cataclysmic stellar events and to understand the origin of many chemical elements.

Nuclear reactions in stars. Stars are born when gas and dust in space condenses under its own gravity to a large body with a central density and temperature sufficiently high to ignite fusion reactions between atomic nuclei. The carbon-nitrogen-oxygen (CNO) cycle is a catalytic sequence of nuclear reactions that generates energy, making massive stars shine and balancing the crushing inward pressure of gravity. Carbon, nitrogen, and oxygen "seed" nuclei capture four protons (the fuel) and eject an alpha particle, liberating energy in the process (**Fig. 1***a*). Typical tempera-
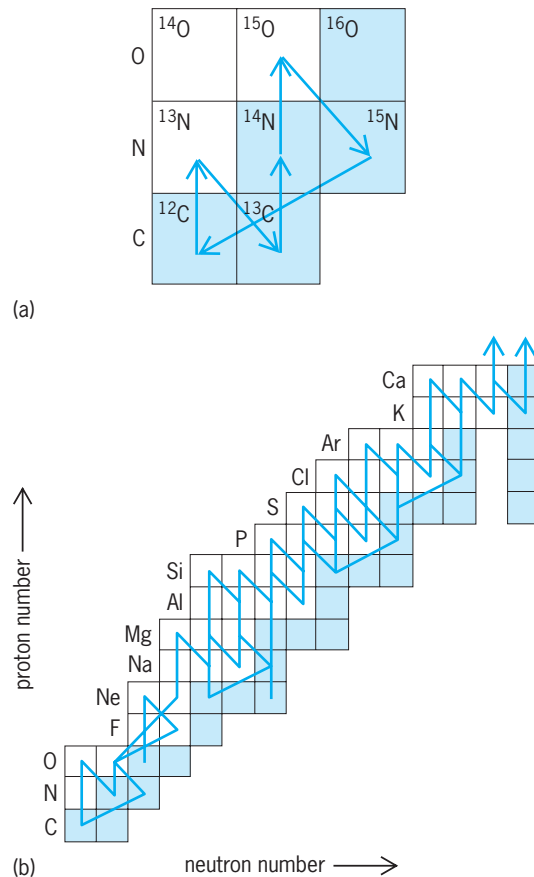

(a)

(b)

Fig. 1. Sequences of nuclear reactions occurring in stars. The shaded boxes represent individual stable nuclei, the white boxes radioactive nuclei. The arrows represent nuclear reactions leading to transforming one nucleus to another. (*a*) Carbon-nitrogen-oxygen (CNO) cycle, occurring in nonexploding stars. (*b*) Reactions on radioactive isotopes occurring in stellar explosions.

tures and densities at the core of a massive star where these nuclear reactions occur may be approximately 10^7 kelvins and 100 grams per cubic centimeter, respectively. Under these conditions, the rates of such nuclear fusion reactions (which have been measured in the laboratory) are extremely slow. Some CNO cycle reactions transform catalytic seed nuclei into radioactive isotopes of different elements, which then spontaneously decay to a stable nucleus in typically a few to a few hundred seconds, well before they will undergo any additional fusion reactions. *See* CARBON-NITROGEN-OXYGEN CYCLES; NUCLEOSYNTHESIS; STELLAR EVOLUTION.

Stellar explosions. During stellar explosions such as novae or x-ray bursts, a different sequence of nuclear reactions occurs. Nova explosions are thought to occur in binary star systems (two stars orbiting each other) when one star of the pair—a cold compact white dwarf star—is able to gravitationally pull material (mainly hydrogen) off the surface of its close binary companion in a phenomenon known as accretion. This situation is not uncommon: there are approximately 40 novae in the Milky-Way Galaxy each year. After approximately 10,000 years, enough hydrogen has collected, or accreted, on the white dwarf surface to raise the temperatures and densities to 10^8 kelvins and 100–10,000 grams per cubic centimeter, respectively. These conditions are sufficient to ignite nuclear fusion reactions between the accreted hydrogen and the carbon and oxygen (or in some cases oxygen, neon, and magnesium) heavy nuclei which make up the white dwarf. The fusion reactions generate energy, which raises the temperature of the accreted layer, which exponentially increases the rates of all the fusion reactions. This becomes a runaway thermonuclear explosion where the temperatures may reach 3–4×10^8 kelvins and the energy generated by the nuclear reactions causes ejection of a sizable fraction of the accreted layer off the surface of the white dwarf. The energy released in these enormous thermonuclear explosions can be up to 10^{31}–10^{38} joules, in a thousand seconds—10^{24} times the energy released in powerful thermonuclear explosions on Earth.

Since nuclear fusion reaction rates increase exponentially with increasing temperature, they occur approximately 10^{16} times faster in stellar explosions than in a normal star. This is so fast that protons can be captured by radioactive isotopes before they have a chance to spontaneously decay. Sequences of nuclear reactions occurring in novae are therefore different from those in ordinary stars (Fig. 1*b*). Heavier nuclei, perhaps as high as mass 40 (calcium) and beyond, can be assembled in novae, ejected into space, and eventually incorporated into new stars and planets; even human bodies are made up of this "stardust." The high rate of nuclear reactions in novae also means energy is generated faster, influencing many aspects of the explosion. A similar situation is thought to occur in x-ray bursts, where the accretion from a companion star onto the surface of a neutron star occurs. Here, the temperatures can peak at more than 10^9 kelvins, the densities may reach as high as

10^6 grams per cubic centimeter, and isotopes up to mass 100 (tin) and beyond may be synthesized. *See* CATACLYSMIC VARIABLE; NOVA.

Measurements. Until recently, direct measurement of the fusion reactions occurring in stellar explosions was impossible. This is because beams of radioactive nuclei were not available, and targets of these nuclei would spontaneously decay before a measurement could be made. Computer models of nuclear burning in stellar explosions therefore had to rely on theoretical estimates of many crucial reaction rates. In some cases, these estimates have been shown to be incorrect by many orders of magnitude. Because of these uncertainties, critical comparisons of theory to astrophysical observations to determine the temperatures, densities, and duration of these events, as well as the origin of the elements in the world, were difficult to make.

The availability of beams of radioactive nuclei in nuclear physics laboratories has enabled experimental determinations of the nuclear reactions that drive stellar explosions. One approach to radioactive beam production, pioneered at CERN (Switzerland), is the isotope separator on-line (ISOL) technique, used at Louvain-la-Neuve (Belgium) and at Oak Ridge National Laboratory (United States) [**Fig. 2***a*]. One accelerator bombards a target with a beam of stable nuclei, and a small number of the radioactive atoms of interest are produced through nuclear reactions. These atoms are transported, by various techniques, including thermal diffusion, to an ion source where they are ionized (removing or adding electrons to give atoms an electrical charge) and extracted. The radioactive ions are then mass-separated from other ions and accelerated to energies needed for nuclear physics experiments by a second accelerator. The ISOL technique can produce very high beam qualities, purities, and intensities; the disadvantages are that only a few radioactive beam species can be generated from each combination of production target and primary beam, and that beams with short lifetimes (less than 1 s) are difficult to produce. *See* ION SOURCES; MASS SPECTROSCOPE.

To measure the fusion rate between radioactive heavy ions and hydrogen nuclei, a number of techniques can be used once a radioactive beam is produced. First, a beam of radioactive nuclei at a specific energy is directed at a thin (gas or foil) target containing hydrogen or helium. A direct measurement technique then involves counting the products of the fusion reaction (nuclei and protons) in sophisticated detection systems (**Fig. 3**). One effective system consists of a mass separator followed by charged particle detectors. By varying the energy of the radioactive beam, the probability of fusion as a function of energy is determined. Combined with a calculation of the relative energy distribution of the particles at a given temperature in a star, the fusion rate is determined as a function of temperature. Indirect techniques are also used to determine reaction rates, such as measuring the scattering of the heavy-ion radioactive beam off protons or alphas to discover resonances that may enhance the fusion re-
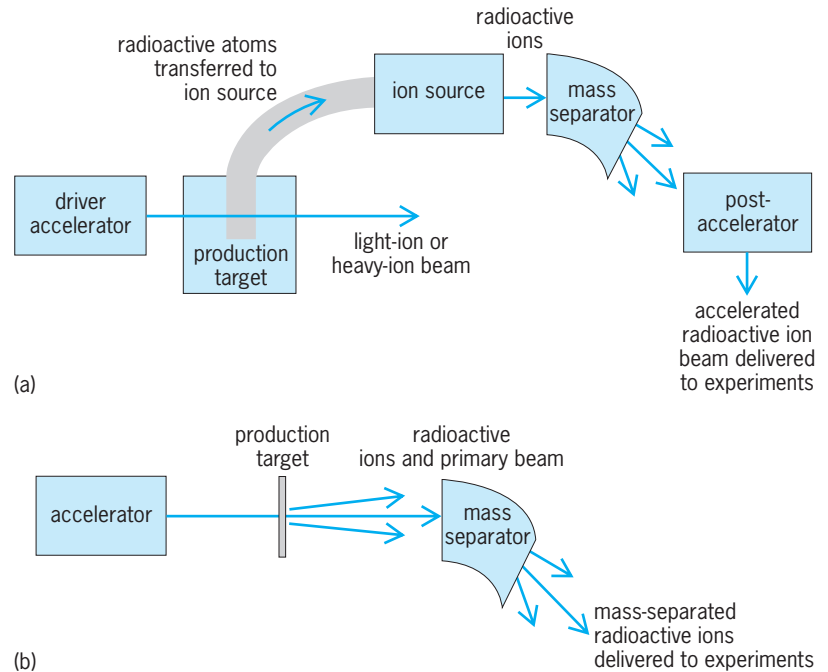

Fig. 2. Techniques for producing beams of radioactive nuclei. (*a*) Isotope separator on-line (ISOL) technique. (*b*) Projectile fragmentation technique.

action. Reactions that have been studied with ISOL beams include the fusion of the unstable ^{13}N, ^{17}F, ^{18}F, and ^{19}Ne nuclei with protons and ^{18}Ne with alphas. These reactions are all important for understanding element production in nova explosions.

A complementary radioactive beam production technique is projectile fragmentation, used at Michigan State University (United States), RIKEN (Japan), GANIL (France), and GSI (Germany) [Fig. 2*b*]. When a high-energy beam of stable heavy ions passes through a thin target, the beam particles (projectiles) can break up into fragments—some of which are the radioactive isotope of interest. The desired fragments are then mass-separated from other ions and steered toward a target to undergo the reaction of interest. The projectile fragmentation technique can produce beams of very short lifetimes (10^{-6} s or less), and the same setup can be used to produce many different
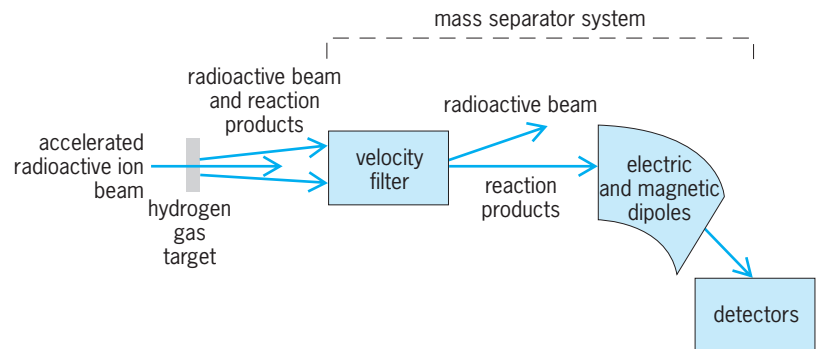

Fig. 3. Technique to measure a thermonuclear fusion reaction occurring in stellar explosions. A radioactive beam bombards a target with hydrogen (protons), and the products of a fusion reaction (as well as unreacted beam particles) enter a mass separator. The products of the thermonuclear reaction are steered onto a detector for counting, while the unreacted beam particles are deflected away.

beam species; the disadvantages are that high beam quality, purity, and intensity are difficult to obtain. Reactions studied with projectile fragmentation beams include the fusion of ^{7}Be and ^{13}N with protons, both by measuring the time-reversed reactions (for example, ^{8}Be breaking up to a ^{7}Be plus a proton). Both ISOL and projectile fragmentation techniques have also been used to measure the properties (masses, spins, parities, energy levels, and decay properties) of nuclei that are important in astrophysics calculations. Other techniques for producing radioactive beams have been used at Notre Dame University and Argonne National Laboratory in the United States. All together, these measurements enable nuclear physicists to directly measure some of the nuclear reactions that power cataclysmic stellar explosions. *See* NUCLEAR FUSION; NUCLEAR REACTION; PARTICLE ACCELERATOR.

Michael S. Smith

Bibliography. D. Arnett, *Supernovae and Nucleosynthesis*, Princeton University Press, 1996; D. D. Clayton, *Principles of Stellar Evolution and Nucleosynthesis*, University of Chicago Press, 1983; C. E. Rolfs and W. S. Rodney, *Cauldrons in the Cosmos: Nuclear Astrophysics*, University of Chicago Press, 1988.

Radioactive fallout

Whenever radioactive materials become airborne, either from a nuclear device detonation or from a nuclear release accident, the resultant contaminated atmospheric plume will ultimately return radioactivity to the Earth's surface. Material settling from the radioactive plume and its subsequent surface deposition is known as radioactive fallout.

Radioactivity. Radioactive materials consist of unstable atoms which emit gamma rays, beta particles, or alpha particles. These emissions—rays and particles—are unique in that they cause ionizations in neighboring atoms. The energy of the emitted radiation and subsequent ionizations can be a cause of concern if absorbed by living systems. *See* ALPHA PARTICLES; BETA PARTICLES; GAMMA RAYS; RADIOACTIVITY.

Nuclear bombs. Nuclear explosions (that is, atomic bombs) depend on the very rapid uncontrolled fissioning of a critical mass of uranium-235 (^{235}U) or plutonium-239 (^{239}Pu). Atomic fission occurs when a ^{239}Pu or ^{235}U atom absorbs an incoming neutron, causing the atom's nucleus to become unstable and to split into two or more parts, called daughters or fission products, each usually radioactive. In addition, the fission reaction releases heat and neutrons. The latter interacting with adjacent plutonium or uranium can induce further fissions in a very short time span (microseconds). When uncontrolled, this chain reaction causes an explosion. The process releases massive amounts of energy, usually measured in kilotons (thousands of tons) of TNT explosive equivalent. Fission bombs can release thousands of kilotons of energy, and the larger thermonuclear or hydrogen bombs can release millions of tons of TNT equivalent (that is, megaton explosions). *See* NUCLEAR EXPLOSION; NUCLEAR FISSION.

Atomic power. At nuclear power plants, the neutron concentration in the fuel is carefully regulated, and the rate of nuclear fission is slowed in a controlled chain reaction. Control is achieved by inserting rods of neutron-absorbing material, which reduces the flux of free neutrons available for further interaction with the fuel. The fissionable uranium in fuel is more dilute than that in a bomb, and the fissioning rate is carefully controlled. Heat from fission is extracted by the nuclear reactor's coolant to make the steam which is converted to electric power, or the coolant is used to generate electricity through direct conversion turbines. *See* CHAIN REACTION (PHYSICS); NEUTRON; NUCLEAR POWER.

Fission products. Fissioning of uranium and plutonium produces isotopes of about 70 different atoms; each atom may have several different isotopic forms. Examples of daughter products are the isotopes of elemental strontium, which are efficiently produced in nuclear fission. The isotope strontium-90 (^{90}Sr) has a 28-year half-life, while ^{89}Sr is produced in slightly higher concentrations but has a half-life of only 50 days. In addition to the differing half-lives, each isotope emits a unique radiation spectrum. The two strontium isotopes emit beta particles of different energies. Other isotopes such as cesium-137 (^{137}Cs, half-life 30 years) and iodine-131 (^{131}I, half-life 8 days) emit both beta particles and gamma rays. *See* CESIUM; IODINE; ISOTOPE; STRONTIUM.

Other radioactive materials. Spent nuclear fuel contains mainly its original uranium, an inventory of fission products, and small amounts of activation products and actinide elements. [Activation products are stable atoms made radioactive (activated) by neutron bombardment in and near the nuclear fuel.] Atom bomb debris also contains fission products, activation products, and unfissioned uranium or plutonium, all of which are radioactive. *See* ATOMIC BOMB; NUCLEAR FUELS.

Atmospheric fallout. Following a major destructive disassembly of a nuclear reactor, such as at Chernobyl, Ukraine, in April 1986, some of the fuel can be released in the form of particles and gases. If the accident is large enough, a fraction of the release becomes airborne, as was the case with Chernobyl. In this disaster, radioactivity injected high into the atmosphere spread over and covered most of the Northern Hemisphere. A surface or air burst of an atomic weapon produces the infamous "mushroom cloud" containing radioactive materials which also can be transported for long distances. *See* NUCLEAR REACTOR.

Global fallout. Soon after the end of World War II, the United States and the Soviet Union began an atomic weapons arms race. France, the United Kingdom, and China joined in, and many aboveground bomb tests were executed. Within 10 years, the amount of radioactivity lofted into the atmosphere began to reach alarming levels, and political pressure increased to halt aboveground atomic-bomb testing. In 1962, a limited test ban treaty forbidding

atmospheric nuclear weapons testing was signed by most nations possessing atomic weapons. Fallout levels began to drop, and today are only a few percent of their peak concentrations in 1962–1964. Most of the atmospheric testing was halted after 1962, but because some of the smallest particles were lofted to the highest altitudes and settle back to Earth at very slow rates, atmospheric fallout is still present. Most of the nuclear detonations occurred in the Northern Hemisphere; in Nevada and the Pacific for the United States, in Kazakhstan for the Soviet Union, in Lop Nor for China, and in Algeria for France. In the Southern Hemisphere, a few tests were conducted by the United Kingdom in Australia's desert and by France on some of its South Pacific islands. Tracking the radioactive clouds from atmospheric tests yielded valuable information on the dispersion patterns and particle size distributions of fallout. Tropospheric and stratospheric concentrations generally followed a westerly trajectory. Little Northern Hemisphere fallout migrated to the Southern Hemisphere; the traces generally stayed in the latitudes in which they were first injected. *See* STRATOSPHERE; TROPOSPHERE.

Local fallout. Most of the fallout from early atomic weapons testing consisted of larger, heavier particles and those smaller particles which agglomerated onto dust particles. These particles settled out close to their source and were at dangerous concentrations adjacent to detonation sites. During the early tests in Nevada, plumes of radioactivity containing mainly fine particles and volatile materials rather than the heavier particulate fractions, blanketed much of the United States. Dose reconstructions of the fallout plumes have shown that many areas of the United States received significant radioactive iodine exposures. Some exposures were high enough to have caused an increase in thyroid cancer risk. Tests at Semipalatinsk, Kazakhstan, produced similar fallout clouds.

A chemical explosion in 1957 (the "Kyshtym accident" at a nuclear materials processing facility in Cheliabinsk in the southern Ural Mountains of Russia) ejected a large, highly radioactive plume of long-lived radionuclides containing about 20 million curies, mainly ^{90}Sr and ^{137}Cs. The resultant "East Urals Radioactive Trace" (EURT) contaminated 25,000 km^2 (10,000 mi^2) of land, exposed about a quarter million people, and caused thousands to be relocated. In 1993, a similar but much smaller explosion in the Tomsk-7 nuclear materials processing facility near Tomsk in Siberia contaminated about 250 km^2 (100 mi^2).

Fallout factors. Atmospheric fallout can be scavenged by rainfall. Wet deposition, involving washing out of atmospheric fallout, can increase local deposition patterns. This was the case following the Chernobyl accident, where local rainfall in Belarus, Ukraine, and Russia washed high concentrations of radioactive iodine and cesium out of the plume and onto the spring pasture. Radioactive iodine and cesium are relatively volatile and were more easily "boiled" out of Chernobyl's burning core. Cesium

is a congener of potassium and therefore is fairly uniformly distributed throughout the body once inhaled or ingested. The result is a whole-body radiation dose. Furthermore, the energetic gamma-ray emission from ^{137}Cs adds a source of external radiation from surface deposits on the ground. These factors, in addition to its long half-life and relatively high concentration, make ^{137}Cs the major long-term contamination concern from fallout. For example, although the Chernobyl accident occurred in 1986, precautions must still be taken against potential intake doses of ^{137}Cs: inhalation doses can occur when burning wood from contaminated trees, and consuming mushrooms grown in contaminated forests delivers an ingestion dose.

Resuspension of surface radioactivity concentrations can cause secondary local radioactive fallout. Simple vehicle traffic over or next to a contaminated surface causing particle resuspension is one example. A more extreme example was produced following a prolonged drought in the Ural Mountains of Russia, which lowered the water level in a radioactivity storage pond. A sudden strong wind storm in 1967 resuspended fallout from the dry lake bed, which contained over a half million curies (1.8×10^{16} becquerels), adding further contamination to the 1957 EURT.

Radiation risks. Levels of ^{131}I in the plumes of radioactive fallout are of particular concern. With an 8 day half-life and a strong beta- and gamma-ray emission, this radionuclide concentrates almost exclusively in the thyroid gland. Recent dose reconstructions in the United States show that the 1950s and 1960s fallout radiation doses from ^{131}I were large enough to have increased the risk for thyroid cancer, especially in children.

In the 1950s, large-scale Pacific tests deposited significant concentrations of weapons fallout on some of the Marshall Islands. Seventeen of the nineteen children under 10 years of age received thyroid doses of about 1000 rads (10 grays) and were later found to have developed thyroid lesions (one became malignant, the others were treated surgically).

Following its release into the environment during the Chernobyl accident, radioactive iodine was inhaled and, more importantly, incorporated into milk following fallout onto dairy herd pasture. Subsequent consumption of this milk caused significant thyroid doses. Over a thousand children living downwind of Chernobyl, most of whom were infants at the time of the accident, have developed thyroid cancer, when only a dozen cases might normally have been expected. This was probably caused by the consumption of large amounts of contaminated milk by infants, and the small size of the thyroid at that stage of development. *See* THYROID GLAND.

In 1955, the United Nations established the Scientific Committee on the Effects of Atomic Radiation, due to concern over possible risks from fallout. It issues comprehensive reports about every 5 years and has collected and documented the world's literature on radioactive fallout and, more recently, on

possible radiation consequences. *See* RADIATION IN-JURY (BIOLOGY). Marvin Goldman

Bibliography. L. R. Anspaugh et al., The global impact of the Chernobyl reactor accident, *Science*, 242:1513–19, 1988; M. Eisenbud, *Environmental Radioactivity*, 3d ed., 1987; M. Goldman, The Russian radiation legacy: Its integrated impact and lessons, *Environ. Health Perspect.*, 105:1385–1391, 1997; J. N. Stannard, *Radioactivity and Health*, 1988; United Nations, *Sources, Effects and Risks of Ionizing Radiation*, 1988.

Radioactive minerals

Minerals that contain uranium (U) or thorium (Th) as an essential component of their chemical composition. Examples are uraninite (UO_2) or thorite ($ThSiO_4$). There are radioactive minerals in which uranium and thorium substitute for ions of similar size and charge. There are approximately 200 minerals in which uranium or thorium are essential elements, although many of these phases are rare and poorly described. These minerals are important, as they are found in ores mined for uranium and thorium, most commonly uraninite and its fine-grained variety, pitchblende, for uranium. Thorite and thorogummite are the principal ore minerals of thorium. Minerals in which uranium and thorium occur in trace amounts, such as zircon ($ZrSiO_4$), are important because of their use in geologic age dating. The isotope uranium-238 (^{238}U) decays to lead-206 (^{206}Pb); ^{235}U decays to ^{207}Pb; ^{232}Th decays to ^{208}Pb; thus, the ratios of the isotopes of uranium, thorium, and lead can be used to determine the ages of minerals that contain these elements. *See* DATING METHODS; GEOCHRONOMETRY; LEAD ISOTOPES (GEOCHEMISTRY); THORITE; URANINITE.

Classification. Typically, the uranium and thorium minerals are classified according to their principal anions: oxides, carbonates, sulfates, molybdates, phosphates, vanadates, and silicates. Thorium exists in a single valence state (Th^{4+}); uranium, in two valences (U^{4+} and U^{6+}). The U^{6+} ion exists as the uranyl ion (UO_2^{2+}), which is soluble; the U^{4+} ion is much less soluble. The double oxidation state of uranium results in the dissolution (on oxidation), transport (under oxidizing conditions), and precipitation (under reducing conditions), which leads to the formation of uranium deposits. The roll-front deposits in the sandstones of the Colorado Plateau are an example. *See* ORE AND MINERAL DEPOSITS.

Properties. The multiple oxidation states of uranium lead to great complexity in the structures of the radioactive minerals. Those in which the principal oxidation state is IV (U^{4+} and Th^{4+}) are dark in color or opaque and do not fluoresce in ultraviolet light. The lower oxidation states are characteristic of minerals formed under reducing conditions, such as pegmatites or hydrothermal deposits. The U^{6+} minerals comprise the uranates and the uranyls. The uranates ($M_2^{+}U_x^{6+}O_{3x+1}$) are red, orange, or brown, with a rare orange or red fluorescence; the uranyls contain

the $(UO_2)^{2+}$ ion. These compounds are bright lemon yellow, yellow, orange, or green and commonly fluoresce in ultraviolet radiation.

The U^{6+} phases form as low-temperature alteration products in near-surface deposits. Weathering may lead to multicolored sequences of these minerals up to tens of meters in thickness. The classic uranium deposits of the Shaba District of Zaire are typical. In addition to the uranyl ion in these highly oxidized phases, water of hydration is common. The great number of possible phases, the variable uranium oxidation states, and the water of hydration make these phases extremely difficult to identity. Some exist in a mixed oxidation state. An important phase in this regard is ianthinite [$(UO_2)(5UO_3)(10H_2O)$], but even apparently simple phases such as uraninite actually have complex formulas.

Radioactive decay of ^{238}U, ^{235}U, and ^{232}Th results in the emission of alpha particles (helium nuclei), alpha-recoil nuclei, beta particles (electrons), and gamma radiation (x-rays). Although the beta particles and gamma radiation can cause discoloration due to ionization, the alpha-decay event leads to extensive atomic displacements (approximately 1000 displaced atoms per decay event), until in some minerals the structure becomes amorphous, the metamict state. Some phases such as zircon are susceptible to radiation damage and may become completely amorphous. Other phases, such as uraninite, retain their crystallinity by annealing, despite doses that may be as high as 50 displacements per atom. Minerals with U^{6+} are seldom metamict, as they are generally of secondary origin. Metamict minerals are glassy in appearance, are optically isotropic, and have a conchoidal fracture, reduced hardness (25%), lower density (17%), and an increased susceptibility to chemical alteration. The changes in physical and chemical properties that are due to alpha-decay events are of interest because of potential radiation effects on nuclear waste forms used in radioactive waste disposal. *See* ALPHA PARTICLES; AMORPHOUS SOLID; BETA PARTICLES; GAMMA RAYS; METAMICT STATE; PLEOCHROIC HALOS; RADIOACTIVE WASTE MANAGEMENT; URANIUM; THORIUM. Rod Ewing

Bibliography. B. C. Chakoumakos et al., Alpha-decay—Induced fracturing in zircon: The transition from the crystalline to the metamict state. *Science*, 236:1556–1559, 1987; R. J. Finch and R. C. Ewing, The corrosion of uraninite under oxidizing conditions, *J. Nucl. Mat.*, 190:133–156, 1992; C. Frondel, *Systematic Mineralogy of Uranium and Thorium*, Geol. Surv. Bull. 1064, 1958; E. W. Heinrich, *Mineralogy and Geology of Radioactive Raw Materials*, 1958.

Radioactive tracer

A radioactive isotope which, when injected into a chemically similar substance or artificially attached to a biological or physical system, can be traced by radiation detection devices. Many problems in biology, medicine, and industrial engineering not amenable

to other approaches can be solved by the use of these tracers. *See* RADIOACTIVITY; RADIOACTIVITY AND RADIATION APPLICATIONS; RADIOISOTOPE; RADIOISOTOPE (BIOLOGY).

Uses in industry. Radioactive tracers are useful in industrial applications, such as the testing of wear and corrosion in mechanical components. Piston ring wear within an engine can be determined accurately by transforming some of the carbon within the ring into carbon-14. This is done by placing the ring in an intense neutron flux. The piston ring is then installed in the engine and, after a period of use, samples of the engine oil are removed and tested for carbon-14 content. In this way amounts of wear much smaller than could be found by weight measurements and other techniques can be determined.

On a larger scale, the entire operation of an industrial plant can be tested and controlled by tracer techniques. In one industrial survey, sulfur-35, prepared in an appropriate chemical form, was introduced into a batch of coke entering an iron-smelting plant. The fate of the sulfur within the plant and its presence in the iron were studied by taking samples and measuring the iron's sulfur-35 content. It is entirely possible that similar techniques could be used in the routine operation and monitoring of many industrial processes. Such applications have been restricted in the past chiefly because of the difficulty of hazard control, but with more sensitive detectors such procedures may become commonplace.

Uses in biology. The simplest radioactive tracer studies consist of the tagging of a biological entity with a radioactive isotope (radioisotope). The entity is then tracked by following the radiation from the isotope. Such a method has been used to study the nocturnal behavior of bats. Small sources of cobalt-60 were attached to the legs of the bats. A radiation detector connected to an automatic recorder registered the presence or absence of the bats from their nests at night. In a similar manner, cockroach migration in an urban sewer system has been studied, as well as behavior of various insects and small animals.

The operation becomes more complex when a large number of biological particles are labeled, for example, in the tagging of red blood cells or bacteria. When the labeled substance is injected into an animal, it is impossible to follow the individual labeled particles, but their average movement can be tracked by observations of the radiation. In this way it is possible to measure the average lifetime of a red blood cell or the diffusion rate of bacteria.

Finally, a radioisotope of a particular element can be used to tag that element. Phosphorus-32 can be introduced into the soil where a plant is growing, and the amount of phosphorus absorbed and its distribution throughout the plant can be studied.

In one of the earliest tracer studies, G. von Hevesy used radioactive lead to trace the pathways of lead metabolism in plants. It was essential in this study, as in any other tracer experiments of this type, that the chemical behavior of the radioisotope be identical to that of the stable isotope in order for it to circulate normally within the biological entity being studied. For most isotopes this is essentially true.

In most biological tracer experiments, the radio-isotope is introduced into the system and its radiation subsequently measured with Geiger-Müller counters or scintillation detectors. Extremely soft (low-intensity) radiations can be detected by the use of photographic film. In a typical experiment using this technique, part of a plant containing a radioactive isotope is sectioned and placed next to the film. The radiation from the plant darkens the film, and the pattern on the film shows the distribution of the radioisotopes in the plant. Such pictures are called autoradiographs. Autoradiographs can be taken of cellular and even subcellular particles. Studies have been made of the reproduction of chromosomes using such techniques. Similar studies can be carried out in animals. *See* AGRICULTURAL SCIENCE (ANIMAL); AUTORADIOGRAPHY; GEIGER-MÜLLER COUNTER; PHOTOSYNTHESIS; SCINTILLATION COUNTER.   Gordon L. Brownell

Uses in medicine. A radioactive atom can be attached to a molecule or more complex substance, which can then be used to examine a chemical reaction in a test tube, or it can be administered to a patient by ingestion or injection and subsequently be incorporated into a biochemical process. The radioactive emissions from the radioactive atom can be used to track (trace) the behavior of the labeled molecule or substance in biological processes by means of medical imaging, utilizing techniques such as positron emission tomography (PET) or single-photon-emission computed tomography (SPECT). *See* MEDICAL IMAGING.

The tracer principle makes it possible to characterize the dynamic state of biological systems, including the living human body. Radioactive tracers revolutionized biochemistry and are widely used in the practice of medicine as well as in biomedical research.

The branch of medicine that uses radioactive tracers in the care of patients is called nuclear medicine. Radiotracers of practically every element can be produced in nuclear reactors or cyclotrons. *See* NUCLEAR MEDICINE.

An example of an early application of the tracer principle in medicine is the use of radioactive iodine (iodine-131), discovered in the late 1930s, to detect abnormalities in thyroid gland metabolism. The rate of accumulation of radioiodine by the thyroid can be measured with a radiation detector, making it possible to detect abnormalities in iodine metabolism, such as reduced thyroid function (hypothyroidism) or increased function of the thyroid (hyperthyroidism). Another early use of biological tracers involved measuring the life-span of red blood cells, platelets, and white blood cells. *See* THYROID GLAND DISORDERS.

In nuclear medicine, radioactive tracers are measured within the body by using radiation detectors directed at the body from the outside. Nuclear medicine also uses measurements of radioactive tracers in body fluids (such as blood and urine) in test

tubes. The emission of gamma rays from radioactive tracers within the human body makes possible the measurement of regional function and biochemistry in practically every organ.

In a nuclear medicine procedure, small amounts of radioactive chemicals are injected into the patient's bloodstream, typically via an arm vein. Then the amount of radioactivity in different regions of the body or within an organ is measured by means of externally placed radiation detectors. The selection of a specific radioactive tracer depends on the particular bodily function and regional biochemical reactions that are to be studied.

Radioactive tracers provide an entirely new way of looking at disease. They can reveal characteristic patterns of blood flow, biochemical abnormalities, and the location, severity, and extent of disease in the brains of patients.

Radioactive tracers play a major role in the study not only of the brain but of practically every other organ, including the heart. Radioactive tracers are used as part of the diagnostic process. Three radionuclides—carbon-14, tritium (hydrogen-3), and phosphorus-32—remain the backbone of modern biomedical sciences, especially in the fields of biochemistry, molecular biology, and genetics.

The cell remains the basic unit of biology, but with radioactive tracers one can examine intracellular chemistry as well as the means of communication among cells and organs. The exquisite sensitivity with which radioactivity can be measured (there are instruments sensitive enough to detect the decay of a single radioactive atom) makes it possible to quantify exceedingly low concentrations of chemical substances in the body, not only in body fluids such as blood but also in the organs themselves.

To carry out such studies, persons are administered radioactive tracers that take part in specific chemical reactions in different organs of the body. The choice of the radioactive substance is determined by the chemical or physiologic process to be examined. The movement of the tracer is displayed in a series of scans or images that reveal regional biochemistry. Positron-emitting radiotracers are usually made in cyclotrons, chiefly by using carbon-11 or fluorine-18 and a PET scanner. Other radioactive tracers are imaged with a SPECT scanner.

Images from PET or SPECT instruments are cross-sectional pictures of the distribution of radiotracers within the body at various times after a radioactive tracer is injected into a vein. These cross sections, or tomographic slices, provide information relating to slices of the person's body. Mathematical models are used to calculate regional concentrations of chemical concentrations and reaction rates, which are then compared from one region of the body to another.

Radioactive tracers reveal abnormalities in the chemical workings of the body, whereas computer tomography (CT) detects abnormalities in anatomy. PET can often reveal disease before anatomical changes show up on a CT scan. Since biochemical problems often occur before anatomical changes,

PET can be used to diagnose certain diseases at an earlier stage than was possible in the past.

Henry N. Wagner, Jr.

Radioactive waste management

The treatment and containment of radioactive wastes. These wastes originate almost exclusively in the nuclear fuel cycle and in the nuclear weapons program. Their toxicity requires careful isolation from the biosphere. Their radioactivity is commonly measured in curies (Ci). The curies, chosen to approximate the activity of 1 gram of radium-226 (^{226}Ra), is equal to 3.7×10^{10} becquerels. The becquerel (Bq), the SI unit of activity (radioactive disintegration rate), is the activity of a radionuclide decaying at the rate of one spontaneous nuclear transition per second. Considering its toxicity, the curie is a rather large unit of activity. A more appropriate unit is the microcurie (1 μCi $= 10^{-6}$ Ci), but the nanocurie (1 nCi $= 10^{-9}$ Ci) and picocurie (1 pCi $= 10^{-12}$ Ci) are also frequently used. A common unit of nuclear generating capacity is the gigawatt (electric) [GW(e)], equal to 10^9 watts of electric power, as opposed to thermal power. *See* UNITS OF MEASUREMENT.

Radioactive wastes are classified in four major categories: spent fuel elements and high-level waste (HLW), transuranic (TRU) waste, low-level waste (LLW), and uranium mill tailings. Minor waste categories, such as radioactive gases produced during reactor operation, radioactive emissions resulting from the burning of uranium-containing coal, or contaminated uranium mine water, will not be discussed.

Spent fuel elements arise when uranium is fissioned in a reactor to generate energy. The fuel elements needed for the production of 1 GW(e)-year of electrical energy contain 40 metric tons (44 short tons) of uranium; the spent fuel contains 1 metric ton (2200 lb) of fission-product nuclides, and also transuranic nuclides such as plutonium and americium produced by neutron capture in uranium nuclei. Spent fuel elements arising in the civilian energy program will not be chemically reprocessed. In the United States, weapons-grade plutonium for nuclear explosives is produced in special reactors. In order to extract this plutonium, the spent fuel must be chemically reprocessed. The resulting high-level waste contains most of the fission products and transuranic elements, including residual plutonium. Transuranic waste, arising mainly during this reprocessing, is now defined as solid material contaminated to greater than 100 nCi/g (3.7×10^6 Bq/kg) with certain alpha-emitting radionuclides. (Prior to 1984, this limit had been set at 10 nCi/g.) Uranium mill tailings are the residues of the chemical extraction of uranium from the ore. Finally, low-level waste is a very broad category of wastes, covering almost every form of radioactive waste not falling into the other categories. Projections of the radioactivity and the volume of the different waste categories (**Figs. 1** and **2**) assume no new nuclear plants will be ordered, leading to an installed civilian nuclear
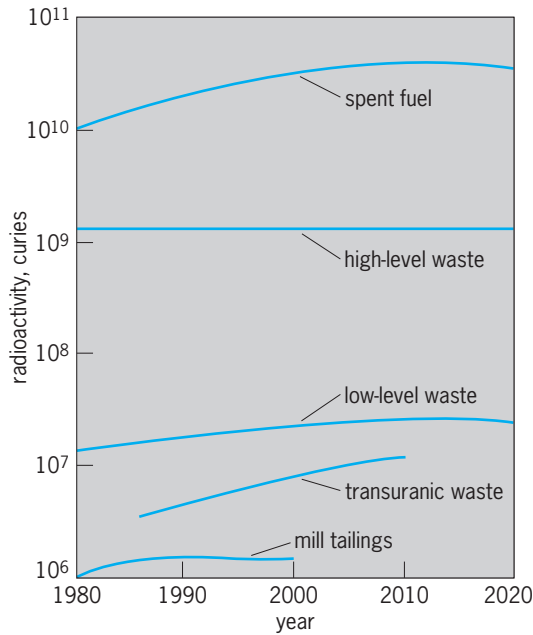
Fig. 1. Past and expected future accumulations of the four major waste categories: high-level waste and spent fuel, transuranic waste from the weapons program, low-level waste, and uranium mill tailings. The radioactivity in the mill tailings is the sum of the radioactivity of all uranium daughters. (*After U.S. Department of Energy, Spent Fuel and Radioactive Waste Inventories, Projections, and Characteristics, Rep. DOE/RW-0006, Rev. 5, Nov. 1989*)
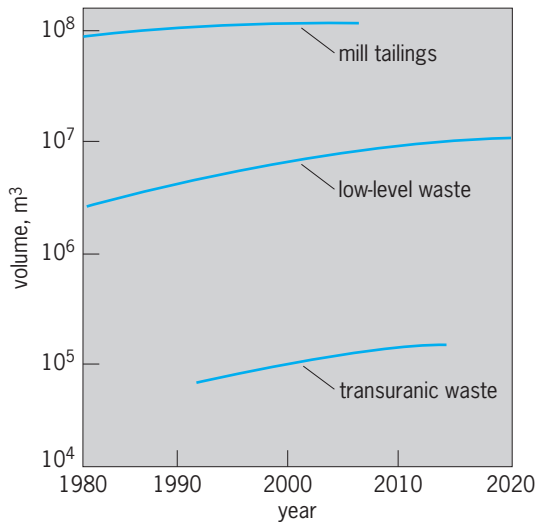

Fig. 2. Volume of the accumulated wastes shown in Fig. 1, except for that of the high-level waste and spent fuel, which is very small (less than 10,000 m³). 1 m³ = 35 ft³.

generating capacity of 102.5 GW(e) in the year 2000, and 51.6 GW(e) in 2020, as the plants reach the end of their lifetimes [1989: 97.5 GW(e)]. Considerably more radioactivity has been produced in the civilian nuclear energy program than in the weapons program. By the year 2020, the spent fuel is expected to consist of 75,000 metric tons (83,000 short tons) of uranium containing transuranic and fission-product nuclides. *See* ATOMIC BOMB; NUCLEAR FISSION; NUCLEAR FUEL CYCLE; NUCLEAR FUELS REPROCESSING; TRANSURANIUM ELEMENTS.

Since no practical methods exist to detoxify radioactive nuclides, protection against their harmful radiation must rely on their isolation from the biosphere until their radioactivity has decayed. Because each of the waste categories poses different problems, they will be discussed separately. The management of the radioactive wastes that arise during the dismantling of nuclear facilities will also be discussed.

Spent fuel and high-level waste. Most of the existing radioactivity is contained in this waste (Fig. 1). To identify and compare the major contributors to the radiotoxicity of the different nuclides, the water dilution volume is defined as the volume of water required to dilute these nuclides to acceptable concentrations, according to guidelines for occupational exposure, as specified by the International Commission of Radiological Protection (1979). The water dilution volumes for the most important isotopes contained in spent nuclear fuel have been calculated (**Fig. 3**; for high-level wastes, the results would be
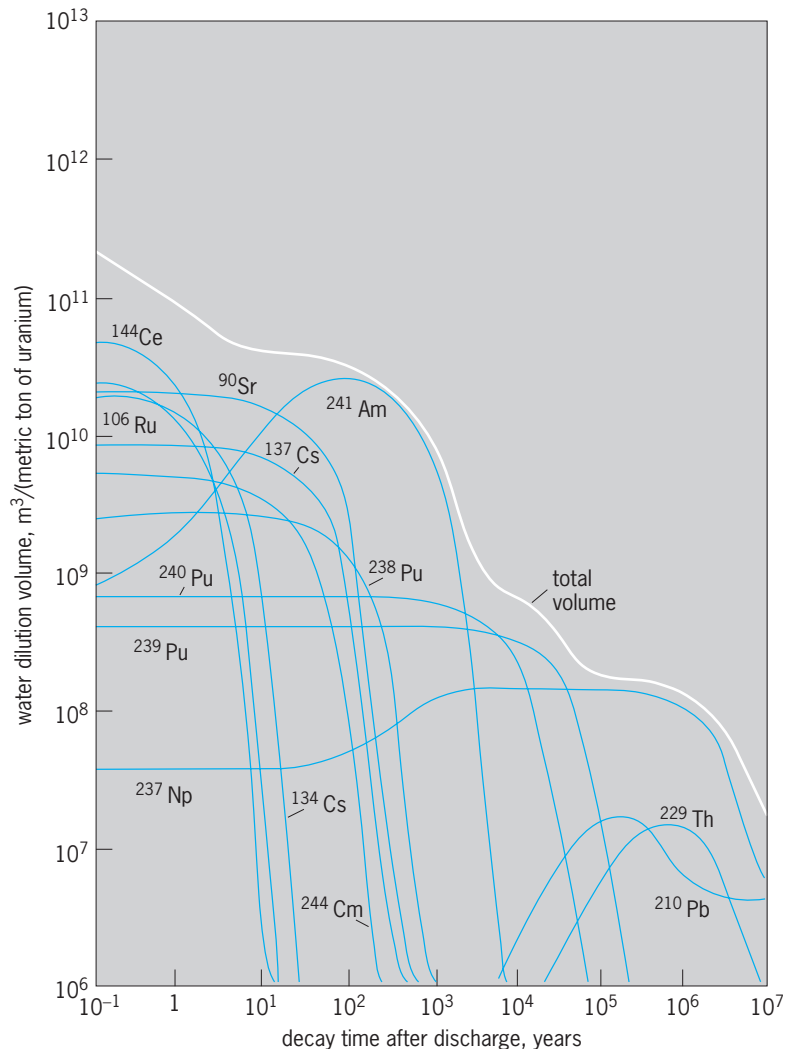

Fig. 3. Water dilution volume, as explained in the text, of spent fuel containing 1 metric ton (2200 lb) of uranium. The total amount of fuel needed to generate 1 GW(e)-year of electric energy is approximately 40 metric tons. The concentration of some of the transuranics increases for some time, as their parent nuclides decay. 1 m³ = 35 ft³. (*After Board on Radioactive Waste Management, National Research Council, A Study of the Isolation System for Geologic Disposal of Radioactive Wastes, National Academy Press, 1983*)

qualitatively similar, except for a reduced contribution by transuranics.) For the first 100 years, the toxicity is dominated by the beta- and gamma-emitting fission products [such as strontium-90 (^{90}Sr) and cesium-137 ($^{137}137$), with half-lives of approximately 30 years]; thereafter, the long-lived, alpha-emitting transuranium elements [for example, plutonium-239 (^{239}Pu), with a half-life of 24,000 years] and their radioactive decay daughters [for example, americium-241 (^{241}Am), with a half-life of 432 years, a daughter of plutonium-241 (^{241}Pu), with a half-life of 13 years] are important. Burial in geologic formations at a depth of 500–1000 m (1600–3200 ft) appears at present the most practical and attractive disposal method, although as of 1994 none of this waste had been disposed of in this way anywhere in the world. Most of the commercial spent fuel is stored in water-cooled basins at the reactor sites. Construction of some interim or emergency storage facilities aboveground for spent fuel has been considered. Such storage would further increase the risks of transportation accidents. The military reprocessed high-level waste is stored in tanks in liquid or solid form. Incorporation of high-level waste into glass blocks suitable for burial was expected to start in the United States in 1996. It is expected that by 2020 only 36% of the existing high-level waste, as measured by its radioactivity, will have been prepared in this form.

Geology as a predictive science is still in its infancy, and many of the parameters entering into model calculations of the long-term retention of the waste in geologic media are questionable. The major single problem is the heating of the waste and its surrounding rock by the radioactive decay heat. This heating can accelerate the penetration of groundwater into the repository, the dissolution of the waste, and its transport to the biosphere. Much effort has been devoted to the development of canisters to encapsulate the spent fuel elements or the glass blocks containing high-level waste, and of improved waste forms and overpacks that promise better resistance to attack by groundwater. The suitability of rock salt, which for 30 years had been widely considered the preferred host rock because of its superior heat dissipative properties and because of its plasticity (which would help to seal cracks), is now questioned because of several shortcomings. The major drawback is that rock salt is a natural resource with a large number of uses. It is impossible to assure that drilling into or through a salt formation housing a nuclear waste repository will never happen. The consequences of such drilling activities add so much to the complexity of predicting the safe containment of the nuclear waste that they seem to outweigh the geophysical advantages that salt might have over other rocks that are not natural resources and not water-soluble.

In the United States, the suitability of a disposal site in tuff, adjacent to the Nevada Nuclear Test Site (Yucca Mountain), is being studied. As planned, this repository could accept all the spent fuel elements, 70,000 metric tons (77,000 short tons), produced up to the year 2020 [used to generate approximately 2000 GW(e)-year of electrical energy]. If approved, the site would not be operational until 2010 at the earliest. *See* TUFF.

Transuranic waste. Although the radioactivity of the transuranic wastes is considerably smaller than that of high-level waste or spent fuel (Fig. 1), the high radiotoxicity and long lifetime of these wastes also require disposal in a geologic repository. Present plans call for disposal in bedded salt at the Waste Isolation Pilot Project (WIPP) in Carlsbad, New Mexico. Suitability for permanent disposal is to be tested over a period of approximately 5 years. During its expected subsequent period of operation, until 2018, the WIPP would accept all transuranic wastes generated up to that time. Waste with less than 100 nCi/g (3.78 Bq/kg) of transuranic elements will be treated as low-level waste.

Uranium mill tailings. Uranium is naturally radioactive, decaying in a series of steps to stable lead. It is currently a rare element, averaging between 0.1 and 0.2% in the mined ore. At the mill, the rock is crushed to fine sand, and the uranium is chemically extracted. The residues, several hundred thousand cubic meters for the annual fuel requirements of a 1-GW(e) reactor, are discharged to the tailings pile (**Fig. 4**). The tailings contain the radioactive daughters of the uranium. The long-lived isotope thorium-230 (^{230}Th, half-life 80,000 years) decays into radium-226 (^{226}Ra, half-life 1600 years), which in turn decays to radon-222 (^{222}Rn, half-life 3.8 days). Radium and radon are known to cause cancer, the former by ingestion, the latter by inhalation. Radon is an inert gas and thus can diffuse out of the mill tailings pile and into the air. It has been estimated by the Environmental Protection Agency that a person living 500 m (1600 ft) away from an unprotected tailings pile containing 280 pCi/g (1×10^4 Bq/kg) of ^{226}R would have a 30% higher chance of lung cancer from the radon gas it produces than the average person. (The average radium concentration in the existing tailings piles is twice as high, and hence, for these piles, the estimate should be doubled.) Ground-water pollution by radium that has leached from the pile has also been

Fig. 4. Partial view of an operating mill tailings pile (Homestake Mining Company, Milan, New Mexico). The pile measures approximately 1000 m × 1000 m (3000 ft × 3000 ft), is 13 m (42 ft) high on average, and is entirely unprotected. In its 1.7×10^7 m^3 (6×10^8 ft^3), it contains 8000 Ci (2.9×10^{14} Bq) of ^{226}Ra. The extracted uranium was enough to generate 100 GW(e)-year of electric energy in a reactor.

observed around tailings piles, but its health effects are more difficult to estimate, since the migration in the ground water is difficult to assess and also highly site-specific. *See* RADIOACTIVITY; RADIUM; RADON; URANIUM.

Although the radioactivity contained in the mill tailings is very small relative to that of the high-level waste and spent fuel, it is comparable to that of the transuranic waste (Fig. 1). It is mainly the dilution of the thorium and its daughters in the large volume of the mill tailings (Fig. 2) that reduces the health risks to individuals relative to those posed by the transuranium elements in the transuranic wastes. However, this advantage is offset by the great mobility of the chemically inert radon gas, which emanates into the atmosphere from the unprotected tailings. New mill tailings piles will be built with liners to protect the ground water, and will be covered with earth and rock to reduce atmospheric release of the radon gas. None of these measures provides protection on the time scales required for the ^{230}Th to decay. Permanent disposal methods, like chemical extraction of the thorium for disposal in a geologic repository, or burial of the tailings in deep mines, are not planned at this time. Because of the enormous volume of the tailings, a permanent disposal—should this be considered necessary—would present serious technical as well as economic problems.

Low-level wastes. By definition, practically everything that does not belong to one of the three categories discussed above is considered low-level waste. This name is misleading because some wastes, though low in transuranic content, may contain very high beta and gamma activity. For example, ion-exchange resins or activated components from reactors, which have radioactivities of hundreds of curies per cubic meter, may even require biological shielding during handling and transport. Of the 130,000 m^3 (4.6×10^6 ft^3) of low-level waste generated in 1988, about 70% originated in the weapons program, 20% in the commercial nuclear fuel cycle, 7% in industry, and 3% for research and medical purposes. The radioactivity contained in the last category was very small, less than 1% of the total radioactivity contained in the low-level waste, and is also mostly short-lived. The radioactivity of the wastes generated for medical purposes alone was less than 0.1%.

The current method of low-level waste disposal is shallow-land burial, which is relatively inexpensive but provides less protection than a geologic repository. [Prior to 1970, a total of 10^5 Ci (3.7×10^{15} Bq) of low-level waste generated in the United States weapons program was also disposed of by ocean dumping and prior to 1983, 1.3×10^6 Ci (4.8×10^{16} Bq) was also injected with grout into hydrofractured shale formations underlying the Oak Ridge National Laboratory. These practices have since been stopped.] At present, there are six major Department of Energy shallow-land burial sites and six commercial sites (only two of which are now operating). While sites in arid environments have generally performed in an acceptable manner, those located in humid environments have commonly not performed as

hoped, and the fact that three of the four commercial sites in the eastern United States are no longer operating illustrates the problems they have encountered. It also points to the difficulties to be expected in developing new regional low-level waste disposal sites, which are to be located in the humid northeastern United States, as required by law. Besides the very large volume, the nonuniformity of the waste presents formidable problems for low-level waste disposal. Current programs focus on improved waste forms, better site selection criteria, and engineering improvements to the site in order to restrict releases.

Decommissioning of nuclear facilities. At the end of their lifetime, nuclear facilities have to be dismantled (decommissioned) and the accumulated radioactivity disposed of. Nuclear power plants represent the most important category of nuclear facilities, containing the largest amounts of radioactive wastes, and will therefore be discussed here. These wastes can be grouped in three classes: neutron-activated wastes, surface-contaminated wastes, and miscellaneous wastes. *See* NUCLEAR POWER; NUCLEAR REACTOR.

The neutron-activated wastes are mainly confined to the reactor pressure vessel and its internal components, which have been exposed to large neutron fluences during reactor operation. These components contain significant amounts of long-lived nontransuranic radioactive isotopes such as niobium-94 (^{94}Nb, an impurity in the stainless steel), which emits highly penetrating gamma rays and has a half-life of 20,000 years. These wastes are unacceptable for shallow-land disposal as low-level wastes. Disposal in a geologic repository is envisioned. Surface-contaminated and miscellaneous radioactive wastes derive mainly from faulty fuel pin claddings, which allow radioactive material to escape from the fuel. While the surface contamination can be effectively removed from smooth surfaces, such decontamination will inevitably increase the amount of the miscellaneous wastes (solvents, filters, and so forth). *See* DECONTAMINATION OF RADIOACTIVE MATERIALS.

Delaying the dismantling of the facility, in a procedure called safe storage or entombment, allows much of the radioactivity to decay. While this reduces the dismantling cost, it will shift the burden to future generations. For a decommissioning scenario assuming a 2-year wait after shutdown before dismantling and decontaminating the facility, the cumulative volume and radioactivity for the 68 reactors expected to be shut down between 1989 and 2020 has been estimated to be 8.4×10^5 m^3 (3.0×10^7 ft^3) and 3.3×10^7 Ci (1.2×10^{18} Bq), respectively. Of this volume, 0.6% will contain 97% of the radioactivity (mainly the neutron-activated reactor internals). The remainder, containing 1×10^6 Ci (3.7×10^{16} Bq), will be buried in shallow-land disposal sites, where it will constitute by the year 2020 approximately 10% of the accumulated volume and 3% of the accumulated radioactivity (Figs. 1 and 2).   Robert O. Pohl

Bibliography. R. E. Berlin, *Radioactive Waste Management*, 1989; Board on Radioactive Waste Management, National Research Council, *A Study of the*

Isolation System for Geologic Disposal of Radioactive Wastes, 1983; B. W. Burton et al., *Overview Assessment of Nuclear Waste Management*, Los Alamos Sci. Lab. Rep. LA-9395-MS, August 1982; R. D. Lipschutz, *Radioactive Waste: Politics, Technology, and Risk*, 1980; A. G. Milnes, *Geology and Radwaste*, 1985; U.S. Department of Energy, *Spent Fuel and Radioactive Waste Inventories, Projections, and Characteristics*, Rep. DOE/RW-0006, Rev. 9, March 1994; U.S. Environmental Protection Agency, *Draft Environmental Impact Statement for Standards for the Control of By-product Materials from Uranium Ore Processing* (40 CFR 192), Rep. EPA 520/1-82-022, March 1983.

Radioactivity

A phenomenon resulting from an instability of the atomic nucleus in certain atoms whereby the nucleus experiences a spontaneous but measurably delayed nuclear transition or transformation with the resulting emission of radiation. The discovery of radioactivity by Henri Becquerel in 1896 was an indirect consequence of the discovery of x-rays a few months earlier by Wilhelm Roentgen, and marked the birth of nuclear physics. *See* X-RAYS.

On the other hand, nuclear physics can also be said to begin with the proposal by Ernest Rutherford in 1911 that atoms have a nucleus. On the basis of the scattering of alpha particles (emitted in radioactive decay) by gold foils, Rutherford proposed a solar model of atoms, where negatively charged electrons orbit the tiny nucleus, which contains all the positive charge and essentially all the mass of the atom, as planets orbit around the Sun. The attractive Coulomb electrical force holds the electrons in orbit about the nucleus. Atoms have radii of about 10^{-10} m and the nuclei of atoms have radii about 2×10^{-15} m, so atoms are mostly empty space, like the solar system. Niels Bohr proposed a theoretical model for the

Fig. 1.   Decay modes of ^{114}Cs based on *Q* values from the droplet-model formula for nuclear masses. (*After J. H. Hamilton et al., eds., Future Directions in Studies of Nuclei Far from Stability, 1980*)

atom that removed certain difficulties of the Rutherford model. *See* ATOMIC STRUCTURE AND SPECTRA.

However, it was only after the discovery of the neutron in 1932 that a proper understanding was achieved of the particles that compose the nucleus of the atom. The nucleus contains protons that carry positive charge and neutrons with slightly higher mass and zero net electrical charge. These protons and neutrons (called nucleons) are held inside the nucleus by the nuclear force between these particles. This force gives rise to the binding energy of the nucleus, which is the energy required to pull all the protons and neutrons apart. The binding energy makes the mass of a nucleus less than the masses of the *Z* protons and *N* neutrons that make up the nucleus. In all radioactive decays the total number of nucleons, $A = Z + N$, before and after the decay is a constant; that is, the number of nucleons is conserved. Differences in binding energies produce differences in masses which in turn determine what type of radioactive decay can occur. Radioactive decay occurs when the masses of all the particles after the decay are less than the mass of the original radioactive nucleus. *See* NEUTRON; NUCLEAR BINDING ENERGY; NUCLEAR STRUCTURE; NUCLEON; PROTON.

In 1934, Irère Curie and Frédéric Joliot demonstrated that radioactive nuclei can be made in the laboratory. All chemical elements may be rendered radioactive by adding or by subtracting (except for hydrogen and helium) neutrons from the nucleus of the stable ones. Studies of the radioactive decays of new isotopes far from the stable ones in nature continue as a major frontier in nuclear research. The availability of this wide variety of radioactive isotopes has stimulated their use in many different fields, including chemistry, biology, medicine, industry, artifact dating, agriculture, and space exploration. *See* ALPHA PARTICLES; BETA PARTICLES; GAMMA RAYS; ISOTOPE; RADIOACTIVITY AND RADIATION APPLICATIONS.

A particular radioactive transition may be delayed by less than a microsecond or by more than a billion years, but the existence of a measurable delay or lifetime distinguishes a radioactive nuclear transition from a so-called prompt nuclear transition, such as is involved in the emission of most gamma rays. The delay is expressed quantitatively by the radioactive decay constant, or by the mean life, or by the half-period for each type of radioactive atom.

The most commonly found types of radioactivity are alpha, beta negatron, beta positron, electron capture, and isomeric transition (**Table 1**). Each is characterized by the particular type of nuclear radiation which is emitted by the transforming parent nucleus. In addition, there are several other decay modes that are observed more rarely in specific regions of the periodic table (Table 1). Several of these rarer processes are in fact two-step processes (**Figs. 1** and **2**). In addition, there are several other processes predicted theoretically that remain to be verified.

In alpha radioactivity (Table 1) the parent nucleus spontaneously emits an alpha particle. Since the
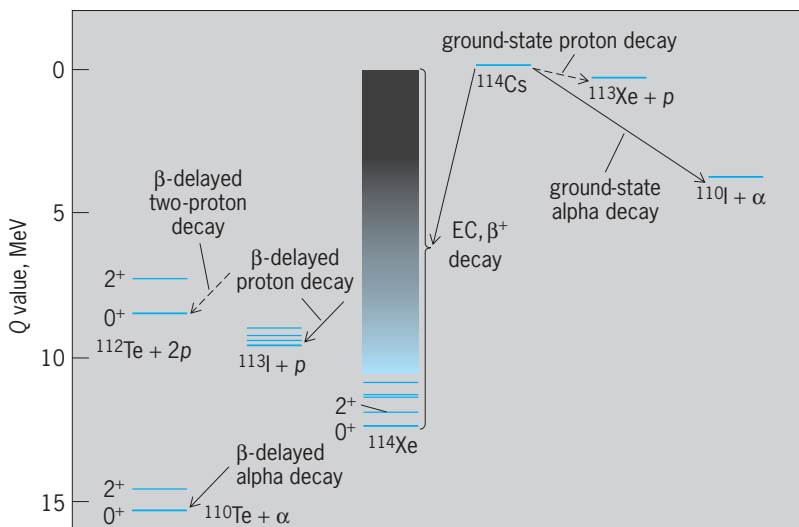
TABLE 1. Types of radioactivity

Type	Symbol	Particles emitted	Change in atomic number, ΔZ	Change in atomic mass number, ΔA	Example
Alpha	α	Helium nucleus	−2	−4	$^{226}_{86}$Ra $\to$ $^{222}_{84}$Rn + α
Beta negatron	β^-	Negative electron and antineutrino[a]	+1	0	$^{24}_{11}$Na $\to$ $^{24}_{12}$Mg + e^- + $\bar{\nu}$
Beta positron	β^+	Positive electron and neutrino[a]	−1	0	$^{22}_{11}$Na $\to$ $^{22}_{10}$Ne + e^+ + ν
Electron capture	EC	Neutrino[a]	−1	0	$^{7}_{4}$Be + e^- $\to$ $^{7}_{3}$Li + ν
Isomeric transition[b]	IT	Gamma rays or conversion electrons or both (and positive-negative electron pair)[c]	0	0	$^{137m}_{56}$Ba $\to$ $^{137}_{56}$Ba + γ or conversion electrons
Proton	p	Proton	−1	−1	$^{151}_{71}$La $\to$ $^{150}_{70}$Yb + p
Spontaneous fission (hot)	SF	Two intermediate-mass nuclei and 1–10 neutrons	Various	Various	$^{238}_{92}$U $\to$ $^{133}_{50}$Sn + $^{103}_{42}$Mo + $2n$
Spontaneous fission (cold)	SF	Two intermediate-mass nuclei (zero neutrons)	Various	Various	$^{252}_{98}$Cf $\to$ $^{106}_{42}$Mo + $^{146}_{56}$Ba
Ternary spontaneous fission (hot)	TSF	Two intermediate-mass nuclei, a light particle (^{2}H, α, up to ^{12}C), and neutrons	Various	Various	$^{252}_{98}$Cf $\to$ $^{100}_{40}$Zr + $^{146}_{56}$Ba + α + $2n$
Ternary spontaneous fission (cold)	TSF	Two intermediate-mass nuclei and a light particle	Various	Various	$^{252}_{98}$Cf $\to$ $^{96}_{38}$Sr + $^{146}_{56}$Ba + $^{10}_{4}$Be
Isomeric spontaneous fission	ISF	Heavy fragments and neutrons	Various	Various	$^{244f}_{95}$Am $\to$ $^{134}_{53}$I + $^{107}_{42}$Mo + $3n$
Beta-delayed spontaneous fission	$(EC + \beta^+)$SF	Positive electron, neutrino, heavy fragments, and neutrons	Various	Various	$^{246}_{99}$Es $\to$ β^+ + ν + $^{246f}_{98}$Cf $\to$ $^{138}_{54}$Xe + $^{107}_{44}$Ru + n
	β^-SF	Negative electron, antineutrino, heavy fragments, and neutrons	Various	Various	$^{236}_{91}$Pa $\to$ β^- + ν + $^{236f}_{92}$U $\to$ $^{139}_{53}$I + $^{94}_{39}$Y + $3n$
Beta-delayed neutron	$\beta^- n$	Negative electron, and antineutrino, neutron	+1	−1	$^{11}_{3}$Li $\to$ β^- + ν + $^{11}_{4}$Be* $\to$ $^{10}_{4}$Be + n
Beta-delayed two-neutron (three-, four-neutron)	$\beta^- 2n\,(3n, 4n)$	Negative electron, antineutrino, and two (three, four) neutrons	+1	−2 (−3, −4)	$^{11}_{3}$Li $\to$ β^- + ν + $^{11}_{4}$Be* $\to$ $^{9(8)}_{4}$Be + $2n(3n)$
Beta-delayed proton	$\beta^+ p$ or $(\beta^+ + EC)p$	Positive electron, neutrino, and proton	−2	−1	$^{114}_{55}$Cs $\to$ β^+ + ν + $^{114}_{54}$Xe* $\to$ $^{113}_{53}$I + p
Beta-delayed two-proton	$\beta^+\,2p$	Positive electron, neutrino, and two protons	−3	−2	$^{22}_{13}$Al $\to$ β^+ + ν + $^{22}_{12}$Mg* $\to$ $^{20}_{10}$Ne + $2p$
Beta-delayed deuteron	β^{-2}_{1}H	Negative electron, antineutrino, and deuteron	0	−2	$^{6}_{2}$He $\to$ β^- + $^{6}_{3}$Li $\to$ $^{2}_{1}$H + $^{4}_{2}$He
Beta-delayed triton	β^{-3}_{1}H	Negative electron, antineutrino, and triton	0	−3	$^{11}_{3}$Li $\to$ β^- + ν + $^{11}_{4}$B* $\to$ $^{8}_{3}$Li + $^{3}_{1}$H
Beta-delayed alpha	$\beta^+ \alpha$	Positive electron, neutrino, and alpha	−3	−4	$^{114}_{55}$Cs $\to$ β^+ + ν + $^{114}_{54}$Xe* $\to$ $^{110}_{52}$Te + α
	$\beta^- \alpha$	Negative electron, antineutrino, and alpha	−1	−4	$^{214}_{83}$Bi $\to$ β^- + ν + $^{114}_{84}$Po* $\to$ $^{210}_{82}$Pb + α
Beta-delayed alpha-neutron	$\beta^- \alpha, n$	Negative electron, antineutrino, alpha, and neutron	−1	−5	$^{11}_{3}$Li $\to$ β^- + ν + $^{11}_{4}$B* $\to$ $^{6}_{2}$He + α + n
Double beta decay	$\beta^- \beta^-$	Two negative electrons and two antineutrinos	+2	0	$^{82}_{34}$Se $\to$ $^{82}_{36}$Kr + $2\beta^-$ + 2ν
	$\beta^+ \beta^+$	Two positive electrons and two neutrinos	−2	0	$^{d130}_{56}$Ba $\to$ $^{130}_{54}$Xe + $2\beta^+$ + 2ν
Double electron capture[d]	EC EC	Two neutrinos	−2	0	$^{d130}_{56}$Ba + $2e^-$ $\to$ $^{130}_{54}$Xe + 2ν
Neutrinoless double beta decay[e]	$\beta^- \beta^-$	Two negative electrons	+2	0	$^{e82}_{34}$Se $\to$ $^{82}_{36}$Kr + $2\beta^-$
Two-proton	$2p$	Two protons	−2	−2	$^{45}_{26}$Fe $\to$ $^{43}_{24}$Cr + $2p$
Neutron[d]	n	Neutron	0	−1	

(cont.)

TABLE 1. Types of radioactivity (*cont*.)

Type	Symbol	Particles emitted	Change in atomic number, ΔZ	Change in atomic mass number, ΔA	Example
Two-neutron[d]	$2n$	Two neutrons	0	-2	
Heavy clusters[f]	$^{14}_{6}C$	$^{14}_{6}C$ nucleus	-6	-14	$^{223}_{88}Ra \rightarrow ^{209}_{82}Pb + ^{14}_{6}C$
	$^{20}_{8}O$	$^{20}_{8}O$ nucleus	-8	-20	$^{227}_{89}Ac \rightarrow ^{207}_{81}Tl + ^{20}_{8}O$
	$^{24}_{10}Ne$	$^{24}_{10}Ne$ nucleus	-10	-24	$^{232}_{92}U \rightarrow ^{208}_{82}Pb + ^{24}_{10}Ne$

[a] The neutrinos and antineutrinos emitted in beta decay are electron neutrinos and antineutrinos.
[b] Excited states with relatively long measured half-lives are called isomeric and are identified by placing the symbol *m* for metastable after the mass number, as in ^{137m}Ba. Excited states with essentially prompt decay are identified by asterisks, as in ^{11}Be*.
[c] Pair emission occurs as an additional competing decay mode when the decay energy exceeds 1.022 MeV.
[d] Theoretically predicted but not established experimentally.
[e] Theoretically possible and predicted in some grand unified theories.
[f] There are other possible clusters in addition to those shown. Cold SF is also a type of cluster radioactivity.

alpha particle is the nucleus of the helium-4 atom, it contains two protons and two neutrons. Thus, the atomic number, or nuclear charge Z, of the decay product is 2 units less than that of the parent, and the nuclear mass A of the product is 4 atomic mass units less than that of the parent, because the emitted alpha particle carries away this amount of nuclear charge and mass. This decrease of 2 units of atomic number or nuclear charge between parent and product means that the decay product will be a different chemical element, displaced by 2 units to the left in a periodic table of the elements. For example, radium has atomic number 88 and is found in column 2 of the periodic table. Its decay product after the emission of an alpha particle is a different chemical element, radon, whose atomic number is 86 and whose position is in column 0 of the periodic table.
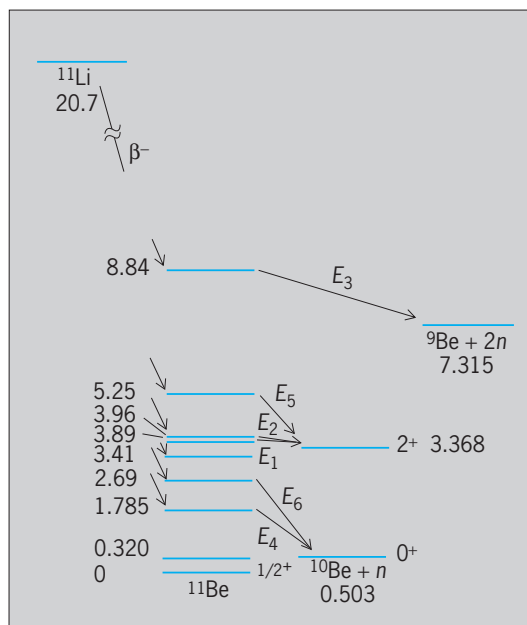

Fig. 2. Observed decay modes of ^{11}Li. Energies in megaelectronvolts. (*After J. H. Hamilton et al., eds., Future Directions in Studies of Nuclei Far from Stability*, 1980)

Transition Rates and Decay Laws

This section covers radioactive decay constant, dual decay, exponential decay law, mean life, and half-period.

Radioactive decay constant. The rate of radioactive transformation, or the activity, of a source equals the number A of identical radioactive atoms present in the source, multiplied by their characteristic radioactive decay constant λ. Thus Eq. (1) holds, where the

$$\text{Activity} = A\lambda \text{ disintegrations per second} \quad (1)$$

decay constant λ has dimensions of s^{-1}. The numerical value of λ expresses the statistical probability of decay of each radioactive atom in a group of identical atoms, per time. For example, if $\lambda = 0.01 \text{ s}^{-1}$ for a particular radioactive species, then each atom has a chance of 0.01 (1%) of decaying in 1 s, and a chance of 0.99 (99%) of not decaying in any given 1-s interval. The constant λ is one of the most important characteristics of each radioactive nuclide: λ is essentially independent of all physical and chemical conditions such as temperature, pressure, concentration, chemical combination, or age of the radioactive atoms. There are a few cases where measurable effects are observed for different chemical combinations. One of the largest observed is a 3.2% change in λ for the 24-s isomer in ^{90}Nb. The half-period is inversely proportional to λ.

The identification of some radioactive samples can be made simply by measuring λ, which then serves as an equivalent of qualitative chemical analysis. For the most common radioactive nuclides, the range of λ extends from $3 \times 10^6 \text{ s}^{-1}$ (for thorium C′) to $1.6 \times 10^{-18} \text{ s}^{-1}$ (for thorium).

Dual decay. Many radioactive nuclides have two or more independent and alternative modes of decay. For example, ^{238}U can decay either by alpha-particle emission or by spontaneous fission. A single atom of ^{64}Cu can decay in any of three competing independent ways: negatron beta-particle emission, positron beta-particle emission, or electron capture. When two or more independent modes of decay are possible, the nuclide is said to exhibit dual decay.

The competing modes of decay of any nuclide have independent partial decay constants given by the probabilities $\lambda_1, \lambda_2, \lambda_3, \ldots$ per second, and the total probability of decay is represented by the total decay constant λ, defined by Eq. (2). If there are A

$$\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \cdots \qquad (2)$$

identical atoms present, the partial activities, as measured by the different modes of decay, are $A\lambda_1$, $A\lambda_2$, $A\lambda_3$, ..., and the total activity $A\lambda$ is given by Eq. (3).

$$A\lambda = A\lambda_1 + A\lambda_2 + A\lambda_3 + \cdots \qquad (3)$$

The partial activities, $A\lambda_1, \ldots$, such as positron beta particles from ^{64}Cu, are proportional to the total activity, $A\lambda$, at all times.

The branching ratio is the fraction of the decaying atoms which follow a particular mode of decay, and equals $A\lambda_1/A\lambda$ or λ_1/λ. For example, in the case of ^{64}Cu the measured branching ratios are $\lambda_1/\lambda = 0.40$ for negatron beta decay, $\lambda_2/\lambda = 0.20$ for positron beta decay, and $\lambda_3/\lambda = 0.40$ for electron capture. The sum of all the branching ratios for a particular nuclide is unity.

Exponential decay law. The total activity, $A\lambda$, equals the rate of decrease $-dA/dt$ in the number of radioactive atoms A present. Because λ is independent of the age t of an atom, integration of the differential equation of radioactive decay, $-dA/dt = A\lambda$, gives Eq. (4), where ln represents the natural log-

$$\ln \frac{A}{A_0} = -\lambda(t - t_0) \qquad (4)$$

arithm to the base e, and A atoms remain at time t if there were A_0 atoms initially present at time t_0. If $t_0 = 0$, then Eq. (4) can be rewritten as the exponential law of radioactive decay in its most common form, Eq. (5). The initial activity at $t = 0$ was $A_0\lambda$, and

$$A = A_0 e^{-\lambda t} \qquad (5)$$

the activity at t, when only A atoms remain untransformed, is $A\lambda$. Because λ is a constant, the fractional activity $A\lambda/A_0\lambda$ at time t and the fractional amount of radioactive atoms A/A_0 are given by Eq. (6). In cases

$$\frac{A\lambda}{A_0\lambda} = \frac{A}{A_0} = e^{-\lambda t} \qquad (6)$$

of dual decay, the partial activities $A\lambda_1$, $A\lambda_2$, ... also decrease with time as $e^{-\lambda t}$, not as $e^{-\lambda_1 t} \ldots$, because $A\lambda_1/A_0\lambda_1 = A/A_0 = e^{-\lambda t}$ where λ is the total decay constant. This is because the decrease of each partial activity with time is due to the depletion of the total stock of atoms A, and this depletion is accomplished by the combined action of all the competing modes of decay.

Mean life. The actual life of any particular atom can have any value between zero and infinity. The average or mean life of a large number of identical radioactive atoms is, however, a definite and important quantity.

If there are A_0 atoms present initially at $t = 0$, then the number remaining undecayed at a later time t is $A = A_0 e^{-\lambda t}$, by Eq. (5). Each of these A atoms has

a life longer than t. In an additional infinitesimally short time interval dt, between time t and $t + dt$, the absolute number of atoms which will decay on the average is $A\lambda dt$, and these atoms had a life-span t. The total L of the life-spans of all the A_0 atoms is the sum or integral of $tA\lambda \, dt$ from $t = 0$ to $t = \infty$, which is given by Eq. (7). Then the average lifetime L/A_0, which is called the mean life τ, is given by Eq. (8),

$$L = \int_0^\infty tA\lambda \, dt = \int_0^\infty tA_0\lambda e^{-\lambda t} dt = \frac{A_0}{\lambda} \qquad (7)$$

$$\tau = \frac{1}{\lambda} \qquad (8)$$

where λ is the total radioactive decay constant of Eq. (2). Substitution of $t = \tau = 1/\lambda$ into Eq. (6) shows that the mean life is the time required for the number of atoms, or their activity, to fall to $e^{-1} = 0.368$ of any initial value.

Half-period (half-life). The time interval over which the chance of survival of a particular radioactive atom is exactly one-half is called the half-period T (also called half-life, written $T_{1/2}$). From Eq. (4), Eq. (9) is

$$-\ln \frac{A}{A_0} = \ln \frac{A_0}{A} = \ln 2 = 0.693 = \lambda T \qquad (9)$$

obtained. Then the half-period T is related to the total radioactive decay constant λ, and to the mean life τ, by Eq. (10). For mnemonic reasons, the half-period

$$T = \frac{0.693}{\lambda} = 0.693\tau \qquad (10)$$

T is much more frequently employed than the total decay constant λ or the mean life τ. For example, it is more common to speak of ^{232}Th as having a half-period of 1.4×10^{10} years than to speak of its mean life of 2.0×10^{10} years or its total decay constant of 1.6×10^{-18} s^{-1}, although all three are equivalent statements of the average longevity of ^{232}Th atoms.

Any initial activity $A_0\lambda$ is reduced to $1/2$ in 1 half-period T, to $1/e$ in 1 mean life τ, to $1/4$ in 2 half-periods $2T$, and so on (**Fig. 3**). The slope of the
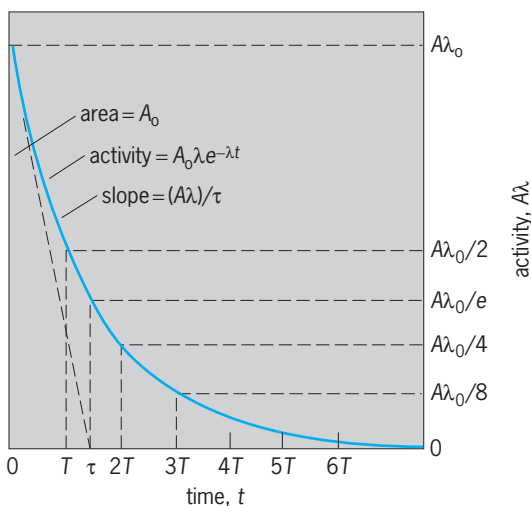

Fig. 3. Graphical representation of relationships in decay of a single radioactive nuclide.

activity curve, or rate of decrease of activity, is $d(A\lambda)/dt = -\lambda dA/dt = -\lambda(A\lambda)$. Thus the initial slope is $-\lambda(A_0\lambda) = -(A_0\lambda)/\tau$. The area under the activity curve, if integrated to $t = \infty$, is simply A_0, the total initial number of radioactive atoms. Also, the initial activity $A_0\lambda$, if it could continue at a constant value for one mean life τ, would exactly destroy all the atoms because $(A_0\lambda)\tau = A_0$.

Radioactive Series Decay

In a number of cases a radioactive nuclide A decays into a nuclide B which is also radioactive; the nuclide B decays into C which is also radioactive, and so on. For example, $^{232}_{90}$Th decays into a series of 10 successive radioactive nuclides. Substantially all the primary products of nuclear fission are negatron beta-particle emitters which decay through a chain or series of two to six successive beta-particle emitters before a stable nuclide is reached as an end product. *See* NUCLEAR FISSION.

Let the initial part of such a series be represented by reaction (11), where radioactive atoms of types A,

$$A \xrightarrow{\lambda_A} B \xrightarrow{\lambda_B} C \xrightarrow{\lambda_C} D \xrightarrow{\lambda_D} \cdots \quad (11)$$
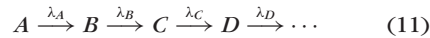
$B, C, D, \ldots$ have radioactive decay constants given by $\lambda_A, \lambda_B, \lambda_C, \lambda_D, \ldots$. Then if there are initially present, at time $t = 0$, A_0 atoms of type A, the numbers A, B, C, $\ldots$ of atoms of types A, B, C, $\ldots$, which will be present at a later time t, are given by Eqs. (12)–(14), and the activities of A, B, C, $\ldots$ are $A\lambda_A$, $B\lambda_B$,

$$A = A_0 e^{-\lambda_A t} \quad (12)$$

$$B = A_0 \frac{\lambda_A}{\lambda_B - \lambda_A}(e^{-\lambda_A t} - e^{-\lambda_B t}) \quad (13)$$

$$C = A_0 \left(\frac{\lambda_A}{\lambda_C - \lambda_A} \frac{\lambda_B}{\lambda_B - \lambda_A} e^{-\lambda_A t} \right.$$
$$+ \frac{\lambda_A}{\lambda_A - \lambda_B} \frac{\lambda_B}{\lambda_C - \lambda_B} e^{-\lambda_B t}$$
$$\left. + \frac{\lambda_A}{\lambda_A - \lambda_C} \frac{\lambda_B}{\lambda_B - \lambda_C} e^{-\lambda_C t} \right) \quad (14)$$

$C\lambda_C, \ldots$ (**Fig. 4**). General equations describing the amounts and activities of any number of radioactive decay products are more complicated and are given in standard texts.

Radioactive equilibrium. In Fig. 4 the ratio $B\lambda_B/A\lambda_A$ of the activities of the parent A and the daughter product B change with time. The activity $B\lambda_B$ is zero initially and also after a very long time, when all the atoms have decayed. Thus $B\lambda_B$ passes through a maximum value, and it can be shown that this occurs at a time t_m given by Eq. (15). The situation in which

$$t_m = \frac{\ln(\lambda_B/\lambda_A)}{\lambda_B - \lambda_A} \quad (15)$$

the activities $A\lambda_A$ and $B\lambda_B$ are exactly equal to each other is called ideal equilibrium, and exists only at the moment t_m.

If the parent A is longer-lived than the daughter B, as occurs in many cases, then at a time which is long compared with the mean life τ_B of B, the

activity ratio approaches a constant value given by Eq. (16), where T_A and T_B are the half-periods of A

$$\frac{B\lambda_B}{A\lambda_A} = \frac{\lambda_B}{\lambda_B - \lambda_A} = \frac{T_A}{T_A - T_B} \quad (16)$$

and of B. When the activity ratio $B\lambda_B/A\lambda_A$ is constant, a particular type of radioactive equilibrium exists. This is spoken of as secular equilibrium if the activity ratio is experimentally indistinguishable from unity, as occurs when T_A is very much greater than T_B.

Equilibrium concepts are applied also between a long-lived parent and any of its decay products in a long series. For example, in a sufficiently old uranium ore, radium ($T = 1620$ years) is in secular equilibrium with its ultimate parent uranium ($T = 4.5 \times 10^9$ years) although there are four intermediate radioactive substances intervening in the series between uranium and radium. Here, secular equilibrium shows that activities of radium and uranium continue to be equal to each other even though the activity of the parent uranium is decreasing with time.

When T_B is comparable with T_A, Eq. (16) shows that the equilibrium ratio will clearly exceed unity; this situation is spoken of as transient equilibrium. For example, in fission-product decay series (17) the

$$^{140}\text{Ba} \rightarrow {}^{140}\text{La} \rightarrow {}^{140}\text{Ce} \quad (17)$$

half-period of ^{140}Ba is 307 h and that of ^{140}La is 40 h. In an initially pure source of ^{140}Ba the activity of ^{140}La starts at zero, rises to a maximum at $t_m = 135$ h [Eq. (15)], then decreases, and after a few hundred hours is in transient equilibrium with its parent, when the ^{140}La activity [by Eq. (16)] is $307/(307-40) = 1.15$ times the activity of its parent ^{140}Ba.
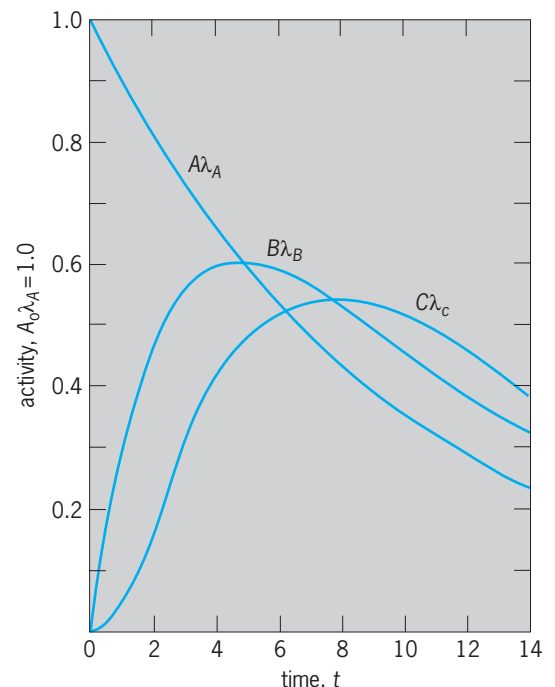

Fig. 4. Growth and decay of the activity $B\lambda_B$ of the daughter product, and $C\lambda_C$ of the granddaughter product, in an initially pure source of a radioactive parent whose activity at $t = 0$ is $A_0\lambda_A$.

TABLE 2. Parent radioactive nuclides found in nature

Nuclide		Percent abundance in nature	Half-period, years	Radioactive transitions observed*	Disintegration energy, MeV*
Atomic number, Z	Mass number, A				
19 K	40	0.0117	1.3×10^9	β^-, β^+, EC	β^- 1.3 – EC 1.5
34 Se	82	9.19	1.1×10^{20}	$\beta^-\beta^-$	3.0
37 Rb	87	27.83	4.8×10^{10}	β^-	0.3
48 Cd	113	12.2	9×10^{15}	β^-	0.3
49 In	115	95.77	5.1×10^{14}	β^-	0.5
52 Te	130	34.49	2×10^{21}	Growth of $^{130}_{54}$Xe†	1.6
57 La	138	0.089	1.1×10^{11}	β^-, EC	β^- 1.0 – EC 1.75
60 Nd	144	23.8	2.1×10^{15}	α	1.9
62 Sm	147	15.07	1.1×10^{11}	α	2.3
62 Sm	148	11.3	8×10^{15}	α	1.99
64 Gd	152	0.20	1.1×10^{14}	α	2.2
71 Lu	176	2.6	3.6×10^{10}	β^-, γ	0.6
72 Hf	174	0.16	2×10^{15}	α	2.5
75 Re	187	62.6	4×10^{10}	β^-	0.003
78 Pt	190	0.013	6×10^{11}	α	3.24
90 Th	232	100	1.4×10^{10}	α	4.08
92 U	235	0.715	7.0×10^8	α, SF	α 4.68
92 U	238	99.28	4.5×10^9	α, SF	α 4.27

*EC = electron capture. The EC energy is between ground states, but, in ^{138}La decay, 1.44 MeV of the EC energy goes to a gamma ray that feeds the ground state.
†Indirect evidence for $\beta^-\beta^-$ decay.

Radioactivity in the Earth. A number of isotopes of elements found in the Earth are radioactive (**Table 2**). All known or theoretically predicted isotopes of elements above bismuth are radioactive. Because the Earth is composed of atoms which were believed to have been created more than 3×10^9 years ago, the naturally occurring parent radioactive isotopes are those which have such long half-periods that detectable residual activity is still observable today. As a general rule, one can detect the presence of a radioactive substance for about 10 half-lives. Therefore activities with $T \lesssim 0.3 \times 10^9$ years should not be found in the Earth. For example, present-day uranium is an isotopic mixture containing 99.3% ^{238}U, whose half-period is 4.5×10^9 years, and only 0.7% of the shorter-lived uranium isotope ^{235}U, whose half-period is 0.7×10^9 years, whereas these isotopes presumably were produced in roughly equal amounts in the Earth a few billion years ago. Geophysical evidence indicates that originally some ^{236}U was present also, but none is found in nature now as expected with its half-period of 0.02×10^9 years. The elements technetium ($Z = 43$) and promethium ($Z = 61$) are not found in the Earth's crust because all their isotopes are radioactive with much shorter half-periods (their longest-lived are $T = 2.6 \times 10^6$ years for ^{97}Tc and $T = 17.7$ years for ^{145}Pm).

Uranium-238 decays through a long series of 14 radioactive decay products before ending as a stable isotope of lead, ^{206}Pb. Some of these members of the ^{238}U decay chain have very short half-periods, so their existence in nature is entirely dependent on the presence of their long-lived parent, and thus is a genealogical accident. For example, radium occurs in nature only in the minerals of its parent, uranium. The decay series of ^{235}U supports 14, and the decay series of ^{232}Th supports 10, short-lived radioactive substances found in nature.

A few of the common elements contain long-lived, naturally radioactive isotopes. For example, all terrestrial potassium contains 0.012% of the radioactive isotope ^{40}K, which has a half-period of 1.3×10^9 years, and emits negatron or positron beta particles (plus decay via electron capture) and gamma rays in a dual decay to stable ^{40}Ca and ^{40}Ar. This isotope is the principal source of radioactivity in a normal human being; each human contains about 0.1 microcurie (3.7×10^3 becquerels) of the radioactive potassium isotope ^{40}K.

Geological age measurements are based on the accumulation of decay products of long-lived isotopes, especially in the cases of ^{40}K, ^{87}Rb, ^{232}Th, ^{235}U, and ^{238}U. See GEOCHRONOMETRY.

Laboratory-produced radioactive nuclei. With particle accelerators and nuclear reactors, the order of 2000 radioactive isotopes not found in detectable quantities in the Earth's crust have been produced in the laboratory since 1935, including those of 26 new chemical elements up to element 118 (as of 2004), with element 117 not known. Earlier titles of induced or artificial radioactivities for these isotopes are misnomers. Many of these now have been identified in meteorites and in stars, and others are produced in the atmosphere by cosmic rays. There are over 5000 isotopes theoretically predicted to exist. As one approaches the place where a proton or neutron is no longer bound in a nucleus of an element (the limits of the existence of that element), the half-periods become extremely short. See TRANSURANIUM ELEMENTS.

For example, carbon-14 is a negatron beta-particle emitter, with a half-period of about 5600 years, which can be produced in the laboratory as the

product of a variety of different nuclear transmutation experiments. Nuclear bombardment of ^{11}B nuclei by alpha particles (helium nuclei) can produce excited compound nuclei of ^{15}N which promptly emit a proton (hydrogen nucleus), leaving ^{14}C as the end product of the transmutation. The same end-product ^{14}C can be produced by bombarding ^{14}N with neutrons, resulting in nuclear reaction (18).

$$^{14}\text{N} + \text{neutron} \rightarrow {}^{15}\text{N}^* \rightarrow {}^{14}\text{C} + \text{proton} \qquad (18)$$
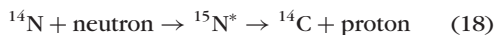
This reaction is easily carried out by using neutrons from nuclear accelerators or a nuclear reactor. This particular transmutation reaction is one which occurs in nature also, because the nitrogen in the Earth's atmosphere is continually bombarded by neutrons which are produced by cosmic rays, thus producing radioactive ^{14}C. Mixing of ^{14}C with stable carbon provides the basis for radiocarbon dating of systems that absorb carbon for times up to about 50,000 years ago (10 half-lives). *See* COSMIC RAYS; NUCLEAR REACTION; NUCLEAR REACTOR; PARTICLE ACCELERATOR; RADIOCARBON DATING.

Radioactive hydrogen, ^{3}H, is also formed in the atmosphere from the ^{14}N + neutron → ^{12}C + ^{3}H reaction. Also, ^{3}H is produced in the Sun, and the Earth's water as well as satelleites show an additional concentration of ^{3}H from the Sun. Over two dozen radioactive products, ranging in half-life from a few days to millions of years, have been identified in meteorites that have fallen to Earth. The carbon and hydrogen burning cycles that produce energy for stars produce radioactive ^{13}N, ^{15}O, ^{3}H. At higher temperatures the radioactivities ^{7}Be and even ^{8}Be ($T \approx 10^{-16}$ s) help burn hydrogen and helium. In addition to the production of radioactive as well as stable isotopes prior to the formation of the solar system, nucleosynthesis continues to go on in stars with the production of many short-lived radioactive atoms by different processes. *See* CARBON-NITROGEN-OXYGEN CYCLES; NUCLEOSYNTHESIS; PROTON-PROTON CHAIN.

The yield of any radioactivity produced in the laboratory is the initial rate of the activity under the particular conditions of nuclear bombardment. When a target material A is bombarded to produce a radioactive product B whose radioactive decay constant is λ_B, the number of atoms B which are present after a bombardment of duration t, and their activity $B\lambda_B$, are given by Eq. (19), where the yield Y has dimen-

$$B\lambda_B = \frac{Y}{\lambda_B}(1 - e^{-\lambda_B t}) \qquad (19)$$

sions equivalent to curies of activity produced per second of bombardment. The yield Y depends on the number of atoms A present in the target, the intensity of the beam of bombarding particles, and the cross section, or probability of the reaction per bombarding particle under the conditions of bombardment.

Radioactive transformation series. As noted in Eqs. (12)–(14), many radioactive substances have decay products which are also radioactive. Thus many long chains or series of radioactive transformations are known. The three naturally occurring transformation series are headed by ^{232}Th, ^{235}U, and ^{238}U (**Fig. 5** and **Table 3**).

Each of the naturally occurring radioactive isotopes in these transformation series has two synonymous names. For example, the commercially important radioisotope whose classical name is mesothorium-1 is now known to be an isotope of radium with mass number of 228 and is designated as radium-228 (^{228}Ra). Table 3 summarizes the names, symbols, and some radioactive properties of these three transformation series. However, these chains are not complete, and their uniqueness or importance as chains is an accident of the very long half-lives of ^{232}Th, ^{235}U, and ^{238}U. For example, element 105 of mass 260 has a succession of seven alpha decays and one electron capture and positron decay to ^{232}Th. The special importance of the chains in Table 3 is related to the fact that they were essentially the only early sources of radioactive materials, and they also play a role in nuclear power. *See* NUCLEAR POWER.
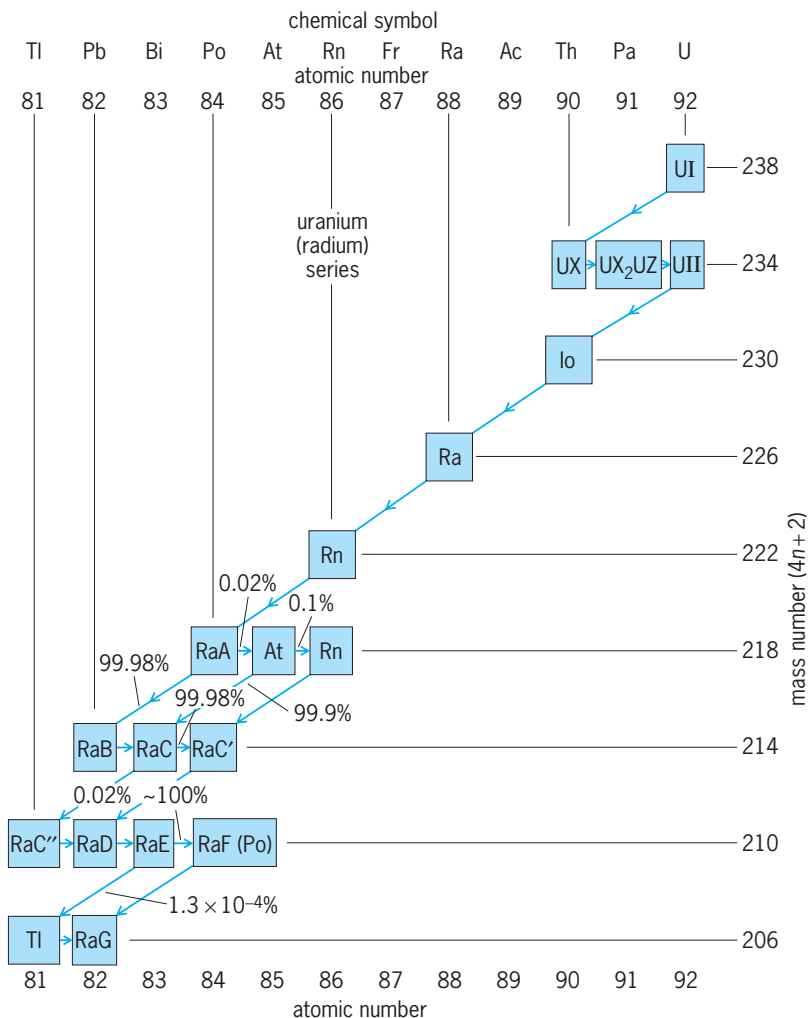

Fig. 5. Main line of decay of uranium series, or $4n + 2$ series, of heavy radioactive nuclides, headed in nature by uranium-238. Each member has a mass number given by $4n + 2$, where n is an integer.

TABLE 3. Names, symbols, and radioactive properties of members of the three naturally occurring radioactive transformation series*

Early name[†]	Early symbol[†]	Atomic number	Mass number	Isotopic symbol	Half-period	Type of decay
Uranium (4n + 2) series						
Uranium I	UI	92	238	^{238}U	4.5×10^9 y	α, SF
Uranium X$_1$	UX$_1$	90	234	^{234}Th	24 d	β^-
Uranium X$_2$	UX$_2$	91	234	^{234m}Pa	1.2 m	IT, β^-
Uranium Z	UZ	91	234	^{234}Pa	6.7 h	β^-
Uranium II	UII	92	234	^{234}U	2.5×10^5 y	α
Ionium	Io	90	230	^{230}Th	8×10^4 y	α
Radium	Ra	88	226	^{226}Ra	1600 y	α
Radon	Rn	86	222	^{222}Rn	3.8 d	α
Radium A	RaA	84	218	^{218}Po	3.0 m	α, β^-
Astatine	At	85	218	^{218}At	1.5 s	β^-, α
Radon	Rn	86	218	^{218}Rn	3.5 ms	α
Radium B	RaB	82	214	^{214}Pb	27 m	β^-
Radium C	RaC	83	214	^{214}Bi	20 m	β^-, α
Radium C′	RaC′	84	214	^{214}Po	1.6×10^{-4} s	α
Radium C″	RaC″	81	210	^{210}Tl	1.3 m	β^-
Radium D	RaD	82	210	^{210}Pb	22 y	β^-
Radium E	RaE	83	210	^{210}Bi	5.0 d	β^-, α
Radium F	RaF	84	210	^{210}Po	138 d	α
Thallium	Tl	81	206	^{206}Tl	4.2 m	β^-
Radium G	RaG	82	206	^{206}Pb	Stable	Stable
Thorium (4n) series						
Thorium	Th	90	232	^{232}Th	1.4×10^{10} y	α
Mesothorium$_1$	MsTh$_1$	88	228	^{228}Ra	5.8 y	β^-
Mesothorium$_2$	MsTh$_2$	89	228	^{228}Ac	6.1 h	β^-
Radiothorium	RdTh	90	228	^{228}Th	1.9 y	α
Thorium X	ThX	88	224	^{224}Ra	3.7 d	α
Thoron	Tn	86	220	^{220}Rn	56 s	α
Thorium A	ThA	84	216	^{216}Po	0.15 s	α
Thorium B	ThB	82	212	^{212}Pb	10.6 h	β^-
Thorium C	ThC	83	212	^{212}Bi	1.0 h	β^-, α
Thorium C′	ThC′	84	212	^{212}Po	3×10^{-7} s	α
Thorium C″	ThC″	81	208	^{208}Tl	3.1 m	β^-
Thorium D	ThD	82	208	^{208}Pb	Stable	Stable
Actinium (4n + 3) series						
Actinouranium	AcU	92	235	^{235}U	7.0×10^8 y	α, SF
Uranium Y	UY	90	231	^{231}Th	26 h	β^-
Protactinium	Pa	91	231	^{231}Pa	3.3×10^4 y	α
Actinium	Ac	89	227	^{227}Ac	22 y	β^-, α
Radioactinium	RdAc	90	227	^{227}Th	19 d	α
Actinium K	AcK	87	223	^{223}Fr	22 m	β^-, α
Actinium X	AcX	88	223	^{223}Ra	11 d	α
Astatine	At	85	219	^{219}At	0.9 m	α, β^-
Actinon	An	86	219	^{219}Rn	4.0 s	α
Bismuth	Bi	83	215	^{215}Bi	7.6 m	β^-
Actinium A	AcA	84	215	^{215}Po	1.8×10^{-3} s	α
Actinium B	AcB	82	211	^{211}Pb	36 m	β^-
Actinium C	AcC	83	211	^{211}Bi	2.2 m	α, β^-
Actinium C′	AcC′	84	211	^{211}Po	0.5 s	α
Actinium C″	AcC″	81	207	^{207}Tl	4.8 m	β
Actinium D	AcD	82	207	^{207}Pb	Stable	Stable

*Radon-223 has been shown to have a very weak radioactive ^{14}C decay branch to ^{209}Pb. Several of the isotopes in these chains, such as $^{234,235,238}U$, ^{226}Ra, and others are predicted to have such very weak, heavy cluster decay branches.

[†]These were the names and symbols used before the different isotopes of these elements were known.

Transformation series are now known for every element in the periodic table except hydrogen. Chains of neutron-rich isotopes have been produced and studied among the products of nuclear fission. Heavy-ion-induced reactions and high-flux reactors have been used to extend knowledge of the elements beyond uranium. The elements from number 93 (neptunium) to 118 (as yet unnamed except for element 117), which have so far not been found on Earth, were made in the laboratory. Both proton- and heavy-ion-induced reactions have extended knowl-edge of chains and neutron-deficient isotopes of the stable elements.

Alpha-Particle Decay

Alpha-particle decay is that type of radioactivity in which the parent nucleus expels an alpha particle (a helium nucleus). The alpha particle is emitted with a speed of the order of 1 to 2×10^7 m/s (10^4 mi/s), that is, about 1/20 of the velocity of light.

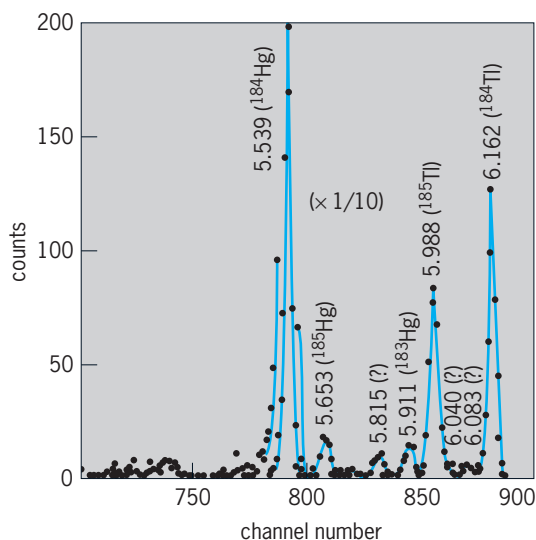In the simplest case of alpha decay, every alpha particle would be emitted with exactly the same

Fig. 6. Alpha groups in the decay of ^{184}Tl ($T = 11$ s) and ^{184}Hg and weak groups from ^{183}Hg and ^{185}Hg, very far off stability (17 neutrons less than the lightest stable thallium isotope). Energies in megaelectronvolts. (*After K. S. Toth et al., Observation of α-decay in thallium nuclei, including the new isotopes ^{184}Tl and ^{185}Tl, Phys. Lett., 63B:150–153, 1976*)

velocity and hence the same kinetic energy. However, in most cases there are two or more discrete energy groups called lines (**Fig. 6**). For example, in the alpha decay of a large group of ^{238}U atoms, 77% of the alpha decays will be by emission of alpha particles whose kinetic energy is 4.20 MeV, while 23% will be by emission of 4.15-MeV alpha particles. When the 4.20-MeV alpha particle is emitted, the decay product nucleus is formed in its ground (lowest energy) level. When a 4.15-MeV alpha particle is emitted, the decay product is produced in an excited level, 0.05 MeV above the ground level. This nucleus promptly transforms to its ground level by the emission of a 0.05-MeV gamma ray or alternatively by the emission of the same amount of energy in the form of a conversion electron and the associated spectrum of characteristic x-rays. Thus in all alpha-particle spectra, the alpha particles are emitted in one or more discrete and homogeneous energy groups, and alpha-particle spectra are accompanied by gamma-ray and conversion electron spectra whenever there are two or more alpha-particle groups in the spectrum.

Geiger-Nuttall rule. Among all the known alpha-particle emitters, most alpha-particle energy spectra lie in the domain of 4–6 MeV, although a few extend as low as 2 MeV ($^{147}_{62}$Sm) and as high as 10 MeV ($^{212}_{84}$Po or ThC′). There is a systematic relationship between the kinetic energy of the emitted alpha particles and the half-period of the alpha emitter. The highest-energy alpha particles are emitted by short-lived nuclides, and the lowest-energy alpha particles are emitted by the very long lived alpha-particle emitters. H. Geiger and J. M. Nuttall showed that there is a linear relationship between log λ and the energy of the alpha particle.

The Geiger-Nuttall rule is inexplicable by classical physics but emerges clearly from quantum, or wave,

mechanics. In 1928 the hypothesis of transmission through nuclear potential barriers, as introduced by G. Gamow and independently by R. W. Gurney and E. U. Condon, was shown to give a satisfactory account of the alpha-decay data, and it has been altered subsequently only in details. The form of the barrier-penetration equations is such that correlation plots of log λ against $1/E$ give nearly straight lines.

Nuclear potential barrier. At distances r which are large compared with the nuclear radius, the potential energy of an alpha particle, whose charge is $2e$, in the field of a residual nucleus, whose charge is $(Z - 2)e$, is $2(Z - 2)e^2/r$. At very close distances this electrostatic repulsion is opposed and overcome by short-range, nuclear, attractive forces. The net potential energy U as a function of the separation r between the alpha particle and its residual nucleus is the nuclear potential barrier.

One of several operating definitions of the nuclear radius R is the distance $r = R$ at which the attractive nuclear forces just balance the repulsive electrostatic forces. At this distance, called the top of the nuclear barrier, the potential energy is about 25–30 MeV for typical cases of heavy, alpha-emitting nuclei (**Fig. 7**).

Inside the nucleus the alpha particle is represented as a de Broglie matter wave. According to wave mechanics, this wave has a very small but finite probability of being transmitted through the nuclear potential energy barrier and thus of emerging as an alpha particle emitted from the nucleus. The transmission of a particle through such an energy barrier is completely forbidden in classical electrodynamics but is possible according to wave mechanics. This transmission of a matter wave through an energy barrier is analogous to the familiar case of the transmission of ordinary visible light through an opaque metal such as gold: if the gold is thin enough, some light does get through, as in the case of the thin gold leaf which is sometimes used for lettering signs on store windows. *See* QUANTUM MECHANICS.

The wave-mechanical probability of the transmission of an alpha particle through the nuclear
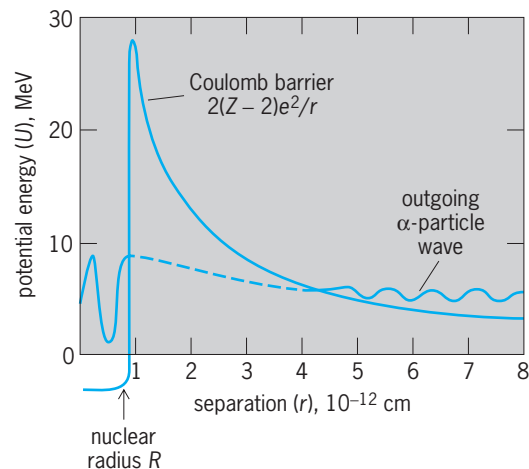

Fig. 7. Schematic of nuclear potential barrier, illustrating emission of an alpha particle as a wave which can be transmitted through the barrier.

potential barrier is very strongly dependent upon the energy of the emitted alpha particle. Analytically the probability of transmission T depends exponentially upon a barrier transmission exponent γ according to Eq. (20). To a good approximation, Eq. (21) holds,

$$T = e^{-\gamma} \qquad (20)$$

$$\gamma = \left(\frac{4\pi^2}{h}\right)\frac{(Z-2)2e^2}{V}$$

$$- \left(\frac{8\pi}{h}\right)[2(Z-2)2e^2MR]^{1/2} \quad (21)$$

where $h = 6.626 \times 10^{-34}$ joule-second is Planck's constant, and M is the so-called reduced mass of the alpha particle. For the alpha decay of ^{226}Ra, the numerical value of γ is about 71: hence $T = e^{-71} = 10^{-31}$. The first term on the right side of Eq. (21) is about 154 and is therefore the dominant term. When this term is taken alone, $e^{-(4\pi^2/h)(Z-2)2e^2/V}$ is called the Gamow factor for barrier penetration.

Inspection of Eq. (21) shows that the barrier transmission decreases with increasing nuclear charge $(Z-2)e$, increases with increasing velocity V of emission of the alpha particle, and increases with increasing radius R of the nucleus. When the experimentally known values of alpha-decay energy are substituted into Eq. (21), with R about 10^{-12} cm and Z about 90, the transmission coefficient $T = e^{-\lambda}$ is found to extend over a domain of about 10^{-20} to 10^{-40}. This range of about 10^{20} is just what is needed to relate the alpha-disintegration energy to the broad domain of known alpha-decay half-periods. Equation (21) thus explains the Geiger-Nuttall rule very successfully (**Fig. 8**). From Eq. (20), one can derive the relationship between the mean life τ in seconds and the alpha-particle energy E_α in MeV, $\ln \tau = AE_\alpha^{-1/2} + B$, where A and B are constants for different parent nuclei.

Since 1970, knowledge of alpha-emitting isotopes has been greatly enlarged through the identification of many isotopes far off stability in the region just above tin and in the broad region from neodymium all the way to uranium. For example, fusion reactions between 290-MeV ^{58}Ni ions and ^{58}Ni and ^{63}Cu targets have been used to produce and study very neutron deficient radioactive isotopes, including 12 alpha emitters between tin and cesium. These results provide important data on the atomic masses of nuclei far from the stable ones in nature. These data test understanding of nuclear mass formulas and their validity in new regions of the periodic table.

Beta-Particle Decay

Beta-particle decay is a type of radioactivity in which the parent nucleus emits a beta particle. There are two types of beta decay established: in negatron beta decay (β^-) the emitted beta particle is a negatively charged electron (negatron); in positron beta decay (β^+) the emitted beta particle is a positively charged electron (positron). In beta decay the atomic number shifts by one unit of charge, while the mass number remains unchanged (Table 1). In contrast to
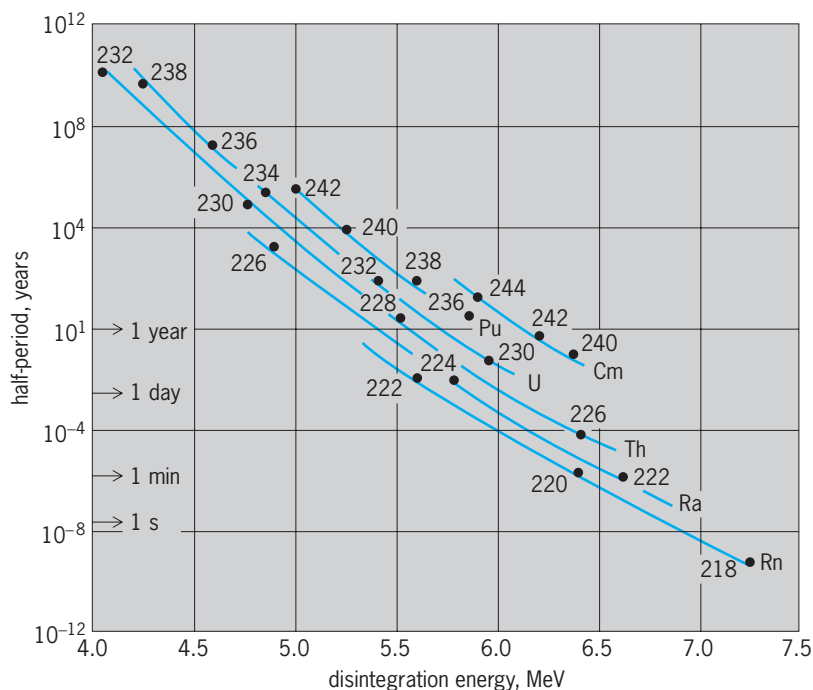

Fig. 8. **Systematics of the broad range of half-periods for alpha-particle decay and their strong dependence on alpha-decay energy and weaker dependence on nuclear charge. Numbers beside experimental points are mass numbers of parent alpha-particle emitters. Lines connect parent isotopes and are drawn using wave-mechanical theory of alpha-particle transmission through nuclear potential barriers.**

alpha decay, when beta decay takes place between two nuclei which have a definite energy difference, the beta particles from a large number of atoms will have a continuous distribution of energy (**Fig. 9**). *See* POSITRON.

For each beta-particle emitter, there is a definite maximum or upper limit to the energy spectrum of beta particles. This maximum energy, E_{max}, corresponds to the change in nuclear energy in the beta decay. Thus $E_{max} = 0.57$ MeV for β^- decay of ^{64}Cu, and $E_{max} = 0.66$ MeV for β^+ decay of ^{64}Cu. For positron decay to occur, the total decay energy must exceed 1.022 MeV (twice the rest energy of the electron). The total decay energy for β^+ decay is then $E_{max}(\beta^+)$ plus 1.022 MeV. As in the case of alpha decay, most beta-particle spectra are not this simple, but include additional continuous spectra which have less maximum energy and which leave the product nucleus in an excited level from which gamma rays are then emitted.

For nuclei very far from stability, the energies of these excited states populated in beta decay are so large that the excited states may decay by proton, two-proton, neutron, two-neutron, three-neutron or alpha emission, or spontaneous fission. In some cases, the energies are so great that the number of excited states to which beta decay can occur is so large that only the gross strength of the beta decays to many states can be studied.

Neutrinos. The continuous spectrum of beta-particle energies (Fig. 9) implies the simultaneous emission of a second particle besides the beta particle, in order to conserve energy and angular momentum for each decaying nucleus. This particle is
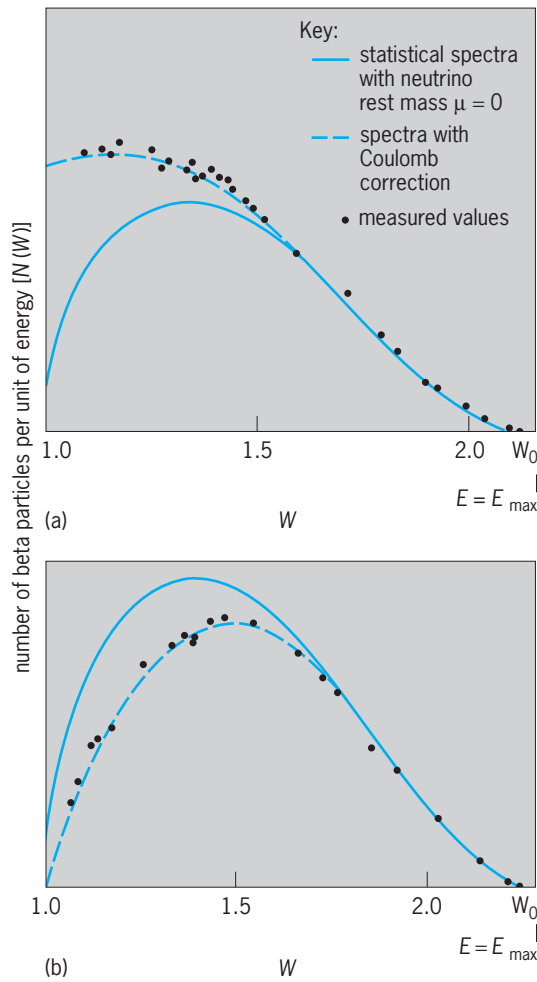
Fig. 9. Spectra of beta particles. The quantity *W* is the total beta-particle energy in units of the electron rest energy. (*a*) β^- decay of ^{64}Cu. (*b*) β^+ decay of ^{64}Cu. (*After L. M. Langer, R. D. Moffat, and H. C. Price, The beta-spectrum of ^{64}Cu, Phys. Rev., 76:1725–1726, 1949*)

time of this transition, two particles, the positron β^+ and the neutrino ν, are created and emitted. The emitted β^+ and ν together carry away the energy $E_{\max}$ of the transition and provide for conservation of energy, momentum, angular momentum, charge, and statistics. Thus positron beta decay is represented by decay (22). Negatron beta decay is a closely

$$p \rightarrow n + \beta^+ + \nu \qquad (22)$$

related process, except that a neutron n changes to a proton p in the nucleus, and a negatron beta particle β^- and its characteristic companion particle, the antineutrino $\bar{\nu}$, are emitted, as in decay (23). The antineutrino is the antiparticle of the neu-

$$p \rightarrow n + \beta^- + \bar{\nu} \qquad (23)$$

trino as the β^+ is the antiparticle of the β^-. The ν and $\bar{\nu}$ have the same properties of zero charge and essentially zero rest mass, and differ only with respect to the direction of alignment of their intrinsic spin along their direction of motion. In most beta-decay contexts, the term "neutrino" includes both its forms, neutrino and antineutrino. *See* ANTIMATTER.

There are, in fact, three classes of neutrinos. The neutrinos emitted in the two types of beta decay [decays (22) and (23)] are called electron neutrinos. In addition, there is a neutrino and antineutrino associated with the mu meson ($\mu\pm$) and neutrinos and

the neutrino. The sum of the kinetic energy of the neutrino and the beta particle equals $E_{\max}$ for the particular transition involved except in the rare cases where internal bremsstrahlung or shake-off electrons are emitted along with the beta particle and neutrino. The neutrino has zero charge and nearly zero rest mass, travels at essentially the same speed as light (3×10^8 m/s or 1.86×10^5 mi/s), and is emitted as a companion particle with each beta particle.

Earlier careful measurements of the beta spectra of ^{3}H established an upper limit for the neutrino rest mass as less than 0.0005 times the rest energy of the electron (**Fig. 10**). In 1980, however, some indirect evidence and a new ^{3}H beta spectra measurement yielded evidence for a rest mass much smaller than this limit, but finite. If the neutrino does have a nonzero rest mass, this will have many consequences, such as the size of the total mass of the universe but will not radically change the general features of the beta decay as presented here.

Two forms of neutrinos are distinguished in beta decay. In positron beta decay, a proton p in the nucleus transforms into a neutron n in the nucleus, thus reducing the nuclear charge by 1 unit. At the
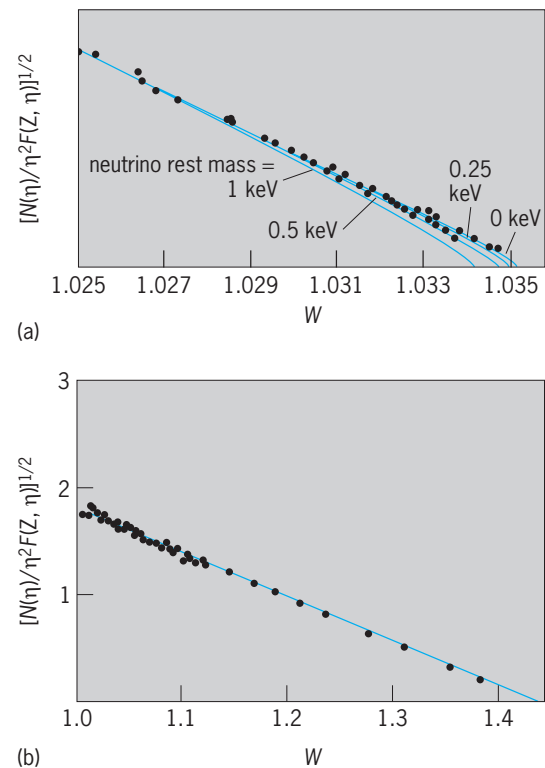

Fig. 10. Kurie plots. The quantity *W* is the total beta-particle energy in units of the electron rest energy; other quantities are defined in text. (*a*) Allowed decay of ^{3}H (*after L. M. Langer and R. J. D. Moffat, The beta-spectrum of tritium and the mass of the neutrino, Phys. Rev., 88:689–694, 1952*). (*b*) Once-forbidden decay of ^{147}Pm (*after J. H. Hamilton, L. M. Langer, and W. G. Smith, The shape of the ^{143}Pr spectrum, Phys. Rev., 112: 2010–2019, 1958*).

antineutrinos associated with the tau ($\tau\pm$). The mu neutrinos are well established. The first evidence for a tau neutrino was reported in 1998. These three particles, electron, mu, and tau, together with their neutrinos, ν_e, ν_μ, and ν_τ, and their respective antiparticles, make up a class of particles called leptons. The number of leptons in a decay or reaction is considered to be conserved; this rule is called lepton number conservation. There are separate conservation numbers for e, μ, and τ. There must be the same net number of each type of lepton on each side of a decay or reaction. For example, since there are no leptons on the left sides of decay (22) and (23), there must be no net leptons on the right sides. Thus decay (22) has on the right side an antielectron (positron) with lepton number $L = -1$ and a neutrino with lepton number $L = 1$. Decay (23) has an electron with $L = 1$ and an antineutrino with $L = -1$, and so the lepton number is $L = 1 - 1 = 0$ leptons on the right side as well. Decays (22) and (23) also conserve nucleon number. *See* LEPTON.

Because of the fact that the neutron rest mass is greater than the proton rest mass, free neutrons can undergo beta decay [decay (23)], but protons must use part of the nuclear binding energy available inside a nucleus to make up the rest mass difference in decay (22).

The interaction of neutrinos with matter is exceedingly feeble. A neutrino can pass all the way through the Sun with little chance of collision. The thickness of lead required to attenuate neutrinos by the factor $^1/_2$ is about 10^{18} m (10^{15} mi), or 100 light-years of lead!

Neutrinos play critical roles in understanding other fundamental questions. One question concerns the rest mass of the neutrino. The current upper limit on the rest-mass energy of the electron neutrino is 2 eV (the rest energy of an electron is 511,000 eV). However, since there are so many neutrinos in the universe (every second every square centimeter of a person's body is bombarded by 10^{10} neutrinos from the Sun), even this small mass can influence the total mass in the universe and whether there is sufficient mass to cause the expansion of the universe to stop and contract. The nature of the neutrino is also important in grand unified theories as outlined below. *See* COSMOLOGY; NEUTRINO; UNIVERSE.

Average beta energy. Charged particles, such as beta particles or alpha particles, are easily absorbed in matter, and their kinetic energy is thereby converted into heat. In beta decay the average energy E_{av} of the beta particles is far less than the maximum energy E_{max} of the particular beta-particle spectrum. The detailed shape of beta-particle spectra and hence the exact value of the ratio E_{av}/E_{max} varies somewhat with Z, E_{max}, the degree of forbiddenness of the transition, and the sign of charge of the emitted beta particle. A rough rule of thumb which covers many practical cases is $E_{av} = (0.40 \pm 0.05)E_{max}$, with slightly higher values for positron beta-particle spectra than for negatron beta-particle spectra. The remaining disintegration energy is emitted as kinetic

energy of neutrinos and is not recoverable in finite absorbers.

There are other processes that carry off part of the energy of beta decay, including internal bremsstrahlung (gamma rays) and shake-off electrons (atomic electrons). The total probabilities for these additional two processes are on the order of 1% or much less per beta decay, and the probability of their emission decreases rapidly with increasing energy so they are mainly low-energy (less than about 50 keV) radiations. In internal bremsstrahlung, through an interaction of the beta particle and the emitting nucleus, part of the decay energy is emitted as a gamma ray. In the shake-off process, part of the beta-decay energy is given to one of the atomic electrons. The gamma rays are not absorbed in matter as easily as the beta particles. In addition, if one tries to absorb the beta particles in matter, the beta particles can interact with the atoms and give off external bremsstrahlung (gamma rays). The number of these gamma rays again is a strongly decreasing function of energy, but their emission extends up to the maximum energies of the beta particles. *See* BREMSSTRAHLUNG.

Fermi theory. By postulating the simultaneous emission of a beta particle and a neutrino, as in reaction (22), E. Fermi developed in 1934 a quantum-mechanical theory which satisfactorily gives the shape of beta-particle spectra (Fig. 9), and the relative half-periods of beta-particle emitters for allowed beta decays. The energy distribution of beta particles in allowed transitions is then given by Eq. (24). Here

$$N(W)\,dW$$

$$= \frac{|P|^2}{\tau_0} F(Z, W)(W^2 - 1)^{1/2}(W_0 - W)^2 W\,dW \quad (24)$$

$N(W)\,dW =$ number of beta particles in energy range W to $W + dW$; $W = 1 + E/(m_0c^2) =$ total energy of beta particle in units of rest energy $m_0c^2 = 0.51$ MeV for an electron ($m_0 =$ electron mass, $c =$ velocity of light); $W_0 = 1 + E_{max}/(m_0c^2) =$ maximum energy of the beta-particle spectrum; $|P|^2 =$ squared matrix element for the transition, and is of the order of unity for allowed transitions; $\tau_0 =$ time constant $\cong 7000$ s; $F(Z,W) =$ complex, dimensionless function involving the nuclear radius, nuclear charge, beta-particle energy, and whether the decay is β^- or β^+.

Physically this distribution function involves the product of the energy W and momentum $(W^2 - 1)^{1/2}$ of the beta particle times the energy $(W_0 - W)$ and the momentum $(W_0 - W)/c$ of the neutrino. The product of these factors gives a "statistical" distribution for the number of beta particles as a function of energy (Fig. 9). The observed spectra show an excess of low-energy β^- and a deficiency of low-energy β^+ particles. This arises because of the Coulomb attraction and repulsion of the nucleus for β^- and β^+. The statistical spectrum is corrected by the Fermi function, $F(Z,W)$, and the new distribution agrees with experiments (Fig. 9).

Equation (24) essentially matches the energy spectra of allowed beta-particle transitions and therefore

furnishes one type of experimental verification of the properties of neutrinos. Its counterpart in terms of the beta-particle momentum spectrum is often used for the analysis of spectra, and is given by Eq. (25).

$$N(\eta)d\eta = \frac{|P|^2}{\tau_0}F(Z, \eta)(W_0 - W)^2\eta^2d\eta \qquad (25)$$

The momentum distribution is much more nearly symmetric than its corresponding energy spectrum. Here $N(\eta)d\eta$ = number of beta particles in the momentum interval from η to $\eta + d\eta$; $\eta = (W^2 - 1)^{1/2}$ = momentum of the beta particle in units of m_0c; and $F(Z,\eta) = F(Z,W)$ of Eq. (24).

Konopinski-Uhlenbeck theory. After the work of Fermi which explained allowed decay, E. J. Konopinski and G. E. Uhlenbeck in 1941 developed the theory of forbidden beta decay. Allowed decays occur between nuclear states which differ in spin by 0 or 1 unit and which have the same parity. Konopinski and Uhlenbeck developed a theory to describe beta decays where energy is available for decay but the allowed selection rules on spin or parity or both are violated. These beta transitions occur at a slower rate and are called forbidden transitions. In 1949 the theory of forbidden beta decay was confirmed by L. M. Langer and H. C. Price. The orders of forbiddenness, which retard the rate of decay, are the following: once-forbidden decay when the change in nuclear spin ΔJ is again 0 or 1 as in allowed decay, but a parity change $\Delta\pi$ occurs; once-forbidden unique decay when $\Delta\pi$ changes and $\Delta J = 2$; n-times forbidden decay when $\Delta J = n$, $\Delta\pi = (-)^n$, where $\Delta\pi = -$ indicates a parity change; and n-times forbidden unique decay when $\Delta J = n + 1$, $\Delta\pi = (-)^n$ [**Table 4**]. In forbidden decays the first-order allowed matrix elements of the Fermi theory in Eq. (24) vanish because of the selection rules on angular momentum and spin. Then the much smaller higher-order matrix elements that can be neglected compared to the large allowed matrix elements come into play. *See* PARITY (QUANTUM MECHANICS); SELECTION RULES (PHYSICS).

Comparative half-lives, fT. The half-period T of beta decay can be derived from Eq. (24) because the radioactive decay constant $\lambda = 0.693/T$ is simply

the total probability of decay, or $N(W) dW$ integrated over all possible values of the beta-particle energy from $W = 1$ to $W = W_0$.

For allowed decays, the matrix elements are not functions of the beta energy and can be factored out of Eq. (24), so Eq. (26) is valid, where f is given by Eq. (27), and the constants include $|P|^2$ of Eq. (24). Equation (26) can be rearranged as Eq. (28).

$$\lambda = \frac{0.693}{T} = \text{constants} \times \text{f} \qquad (26)$$

$$\text{f} = \int_1^W F(Z, W)(W^2 - 1)^{1/2}(W_0 - W)^2 W \, dW \qquad (27)$$

$$\text{f}T = \frac{0.693}{\text{constants}} = \text{comparative half-life} \qquad (28)$$

For different beta decays, T varies over a range greater than 10^{18} and inversely depends on the beta-decay energy in analogy to the Geiger-Nuttall rule for alpha decay. However, Eq. (28) says that the comparative half-life should be a constant. Indeed it is found experimentally that different classes of beta decay do have very similar fT values. It is generally easier to give the $\log_{10}$ fT for comparison. The groups (Table 4) include, in addition to the forbidden decays, three classes of allowed decays: the favored or superallowed decays of nuclei whose structures are very similar so that the matrix element in the denominator of Eq. (28) is large and log fT is small; normal allowed; and allowed I-forbidden where the total angular momentum selection rule holds, but the individual particle that is undergoing beta decay has a change of 2 units of orbital angular momentum. The matrix elements for each degree of forbiddenness get progressively smaller and so log fT values increase sharply with each degree of forbiddenness. The ranges of these fT values for each degree of forbiddenness are in general so well established that measurements of fT values can be used to establish changes in spins and parities between nuclear states in beta decay.

Kurie plots. For allowed transitions, the transition matrix element $|P|^2$ is independent of the momentum η. Then Eq. (25) can be put in the form of Eq. (29).

$$\left(\frac{N(\eta)}{\eta^2 F(Z, \eta)}\right)^{1/2} = \text{const}(W_0 - W) \qquad (29)$$

Therefore a straight line results when the quantity $(N/\eta^2 F)^{1/2}$ is plotted against beta-particle energy, either as W or as E, on a linear scale. Such graphs are called Kurie plots, Fermi plots, or Fermi-Kurie plots. These are especially useful for revealing deviations from the allowed theory and for obtaining the upper energy limit E_{max} as the extrapolated intercept of $N/\eta^2 F$ on the energy axis. Practically all of the results on the shapes of beta-particle spectra are published as Kurie plots, rather than as actual momentum or energy spectra.

When spectral data give a straight line (Fig. 10), then $N(\eta)$ is in agreement with the Fermi momentum

TABLE 4. Selection rules for beta decay and log fT values

| Type	ΔJ	$\Delta \pi$	Log fT	Examples
Allowed (favored)	0 or 1	No	3	n, ^{3}H
Allowed (normal)	0 or 1	No	4 to 7	^{35}S, ^{30}P
Allowed (I-forbidden)	1	No	6 to 9	^{32}P, ^{65}Ni
Once-forbidden	0 or 1	Yes	6 to 8	^{111}Ag, ^{143}Pr
Once-forbidden (unique)	2	Yes	8 to 9	^{42}K, ^{91}Y
Twice-forbidden	2	No	11 to 14	^{36}Cl, ^{59}Fe
Twice-forbidden (unique)	3	No	12 to 14	^{22}Na, ^{60}Co
Third-forbidden	3	Yes	17 to 19	^{87}Rb, ^{138}La
Third-forbidden (unique)	4	Yes	(~18)	^{40}K
Fourth-forbidden	4	No	~ 24	^{115}In
Fourth-forbidden (unique)	5	No		
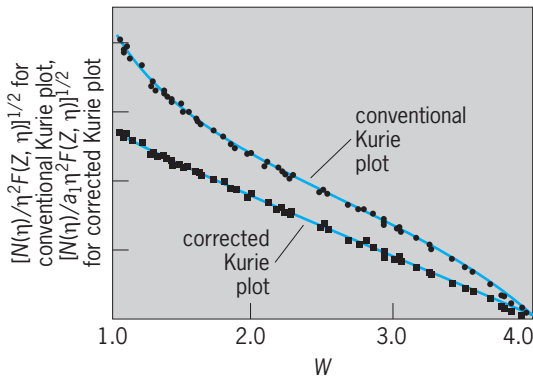
Fig. 11. Once-forbidden spectrum of ^{91}Y: conventional Kurie plot and Kurie plot corrected by the unique shape factor, $a_1 = W^2 - 1 + (W_0 - W)^2$, given by Konopinski and Uhlenbeck, which linearizes the data. The quantity W is the total beta-particle energy in units of the electron rest energy; other quantities are defined in text. (*After L. M. Langer and H. C. Price, Shape of the beta spectrum of the forbidden transition of yttrium 91, Phys. Rev., 75:1109, 76:641, 1949*)

distribution, Eq. (25); and the intercept of this straight line, on the energy axis, gives the disintegration energy E_{max}. For β^+ decay the total decay energy is E_{max} (β^+) plus 1.022 MeV. In Fig. 10a, theoretical curves are given for various values of the neutrino rest mass, and the data points, which are experimental values, lie on the curve corresponding to zero mass.

In addition to allowed decays, all but one known once-forbidden decays have Kurie plots that are essentially linear in energy (Fig. 10b). The once-forbidden unique decays have a pronounced characteristic energy dependence for their matrix elements, and thus the conventional Kurie plot has a characteristic shape that differs from a straight line (**Fig. 11**). When the data are corrected by the unique shape factor, given by Konopinski and Uhlenbeck, a linear Kurie plot is again obtained. This unique shape was the key to the discovery of forbidden beta decay by Langer and Price (Fig. 11). The higher-order forbidden spectra each show different strong energy dependences in their Kurie plots, each characteristic of their degree of forbiddenness.

Double beta decay. When the ground state of a nucleus differing by two units of charge from nucleus A has lower energy than A, then it is theoretically possible for A to emit two beta particles, either $\beta^+\beta^+$ or $\beta^-\beta^-$ as the case may be, and two neutrinos or antineutrinos, and go from Z to $Z \pm 2$. Here two protons decay into two neutrons, or vice versa. This is a second-order process and so should go much slower than beta decay. There are a number of cases where such decays should occur, but their half-lives are of the order of 10^{20} years or greater. Such decay processes are obviously very difficult to detect.

The first direct evidence for two-neutrino double-beta-minus decay of selenium, reaction (30), was

$$^{82}_{34}\text{Se} \rightarrow \,^{82}_{36}\text{Kr} + 2e^- + 2\bar{\nu} \qquad (30)$$
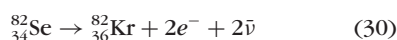
found only in 1987, some 40 years after the first attempt to observe this rare process. The trajectories of the electrons from this decay have been detected in a time-projection chamber. A half-period of $1.08^{+0.26}_{-0.06} \times 10^{20}$ y has been obtained. Previous geochemical and cosmochemical evidence of double beta decay had been only indirect because it consisted of observations of buildup of the noble gases $^{82}_{32}\text{Kr}$ and $^{130}_{54}\text{Xe}$ in the $^{84}_{32}\text{Se}$ and $^{130}_{52}\text{Te}$ samples, respectively. *See* TIME-PROJECTION CHAMBER.

Some unifications of the electroweak and strong forces suggest that the electron neutrino and antineutrino are identical ($\nu_e \equiv \bar{\nu}_e$), that is, that the neutrino is a Majorana particle. There is no experimental evidence at present for this. If true, this allows double beta decay to proceed without the emission of neutrinos. In this neutrinoless double beta decay, the first neutrino absorbs the second one so the decay is decay (31).

$$^{82}_{24}\text{Se} \longrightarrow \,^{82}_{36}\text{Kr} + 2e^- \qquad (31)$$
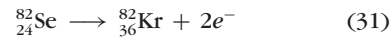
Searches have shown that the half-period of this decay is greater than 2.7×10^{22} y, which is over 200 times larger than the half-period of the ^{82}Se double beta decay with two-neutrino emission. Lepton number is not conserved in decay (31). This process is predicted to occur in some grand unified theories. The experimentally established lower limits of half-lives for neutrinoless double beta decays have been used to derive upper limits for the mass of the Majorana neutrino ($m_{\nu e}c^2 < 2$ eV) and for the right-handedness of the weak interaction (less than 2×10^{-8} of the left-handed strength). *See* GRAND UNIFICATION THEORIES.

Electron-capture transitions. Whenever it is energetically allowed by the mass difference between neighboring isobars, a nucleus Z may capture one of its own atomic electrons and transform to the isobar of atomic number $Z - 1$ (Table 1). Usually the electron-capture (EC) transition involves an electron from the K shell of atomic electrons, because these innermost electrons have the greatest probability density of being in or near the nucleus. *See* ELECTRON CAPTURE.

In EC transitions, a proton p bound in the parent nucleus absorbs an electron e^- and changes to a bound neutron n. The disintegration energy is carried away by an emitted neutrino ν as in transition (32). The residual nucleus may be left either

$$p + e^- \rightarrow n + \nu \qquad (32)$$
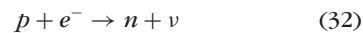
in its ground level or in an excited level from which gamma-ray emission follows.

EC transitions compete with all cases of positron beta-particle decay. EC has an energetic advantage over β^+ decay equivalent to the mass of two electrons, or 1.02 MeV, because in transition (32) one electron mass e^- enters on the left and is available, whereas in decay (22) one electron mass β^+ must be produced as a product of the positron beta-particle decay. For example, in the radioactive decay of $^{64}_{29}\text{Cu}$, twice as many transitions go by EC to $^{64}_{28}\text{Ni}$ as go by positron beta decay to the same decay

product. In the heavy, high-Z elements, EC is greatly favored over the competing β^+ decay, and examples of measurable β^+ decay are practically unknown for Z greater than 80, although there are a large number of examples of electron capture. As the energy for decay increases beyond 1.02 MeV, the probability of β^+ decay increases relative to EC and dominates at several megaelectronvolts of energy.

Several examples are known of completely pure EC radioactivity in which there is insufficient nuclear energy to allow any positron beta-particle decay (total decay energy is less than 1.022 MeV). For example, $^{55}_{26}$Fe emits no positron beta particles, but transforms with a half-period of 2.6 years entirely by EC to the ground level of $^{55}_{25}$Mn. This radioactivity is detectable through the K-series x-rays which are emitted from ^{55}Mn when the atomic electron vacancy, produced by nuclear capture of a K electron, refills from the L shell of atomic electrons. Also, double electron capture, analogous to double beta decay,

is theoretically predicted to exist. Here two atomic electrons are captured and two neutrinos emitted.

Gamma-Ray Decay

Gamma-ray decay involves a transition between two excited levels of a nucleus, or between an excited level and the ground level. A nucleus in its ground level cannot emit any gamma radiation. Therefore gamma-ray decay occurs only as a sequel of one of the processes in Table 1 or of some other process whereby the product nucleus is left in an excited state. Such additional processes include gamma rays observed following the fusion of two nuclei, as occurs in bombarding ^{58}Ni with ^{16}O to form an excited compound nucleus of ^{74}Kr. This compound nucleus first promptly gives off a few particles like two neutrons to leave ^{72}Kr* or two protons to leave ^{72}Se*, both of which will be in excited states which will emit gamma rays. Or one may excite states in a nucleus by the Coulomb force between two nuclei when they pass close to each other but do not touch (their separation is greater than the sum of the radii of the two nuclei). There are also other nuclear reactions such as induced nuclear fission that leave nuclei in excited states to undergo gamma decay. *See* COULOMB EXCITATION.

A gamma ray is high-frequency electromagnetic radiation (a photon) in the same family with radio waves, visible light, and x-rays. The energy of a gamma ray is given by $h\nu$, where h is Planck's constant and ν is the frequency of oscillation of the wave in hertz. The gamma-ray or photon energy $h\nu$ lies between 0.05 and 3 MeV for the majority of known nuclear transitions. Higher-energy gamma rays are seen in neutron capture and some reactions. *See* ELECTROMAGNETIC RADIATION.

Gamma rays carry away energy, linear momentum, and angular momentum, and account for changes of angular momentum, parity, and energy between excited levels in a given nucleus. This leads to a set of gamma-ray selection rules for nuclear decay and a classification of gamma-ray transitions as "electric" or as "magnetic" multipole radiation of multipole order 2^l where $l = 1$ is called dipole radiation, $l = 2$ is quadrupole radiation, and $l = 3$ is octupole, l being the vector change in nuclear angular momentum. The most common type of gamma-ray transition in nuclei is the electric quadrupole (E2). There are cases where several hundred gamma rays with different energies are emitted in the decays of atoms of only one isotope. *See* MULTIPOLE RADIATION.

Mean life for transitions. A reasonably successful approximate theory of the mean life for gamma-ray decay was developed by V. F. Weisskopf in 1951, using the single-particle shell model of nuclei (**Fig. 12**). An E2 transition of about 1 MeV is expected to take place with a mean life, τ_{el}, or mean delay in the upper level, of about 10^{-11} s. Thus most gamma-ray transitions are prompt transitions, in which the mean life of the excited level is too short to be measured easily. The mean life τ_{mag} for magnetic multipoles is of the order of 30 (for $A = 20$) to 150 (for
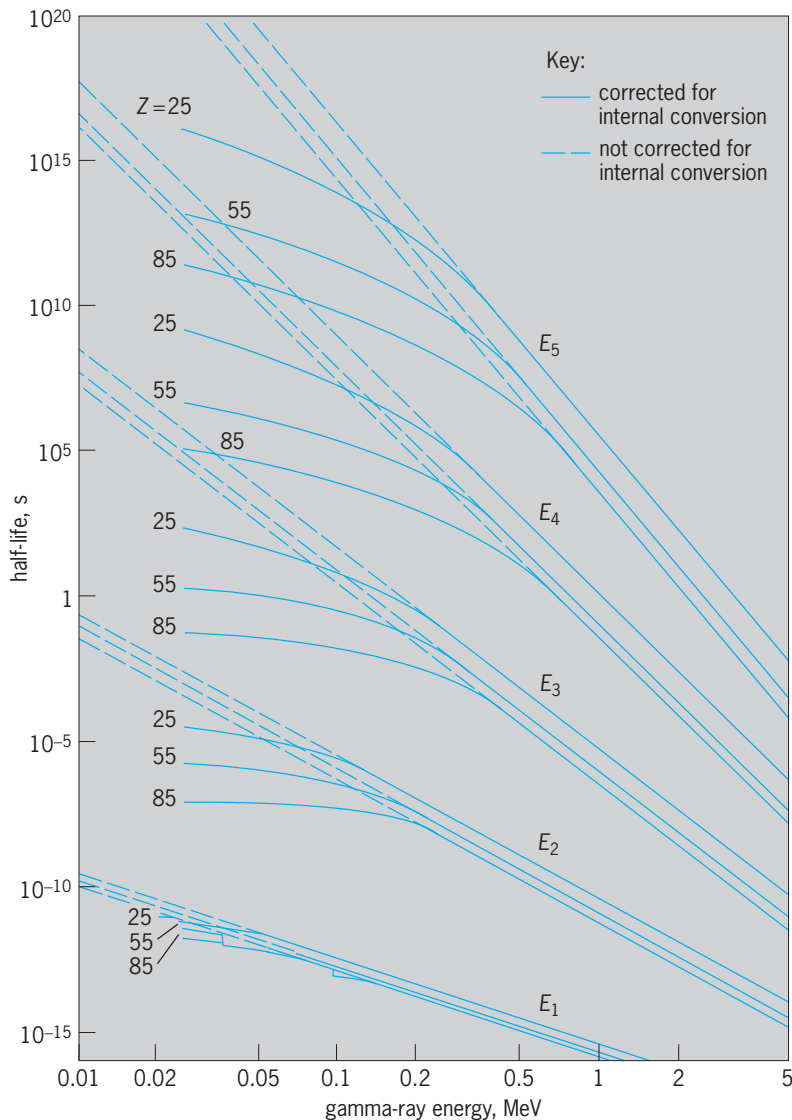

Fig. 12. Theoretical values of half-lives for decay of nuclear levels by emission of gamma rays and conversion electrons, for electric multipoles. (*After A. H. Wapstra, G. J. Nijgh, and R. van Lieshout, Nuclear Spectroscopy Tables, North-Holland, 1959*)

$A = 200$) times longer than τ_{el} for electric multipole transitions (Fig. 12).

At low energies or high Z, or both, the internal conversion process becomes a very important additional mode of decay that markedly shortens the mean lives of the nuclear levels. In addition, in many cases the structure of the nucleus comes into play and alters the observed mean lives considerably compared to those predicted by the Weisskopf theory (Fig. 12). Electric dipole (E1) transitions are generally retarded (longer mean lives) by factors of 10^6 over the Weisskopf estimates (Fig. 12). On the other hand, A. Bohr and B. Mottelson developed a model of collective nuclear motions where E2 transitions are enhanced by factors of 100 or more (shorter τ) over the Weisskopf single-particle estimates, and these predictions are confirmed by experiments. The magnetic dipole (M1) transitions are also often hindered by factors of 100 or more. Measurements of the mean lives for gamma-ray decay provide important tests of nuclear models.

Internal conversion. An alternative type of deexcitation which always competes with gamma-ray emission is known as internal conversion. Instead of the emission of a gamma ray, the nuclear excitation energy can be transferred directly to a bound electron of the same atom. Then the nuclear energy difference is converted to energy of an atomic electron, which is ejected from the atom with a kinetic energy E_i given by Eq. (33). Here B_i is the original atomic

$$E_i = W - B_i \qquad (33)$$

binding energy of the particular electron, which is ejected, and W is the nuclear transition energy which would otherwise have been emitted as a gamma-ray photon having energy $h\nu = W$.

The spectrum of internal conversion electrons is then a series of discrete energies, or "lines," each corresponding to an individual value of B_i, for the K, L (L_1, L_2, L_3), $M, \ldots$ electrons in each shell and subshell of the atom. Thus conversion electron spectra are much more complex than gamma spectra (**Fig. 13**). From the spacing of the E_i values in this conversion electron spectrum, it is possible to assign definitely the atomic number Z of the atom in which the nuclear transition W took place. In this way it is known that the conversion electron and the competing gamma-ray emission are sequels and not antecedents of alpha decay, beta decay, and electron-capture transitions.

The total internal conversion coefficient α_T is the ratio of the number of transitions proceeding by internal conversion, N_{eT}, to the number going by gamma-ray emission, N_γ, for any particular nuclear transformation from an excited level to a lower-lying level, as in Eq. (34). The total internal conversion co-

$$\alpha_T = \frac{N_{eT}}{N_\gamma} \qquad (34)$$

efficient is a sum of the conversion coefficients for each shell [K, $L(L_1 + L_2 + L_3)$, $M(M_1 + \cdots)$, and so

forth], and is given by Eq. (35), where $\alpha_K = N_{eK}/N_\gamma$,

$$\alpha_T = \alpha_K + \alpha_L + \alpha_M + \cdots \qquad (35)$$

$\alpha_L = N_{eL}/N_\gamma = \alpha_{L1} + \alpha_{L2} + \alpha_{L3}$ ($\alpha_{L1} = N_{eL1}/N_\gamma, \ldots$), and N_{eK}, N_{eL} ($N_{eL1}, \ldots$), and so forth are the numbers of electrons ejected from the K, L (L_1, L_2, L_3), $\ldots$ shells, respectively. In general, this probability of internal conversion relative to gamma-ray emission increases with increasing atomic number Z, with increasing multipole order 2^l, and with decreasing nuclear deexcitation energy W. In middle-weight elements, for $W = 1$ MeV, α is of the order of 10^{-2} to 10^{-4}; while for $W = 0.2$ MeV, α is of the order of 0.1 for electric $l = 2$ transitions, and 10 or larger for electric $l = 5$ transitions. When internal conversion electron decay occurs, this process is always followed by the emission of characteristic x-rays of the element and Auger electrons from outer shells when the inner shell vacancy is filled. This emission can include K x-rays, L x-rays and x-rays from higher shells, and K Auger electrons, L Auger electrons, and so forth. *See* AUGER EFFECT.

Radiationless transitions. There are cases where gamma-ray emission is strictly forbidden and conversion electron emission allowed (Fig. 13). This occurs when both nuclear states have zero spin and the same parity. The conversion electrons are called electric monopole radiations, EO. These transitions occur because of the penetration of the atomic electrons into the nuclear volume where they interact directly with the nucleus. EO radiation can occur in principle whenever two states have the same spin and same parity, but in practice, EO decays are found to be very, very small in these cases. There are some exceptions in well-deformed nuclei and in nuclei which have states with quite different shapes. In these cases, EO decays can totally dominate the electron emission for transitions that have no change in spin and that involve decays between states with large differences in their nuclear shapes (Fig. 13).

The EO decays which arise because of the penetration of the atomic electrons into the nuclear volume are thus sensitive measures of changes in shape between two nuclear states, and have played important roles in establishing vibrations of the nuclear shape and the coexistence of states with quite different deformation in the same nucleus. There also are other circumstances where the penetration of the atomic electron into the nuclear volume gives rise to additional contributions to the conversion-electron decay. Again these penetration effects probe details of the structure of the nucleus.

Internal pair formation. When the energy between two states in the same nucleus exceeds 1.022 MeV, twice the rest mass energy of an electron, it is possible for the nucleus to give up its excess energy to an electron-positron pair—a pair creation process. This is a third alternative mode to gamma decay and conversion electron decay. This process becomes more important as the gamma-ray energy increases. It is relatively unimportant below 2–3 MeV of decay energy. *See* ELECTRON-POSITRON PAIR PRODUCTION.
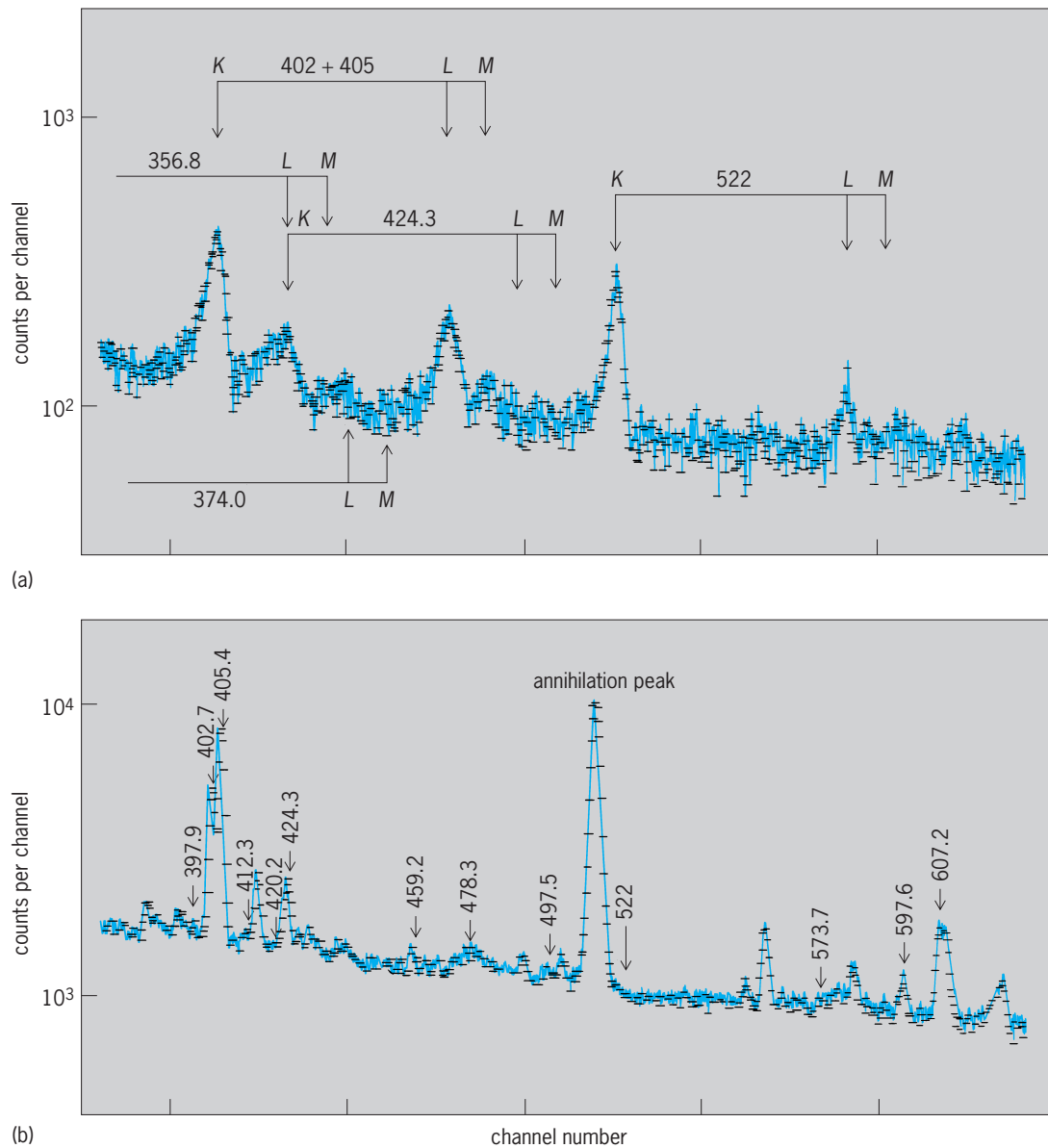
Fig. 13. Spectra from the decay of 30-s half-life ^{186}Tl far off stability (15 neutrons less than the lightest stable thallium isotope). Nuclear transition energies are given in kiloelectronvolts. (*a*) Internal conversion electrons. (*b*) Gamma rays. The strong 522-keV electron transition has no gamma ray associated with it. The strong 511-keV gamma ray is from annihilation of positrons and is not a nuclear transition. (*After J. H. Hamilton et al., Shape coexistence in ^{186}Hg and the decay of ^{186}Tl, Phys. Rev., C16: 2010–2018, 1977*)

Isomeric transitions. Measurably delayed radioactive transitions from an excited level of a nucleus are known as isomeric transitions. The measurably long-lived excited level is called an isomeric or metastable level or an isomer of the ground level. What constitutes an isomer is not well defined. The terminology arose when it was difficult to measure mean lives shorter than 10^{-7} s. States with longer mean lives were isomers. Now mean lives down to 10^{-13} s can be measured for many transitions in different nuclei, but these are not generally called isomers. The break point is simply not defined.

Figure 12 shows that if the excitation energy is small (say, 0.5 MeV or less) and the angular momentum difference l is large (say, $l = 3$ or more), then

the mean life of an excited level for gamma-ray or conversion-electron emission can be of the order of 1 s up to several years.

Most of the long-lived isomers occur in nuclei which have odd mass number A. Then either the number of protons Z in the nucleus is odd, or the number of neutrons N in the nucleus is odd. The frequency distribution of odd-A isomeric pairs, excited level and ground level, displays so-called islands of isomerism in which the odd-proton or odd-neutron number is less than 50 or less than 82. The distribution is one of several lines of evidence for closed shells of identical nucleons at N or $Z = 50$ or 82 in nuclei, and it plays an important role in the so-called shell model of nuclei. *See* MAGIC NUMBERS; NUCLEAR ISOMERISM.

Spontaneous Fission

This involves the spontaneous breakup of a nucleus into two heavy fragments (two intermediate atomic number elements, for example, with $Z = 42$ and 50) and neutrons (Table 1). Spontaneous fission can occur when the sum of the masses of the two heavy fragments and the neutrons is less than the mass of the parent undergoing decay. After the discovery of fission in 1939, it was discovered that isotopes like ^{238}U had very weak decay branches for spontaneous fission, with branching ratios on the order of 10^{-6}. Some isotopes with relatively long half-lives such as ^{252}Cf have large (3.1%) spontaneous fission branching. In these cases, the nucleus can go to a lower energy state by spontaneously splitting apart into two heavy fragments of rather similar mass plus a few neutrons. This process liberates a large amount of energy compared to any other decay mode. Thus, ^{252}Cf has become important in many applications in medicine and industry as a compact energy source, a source of neutrons or as a source of nuclear radiation, since the fragments themselves are left in excited states and so emit gamma rays.

An important isomeric decay mode was discovered in the early 1960s in the very heavy elements, spontaneous fission isomers. Here the nucleus in an excited state, rather than emit a gamma ray or conversion electron, spontaneously breaks apart into two heavy fragments plus neutrons exactly as in spontaneous fission. To identify these isomers, the symbol f is often placed after their atomic mass, for example, $^{244f}_{95}$Am. Their half-lives are generally short, 10^{-3} to 10^{-9} s. It is now understood that these fission isomers are states with much larger deformation than the ground states of these isotopes. The Coulomb barrier against fission is in fact a double-hump barrier with the fission isomers in the valley at large deformation (**Fig. 1**4). The study of these fission isomers has provided important tests of understanding of the behavior and structure of nuclei with very large deformation.

Fission with the emission of neutrons is called hot fission because the fragments have high excitation energy that is carried away by the neutron
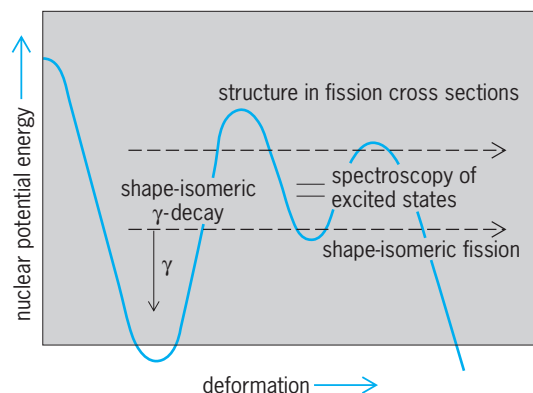
Fig. 14. **Observable consequences of a double-humped nuclear potential barrier against fission. The potential well at the larger deformation gives rise to isomeric spontaneous fission.**

emission. In 1994 cold spontaneous fission was observed for ^{252}Cf and, subsequently, ^{242}Pu, in which no neutrons are emitted, in contrast to hot spontaneous fission, where one to ten neutrons are emitted. Cold ternary spontaneous fission and hot ternary spontaneous fission have also been observed. In ternary spontaneous fission, in addition to the two intermediate-mass nuclei, one light nucleus such as an alpha particle, ^{3}H, or ^{6}Li up to ^{14}C, has been observed. In cold ternary spontaneous fission no neutrons are emitted, while in hot ternary spontaneous fission one or more neutrons are emitted. These rare decay modes provide new insight into the fission process, clustering in nuclei, and the theory of multifragmentation, as discussed below. *See* NUCLEAR FISSION.

Joseph H. Hamilton

Heavy Cluster Decays

Alpha-particle decay and spontaneous fission are two natural phenomena in which an atomic nucleus spontaneously breaks into two fragments, but the fragments are of very different size in one case and almost equal size in the other. On the basis of fragmentation theory and the two-center shell model, new kinds of radioactivities that are intermediate between alpha particle decay and fission were predicted in 1980. Subsequently, it was shown theoretically that the new processes should occur throughout a very broad region of the nuclear chart, including elements with atomic numbers higher than 40. However, experimentally observable emission rates could be expected only for nuclei heavier than lead, in a breakup leading to a very stable heavy fragment with proton and neutron numbers equal or very close to $Z = 82$, $N = 126$ ($^{208}_{82}$Pb or its neighborhood).

For more than 150 kinds of cluster emission, the predicted half-periods of the parent nuclei are shorter than 10^{23} years. The main competitor is always alpha-particle decay. The predicted branching ratios relative to this alpha-particle decay are smaller than 10^{-9}, with a maximum value for $^{14}_6$C radioactivity of $^{223}_{88}$Ra, the first experimentally observed case. Clusters are emitted through fission processes in which the fragments retain compact shape configurations, with a relatively high kinetic energy of about 2 MeV per nucleon.

In 1984, a series of experimental confirmations began with the discovery of $^{14}_6$C radioactivity of $^{223}_{88}$Ra. Initially a semiconductor telescope identification technique was used. In subsequent experiments on $^{14}_6$C emission from other radium isotopes, magnetic spectrometers (a superconducting solenoid and an Enge split pole) allowed suppression of the strong background of alpha particles.

A very promising technique uses solid-state track-recording detectors with special plastic films and glasses that are sensitive to heavier clusters but not to alpha particles. This technique has been applied to the entire range of cluster emissions measured previously and has yielded results in good agreement with theoretically predicted half-lives in the emission of $^{14}_6$C from radium isotopes with mass numbers

222–224 and 226; $^{20}_{8}O$ from $^{288}_{90}Th$; $^{23}_{9}F$ from $^{231}_{91}Pa$; $^{24}_{10}Ne$ from $^{230}_{92}Th$, $^{231}_{91}Pa$, and $^{232-234}_{92}U$; $^{28}_{12}Mg$ from $^{234}_{92}U$; and many others. *See* PARTICLE TRACK ETCHING.

Cluster radioactive decay to excited states of the daughter nucleus or involving excited clusters has been predicted, and experimentally verified.

In cold fission, the fragments are not very deformed or excited, just as in cluster radioactivity. Cold fission has been interpreted, according to the two-center shell model, as cluster emission. As noted above, cold fission has been clearly established in the spontaneous fission of ^{252}Cf and ^{242}Pu, and cold ternary spontaneous fission has also been observed. These decay modes provide tests of a unified theory that includes cold fission, light cluster radioactivity, and alpha decay.

While quite rare, these modes also provide interesting tests of the understanding of the structure of heavy nuclei, including shell and pairing effects, deformation, large-amplitude collective motion, and clustering, with particular emphasis on very heavy clusters like $^{132}_{50}Sn$ and $^{208}_{82}Pb$. Because of its particularly strong shell effects, the $^{208}_{82}Pb$ cluster is involved not only in cluster decays but also in the asymmetric cold fusion process, which has been used in the synthesis of the heaviest elements. *See* ELEMENT 112.

Walter Greiner; Joseph H. Hamilton

Proton Radioactivity

Proton radioactivity is a mode of radioactive decay that is generally expected to arise in proton-rich nuclei far from the stable isotopes, in which the parent nucleus changes its chemical identity by emission of a proton in a single-step process. Its physical interpretation parallels almost exactly the quantum-mechanical treatment of alpha-particle decay. It is also theoretically predicted that one can have the simultaneous emission of two protons—two-proton radioactivity. Although proton radioactivity has been of considerable theoretical interest since 1951 and is expected to be a general phenomenon, for many years only a few examples of this decay mode were observed, because of the narrow range of half-lives and decay energies where this mode can compete with other modes. However, in the late 1990s, experimental techniques using new recoil mass spectrometers, which can separate rare reaction products, and new double-sided silicon strip detectors became available and opened up the discovery of many new proton radioactivities.

The first nuclide found, in 1970, to decay by proton radioactivity was $^{53m}_{27}Co$ (**Fig. 15**), where the *m* (metastate) denotes a (relatively) long-lived isomeric state. Because of its very high angular momentum of 19/2 and odd parity, gamma decay is highly forbidden. This mode of decay is essentially the same as that of β-delayed proton emission discussed below, except that the energy of the excited nuclear level is low, and angular momentum selection rules highly forbid gamma-ray decay so the state lives a relatively long time in comparison to those states populated in beta decay. It was produced in the laboratory by the compound nucleus reactions

$^{16}_{8}O + ^{40}_{20}Ca \rightarrow ^{53m}_{27}Co + p + 2n$ and $^{54}_{24}F + p \rightarrow ^{53m}_{27}Co + 2n$. This 247-ms isomer exhibits two different decay modes: though it predominantly decays by positron (β^+) emission to a similar $19/2^-$ level in $^{53}_{26}Fe$, a 1.5% branch in its decay occurs via direct emission of a 1.59-MeV proton to the $^{52}_{26}Fe$ ground state. The calculated half-life that $^{53m}_{27}Co$ would possess if proton radioactivity were the only decay mode (its partial half-life for this decay branch) is the surprisingly long time of 17 s.

The first two cases of ground-state proton radioactivities were reported in 1982. These were $^{151}Lu \rightarrow ^{150}_{70}Yb + p$ ($T_{1/2} = 85$ ms) and $^{147}_{69}Tm \rightarrow ^{146}_{68}Er + p$ ($T_{1/2} = 0.56$ s) produced in the reactions $^{58}Ni + ^{96}Ru$ and $^{58}Ni + ^{92}Mo$. Others have now been found as heavy-ion experiments have reached isotopes still further off stability, for example, $^{113}_{55}Cs$. Nevertheless, the windows for the observation of direct proton decays are small and, therefore, such decays are very difficult to identify.

As one continues to remove neutrons from the nucleus of a given element with atomic number Z, one reaches a point, called the proton drip line, where for that Z and N a single proton becomes unbound and can drip off the nucleus. The Coulomb barrier (discussed earlier for alpha decay) and the centrifugal (angular momentum) barrier can hold the proton in for a limited time, so the nucleus undergoes radioactive decay with a certain half-life for the emission of a proton. Studies of proton radioactivities probe the limits of stability of proton-rich nuclei. Beyond the proton drip line, proton radioactivity is 100% of the decay. As one moves inside the line to more stable nuclei, proton radioactivity competes with positron decay, and, with the addition of more neutrons, positron beta decay becomes 100% of the radioactivity.

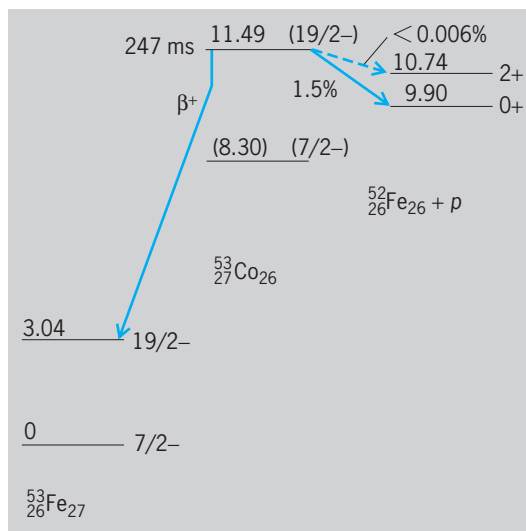There are numerous examples of proton radioactivities with half-lives from hundreds of milliseconds

Fig. 15.   Decay scheme of $^{53m}_{27}Co$. Numbers to left of levels represent energies in megaelectronvolts, relative to ground state of $^{53}_{26}Fe_{27}$. Symbols to right of levels are spin and parity. (*After J. Cerny et al., Further results on the proton radioactivity of $^{53m}_{27}Co$, Nucl. Phys., A188:666–672, 1972*)

down to microseconds. The shortest-lived case is ^{145}Tm ($T_{1/2} = 3.5~\mu$s), discovered in 1998. Most of the newer proton radioactivities have Z between 63 and 82, and most of these proton-radioactive nuclei are spherical. However, an example of proton radioactivity has been found for well-deformed ^{131}Eu, which decays by proton emission to ^{130}Sm, which is likewise well deformed. Fine structure in the proton spectrum is observed with decays to the ground and first excited states of ^{130}Sm to establish the large deformation of this nucleus. Proton fine-structure radioactivities have been observed both in the ^{146}Tm ground state and from an isomeric state ^{146m}Tm. (**Fig. 16**).

Proton radioactivities make it possible to probe the structure of nuclei at the limits and beyond the limits of stability. The half-lives for proton decays are strongly dependent on the energies of the proton and on the angular momentum carried away by the proton. Proton emitters are odd-Z nuclei because in such nuclei the energy to break apart a proton pair in the nucleus is not needed for proton emission to occur. The angular momentum carried away by the proton gives insight into which orbit the proton occupied prior to emission. Moreover, studies in which the proton emission has been used to tag the recoils of a heavy-ion reaction after mass separation have allowed the observation of the gamma rays emitted by the recoil nucleus prior to proton emission (Fig. 16). Thus, it is possible to study excited states in nuclei beyond the limits of proton stability. Both such data test theories of nuclear structure under new extreme conditions. In addition, the energies of the protons provide information about the mass differences of nuclei at the drip lines and so probe mass formulas out to new limits as well.

Two-proton radioactivities from ground states of nuclei are now observed. The first example was the two-proton decay of ^{45}Fe, $^{45}_{26}$Fe $\rightarrow$ $^{43}_{24}$Cr $+ 2p$. Two-proton decay of an excited state in ^{18}Ne to ^{16}O was observed earlier.

Neutron Radioactivity

In very neutron rich nuclei near the boundary line of nucleus stability, one may find nuclei with ground or excited states which are unstable to the emission of one or two neutrons. Here there is no Coulomb barrier to hold the neutron in the nucleus, but one can have a centrifugal barrier that may give rise to one- or even two-neutron radioactivity. These processes for ground states would be very near the limits where nuclei become totally unstable to the addition of a neutron, the neutron drip line, and very difficult to even make much less measure. However, there may be neutron-rich nuclei with high spin isomeric states where the high spin analogous to the one in ^{53m}Co gives rise to a large centrifugal barrier. Such isomeric states may undergo one- or two-neutron radioactivity.

Delayed Particle Emissions

Thirteen types of beta-delayed particle emissions have been observed (Table 1). Over 100 beta-delayed
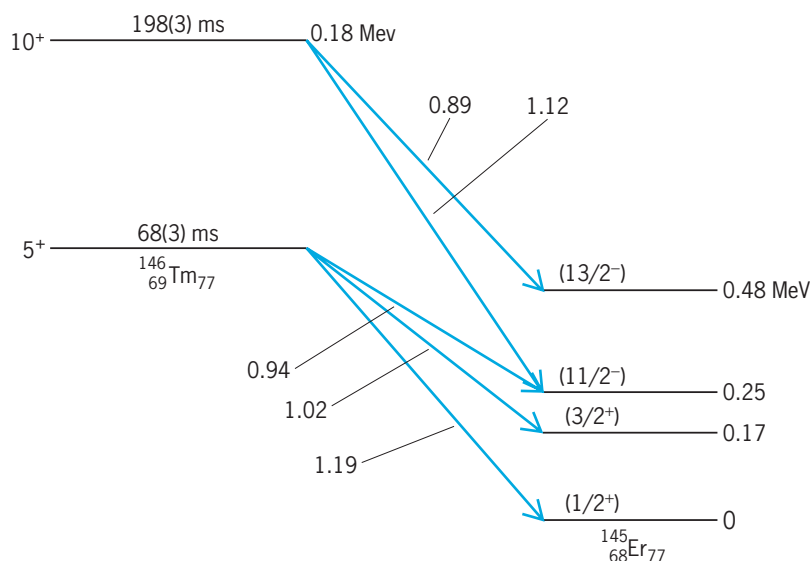

Fig. 16. Proposed decay scheme of the $^{146}_{69}$Tm ground state and of the isomeric state $^{146m}_{69}$Tm. Symbols at left of levels are spin and parity. Numbers above $^{146}_{69}$Tm levels are half-lives. Numbers to right of levels are energies relative to ground state in megaelectronvolts. Decays are labeled with their energies in magelectronvolts. (*After K. P. Rykaczewski et al., Proceedings, Conference on Nuclei at the Limits, Argonne, IL, July 2004*)

particle radioactivities are now known. Theoretically, the number of isotopes which can undergo beta-delayed particle emission could exceed 1000. Thus, this mode, which was observed in only a few cases prior to 1965, is among the important ones in nuclei very far from the stable ones in nature. Studies of these decays can provide insights into the nucleus which can be gained in no other way.

Beta-delayed alpha radioactivity. The β^- decay of ^{214}Bi to ^{214}Po leaves the nucleus in such a high-energy excited state that it can emit an alpha particle and go to ^{210}Pb as an alternative to gamma-ray decay to lower levels in ^{214}Po. This is a two-step process with beta decay the first step. After beta decay the nucleus is in such a highly excited state that it can emit either an alpha particle or gamma ray.

The β^- delayed alpha emission has been found relatively rarely, but in many cases beta (β^+, EC) delayed alpha emission has been discovered. In proton-rich nuclei far from stability, the conditions are more favorable for beta (β^+, EC) delayed alpha emission because of the excess of nuclear charge, and a number of such beta-delayed alpha emitters are now known.

Beta-delayed neutron radioactivity. In 1939, shortly after the discovery of nuclear fission, it was proposed that the delayed neutrons observed following fission were in fact beta-delayed neutrons. That is, after the nucleus fissioned, the beta decay of the neutron-rich fission fragments populated high-energy excited states that could promptly undergo dual decay, emitting either a gamma ray or neutron (Fig. 2). The processes of beta-delayed two- (Fig. 2) and three-neutron emission were discovered in 1979 and 1980 in the decay of ^{11}Li, and β^-2n to $4n$ decays were subsequently observed in other nuclei.

The process of beta-delayed neutron emission is essential for the control of nuclear fission reactors.

When neutrons absorbed by ^{235}U cause the ^{236}U nucleus formed to fission, many of the fission products undergo beta-delayed neutron emission. These neutrons are important in producing more fission events. In a nuclear reactor, the rate of neutron-induced fission must be controlled to prevent the fission reactions from running away and destroying the reactor. The rate of fission depends on the number of neutrons available. The numbers of neutrons can be controlled by moving in and out of the reactor control rods, which contain material with very high neutron absorption rates. Since many of the neutrons emitted in fission are delayed by beta-decay half-lives, these half-lives allow time for the control rods to be mechanically inserted and removed to control the rate of fission. *See* DELAYED NEUTRON.

Beta-delayed proton radioactivities. In addition to proton radioactivity, one can have beta-delayed proton and beta-delayed two-proton radioactivities which again ultimately result in emission of protons from the nucleus. These latter processes also occur in quite proton-rich nuclei with very high decay energies; however, they are complex two-step decay modes whose fundamental first step is beta decay.

Over 40 nuclei ranging from $^{9}_{6}$C to $^{183}_{80}$Hg have been identified to decay by the two-step mode of beta-delayed proton radioactivity. Typical is the decay of $^{33}_{18}$Ar (**Fig. 17**), with a half-life of 173 ms; it was produced by the $^{32}_{16}$S + $^{3}_{2}$He → $^{33}_{18}$Ar + 2n reaction. This isotope decays by superallowed and allowed β^+ decay to a number of levels in its daughter nucleus $^{33}_{17}$Cl, which immediately (in less than 10^{-17} s) breaks up into $^{32}_{16}$S and a proton. More than 30 proton groups arising from the decay of $^{33}_{18}$Ar are observed, ranging in energy from 1 to approximately 6 MeV

and varying in intensity over four orders of magnitude. Although it is normally very difficult to study many β-decay branches in the decay of a particular nuclide—because of the continuous nature of the energy spectrum of the emitted beta particles—it is possible to do so when investigating beta-delayed proton emitters. The observed proton group energies and intensities can be correlated with the levels fed in the preceding beta decay and their transition rates, thereby permitting sensitive tests via beta decay of nuclear wave functions arising from different models of the nucleus. β^+-delayed two-proton decay has also been discovered, as well as β^--delayed deuteron and triton decay.

Beta-delayed spontaneous fission. There are also observed beta-decay processes where the excited nucleus following beta decay has a probability of undergoing spontaneous fission rather than gamma-ray decay. This is the same process as in spontaneous or isomeric spontaneous fission. The excitation energy of the nuclear level provides the extra energy to make fission possible. The nucleus splits into two nearly equal fragments plus some neutrons. This process is like isomeric spontaneous fission except that the lifetime of the nuclear level is so short that the level would not normally be called an isomer.   Joseph H. Hamilton

Bibliography. D. A. Bromley (ed.), *Nuclei Far from Stability*, vol. 8 of *Treatise on Heavy-Ion Science*, Plenum, 1989; J. M. Eisenberg and W. Greiner, *Nuclear Theory*, 3 vols., 3d ed., Elsevier, 1987–1988; R. B. Firestone and V. S. Shirley (eds.), *Table of Isotopes*, 8th ed., Wiley, 1996; E. J. Konopinski, *The Theory of Beta Radioactivity*, Clarendon Press, 1966; D. N. Poenaru (ed.), *Nuclear Decay Modes*, Institute of Physics, 1996; A. Sandulescu and W. Greiner, New radioactivities, *Sci. Amer.*, 262(3):58, March 1990; F. Yang and J. H. Hamilton, *Modern Atomic and Nuclear Physics*, McGraw-Hill, 1996.

Radioactivity and radiation applications

The field in which the subatomic fragments emitted in radioactive decay (alpha particles, beta particles, gamma rays) or produced by high-voltage accelerators (electrons, protons, x-rays) are applied to the problems of science, engineering, industry, and medicine. The techniques are extraordinarily versatile and sensitive and are basically inexpensive. A disadvantage that limits the range and extent of these applications is the health hazard that may be involved. *See* RADIOACTIVITY; RADIOISOTOPE.

Tracer applications. These techniques are based on two principles. First is the chemical similarity of radioactive atoms and other atoms of the same element. Periodically a few of the radioactive atoms decay, emitting some penetrating subatomic fragments that can be detected one by one, usually through their ability to cause ionization. Thus the movement of a particular element can be followed through various chemical, physical, and biological steps. The second principle involves the characteristic half-life and nature of the emitted fragments. This
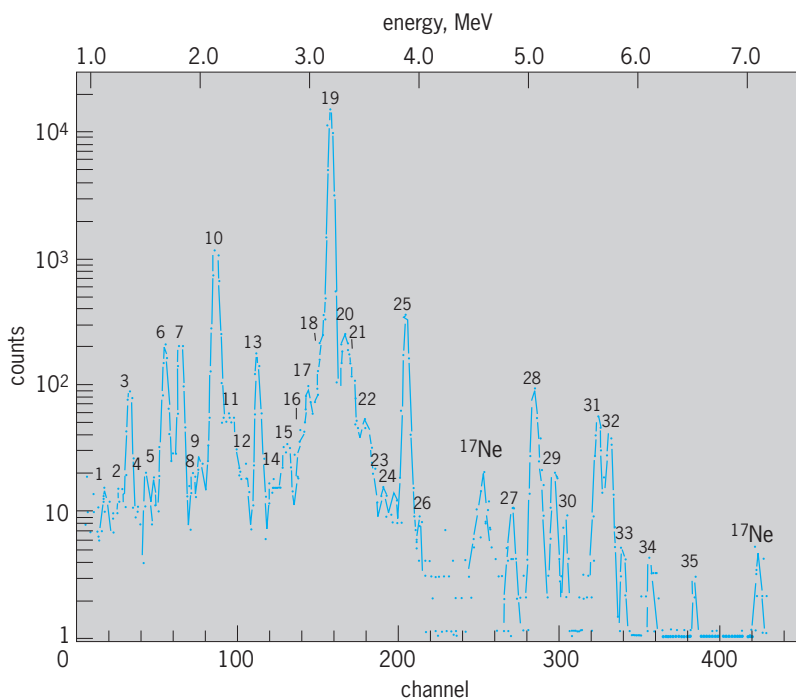

Fig. 17. Spectrum of β^+-delayed protons from the decay of $^{33}_{18}$Ar as observed in a counter telescope; the proton laboratory energy is indicated at the top. Proton groups are numbered 1 through 35. (*After J. C. Hardy et al., Isospin purity and delayed-proton decay: ^{17}Ne and ^{33}Ar, Phys. Rev., C3:700–718, 1971*)

makes a radioactive species unique and thereby detectable above a background of radioactive emitters associated with elements.

The use of tracers in pure and applied research, process control, and medical diagnosis is one of the triumphs of modern science. Examples include the following: (1) A small amount of radioactive additive injected into an opening in a piping system can be used to measure the flow rate throughout the system and to detect the leaks in buried portions of the pipe. (2) In the metals industry, wear studies have been successfully performed using radioactively tagged tool bits and piston rings; radioactivity buildup in the lubricant, on the metal being shaped, or on the wall of the cylinder makes possible a simple and rapid measure of the durability of the tagged component. (3) Tracers added to the bloodstream have been used to monitor the movement of biochemical components throughout the body. (4) Addition of short-lived tracers to industrial waste streams have been used to determine their flow patterns. (5) A good way to determine when mixing is complete in a 50,000-bbl (8000 m^3) tank at an oil refinery is to add a small amount of radioactivity and find out when the activity level in random samples is constant. (6) Important contributions in understanding the mechanism of combustion have been obtained by means of tracer studies. (7) Diffusion in a high-viscosity medium or in a crystal is a slow process whose measurement is tedious and difficult by any technique except the tracer method. For discussions of radioisotope techniques relating to tracer methodology *see* ACTIVATION ANALYSIS; ISOTOPE DILUTION TECHNIQUES; NONDESTRUCTIVE EVALUATION; RADIOACTIVE TRACER; RADIOECOLOGY; RADIOISOTOPE (BIOLOGY).

Penetration and scattering applications. These applications arise from the fact that subatomic fragments can penetrate a thick section of a material, and yet a small fraction of the incident particles can be backscattered by a relatively thin section.

The oldest application of the penetrating properties of energetic ionizing photons is radiography. A more recent extension of this technique is autoradiography. *See* AUTORADIOGRAPHY; RADIOGRAPHY.

Since World War II the penetration and scattering properties of beta particles and gamma rays have been applied in industry in the form of thickness gages, particularly for process control in the continuous production of paper and metal films. In the usual application of a penetration-thickness gage, a beta source is placed on one side of the continuous sheet, and a beta detector is mounted on the other side. The basic absorption law is given by the notation below. Because the number of beta particles

$$N \propto e^{-x/\overline{x}}$$

penetrating the sheet N is an exponential function of the thickness x, small changes in thickness can be detected when a beta source with optimum penetration properties is used. The optimum condition is one in which the thickness of the sheet equals the apparent mean free path of the beta particle, $\overline{x}$, which varies directly with the energy of the beta particle.

A related application is the use of the amount of beta backscatter to measure the thickness of a thin coating on a thick backing of different composition. A backscatter gage using thallium-204 beta particles can easily monitor changes of 4×10^{-5} in. (10^{-4} cm) in the thickness of a tin coating on a steel backing.

An important application of the scattering and penetration of neutrons and gammas is in the field of well logging. *See* WELL LOGGING.

Initiation of chemical reactions. The absorption of small amounts of energy from ionizing particles and ionizing photons have chemical effects that have been the basis of several practical applications. A dose of 1 rad (0.01 gray) is equivalent to the deposition in each gram of the absorber of 100 erg (1 erg = 10^{-7} joule = 7.4×10^{-8} ft-lbf) from a source of radiation. An absorbed whole body dose of 500 rads (5 gray) may be sufficient to cause death in a human, but the same amount of energy in the form of heat would do no more than raise the body temperature by 0.002°F (0.001°C).

The oldest application of this principle is radiation therapy. For example, in cancer therapy the local affected areas are irradiated by external beams of gammas from cobalt-60 or of radiation from accelerators. Radioactive sources have also been administered internally to induce beneficial biochemical reactions in patients with various ailments. Injection of gold-198 in colloidal form into tumor tissue is a common therapeutic procedure. An iodine-131 "cocktail" is useful in the treatment of an overactive thyroid gland. The intravenous injection of phosphorus-32 is a standard method of treating leukemia and other blood diseases. *See* ISOTOPIC IRRADIATION; RADIOLOGY.

A related area is the radiation sterilization of biomedical supplies. There are about 50 installations throughout the world employing ionizing radiation (usually in the form of cobalt-60 gamma rays) to give sterilizing doses (2.5–4 megarads or 25–40 kilograys) to disposable medical supplies. The advantages to this method of biochemical destruction of microscopic life are that (1) unlike steam sterilization, it can be performed at low temperatures on plastics and other thermally unstable materials, and (2) unlike germicidal gases, ionizing radiation can reach every point in the treated product. Radiation-sterilized objects are not radioactive.

The radiation preservation of food is an area of considerable promise. Small doses can inhibit sprouting in potatoes (10,000 rads or 100 Gy), kill insects in wheat (50,000 rads or 500 Gy), and sterilize pork products (3×10^6 rads or 30,000 Gy) but practical applications have been sharply limited due to a cautious role by regulatory authorities in approving such procedures. There is also industrial concern over the possibility of consumer rejection of foods preserved by treatment with ionizing radiation. Despite the short-term problems, there are grounds for optimism in the long run. The energy requirements for radiation preservation of food are far less than those of a preservation technology based on steam sterilization and refrigeration. Rising energy costs as well as increasing laboratory and marketing experience

with irradiated foods should lead to a more favorable attitude.

Although ionizing radiation has not yet proved to be an attractive alternative to nonradiation methods of chemical synthesis, it has assumed a significant role in the modification of plastics. Electron-beam irradiation of wire insulated by polyethylene or poly(vinyl chloride) formulations is an important means of upgrading the properties of the insulation at higher temperatures. The ionizing radiation induces chemical reactions that lead to the cross-linking of polymer chains to form a network called a gel, which, in effect, does not melt and does not dissolve in solvents. Electron-beam irradiations of polyethylene packaging is used to manufacture heat-shrinkable form-fitting pouches for turkeys, hams, electronic parts, and so on. The principle behind the application is as follows. Polyethylene is highly cross-linked by irradiation; when heated above its normal melting point, it behaves like a rubber. Thus a pouch made of cross-linked polyethylene, which is then heated and stretched, retains the stretched configuration when cooled. When a ham is placed in this pouch and the entire package is heated, the now-rubbery pouch shrinks to form a skin-tight covering that retains its shape after cooling. The same principle has been successfully applied to produce heat-shrinkable electrical-insulation components that have widespread industrial use.

Another important application of electron irradiation is the curing of paint applied to plastic or metal backings. Some automobile dashboards are coated with paint that is "cold-cured" by exposure to accelerated electrons. A variety of automobile parts are treated in a similar manner. The process replaces one involving long-heating ovens with high thermal energy losses and with the liberation to the atmosphere of volatile paint components.

Electron machines are used for a radiation process that produces an estimated 2×10^9 ft^2 (2×10^8 m^2) per year of a permanent-press polyester-cotton blend with excellent soil-release properties. The process involves irradiation of the fabric, which has been soaked with a reagent that ultimately provides the permanent-press characteristics. The irradiation provides reactive sites on the fabric to which the reagent molecules attach themselves. The fabric is then exposed to the soil-release agent, which then hooks on to some long-lived reactive sites that persist after the irradiation.

The role of ionizing radiation in sterilization and chemically oriented applications is significant, but it is not a major factor in modern technology. A major motivating force behind the research leading to radiation processes is the potential availability of large inventories of fission products. Nevertheless, the only successful industrial irradiators involve either cobalt-60 or electron accelerators. The fission product, cesium-137, has some potential future value as a gamma source. *See* RADIATION CHEMISTRY.

Radioactive power sources.   Kinetic energy of emissions in radioactive decay can be converted to useful forms of light, heat, and electricity.

A mixture of a phosphor and radionuclide produces a luminous paint that has been used for watch and instrument dials. Such a light source in combination with a photovoltaic cell can provide electricity at a power level up to a microwatt. *See* LUMINOUS PAINT.

The heat produced by betas from strontium-90 on their absorption in a thermoelectric converter has been used to power buoys and remote arctic radio-transmitters. Similar converters of heat to electricity using promethium-147 and plutonium-238 are the microwatt power sources for pacemakers that regulate heart action. Larger versions of the plutonium-238 battery have been proposed for use in an artificial heart.

Although direct conversion to electricity of the energy of alphas and betas has been accomplished, the most efficient and practical converters are based on the conversion to heat in an intermediate step. *See* NUCLEAR BATTERY.   Joseph Silverman

Bibliography. G. Foldiak (ed.), *Industrial Applications of Radioisotopes*, 1986; M. F. L'Annunziata, *Radionuclide Tracers*, 1987; E. Reichmanis, C. W. Frank, and J. H. O'Donnell (eds.), *Irradiation of Polymeric Materials*, 1993; J. W. Spinks and R. J. Woods, *An Introduction to Radiation Chemistry*, 3d ed., 1990; W. M. Urbain, *Food Irradiation*, 1986; P. Von der Hardt and H. Rottger (eds.), *Irradiation Technology*, 1983; R. R. Wolfe, *Radioactive and Stable Isotope Tracers in Biomedicine*, 1992.

Radioactivity standards

Calibrated standard sources of radioactive substances used to determine, by comparison, the strength or activity of samples of the same substances in terms of the number of radioactive atoms they contain or in terms of some figure proportional to this number. The calibration of the standard source in terms of the number of radioactive atoms is usually an elaborate procedure but need only be carried out once, and the calibration may be made at a standardizing laboratory, such as the National Institute of Standards and Technology, having special equipment for the work. Comparison between a sample and the standard is usually made by finding the ratio of the response of an ionization chamber, or other detector of radiation, to the radiation from a sample and from the standard. In each case the intensity of the radiation, and therefore the response of the detector under identical conditions, is proportional to the number of radioactive atoms in the source, because this number is also proportional to the activity or disintegration rate of a source. *See* HALF-LIFE; RADIOACTIVITY.

The situations in which standards of radioactivity can be used are limited by a number of factors. For example, the radioactive species involved must have a half-life long enough that the standards will continue to have sufficient activity for the period of time in which they are to be used. Thus standards of radium, in which the half-life is about 1620 years, have an almost permanent value. Standards of radioisotopes

with very much shorter half-lives, of the order of a day, require preparation and calibration immediately prior to their use.   Leon F. Curtiss; Karl Z. Morgan

Bibliography. *American Institute of Physics Handbook*, 3d ed., 1972; H. Etherington (ed.), *Nuclear Engineering Handbook*, 1958; W. B. Mann, The preparation and maintenance of standards of radioactivity, *Int. J. Appl. Radiat. Isotop.*, 1:3–23, 1956; National Council on Radiation Protection and Measurement, *Handbook of Radioactivity Measurements Procedures*, NCRP Rep. 58, 2d ed., 1985; J. Shapiro, *Radiation Protection*: *A Guide for Scientists and Physicians*, 3d ed., 1990.

Radiocarbon dating

A method of obtaining age estimates on organic materials which has been used to date samples as old as 75,000 years. The method was developed immediately following World War II by Willard F. Libby and coworkers, and has provided age determinations in archeology, geology, geophysics, and other branches of science.

Radiocarbon (^{14}C) determinations can be obtained on wood; charcoal; marine and fresh-water shell; bone and antler; peat and organic-bearing sediments; carbonate deposits such as tufa, caliche, and marl; and dissolved carbon dioxide (CO_2) and carbonates in ocean, lake, and ground-water sources. Each sample type has specific problems associated with its use for dating purposes, including contamination and special environmental effects. While the impact of ^{14}C dating has been most profound in archeological research and particularly in prehistoric studies, extremely significant contributions have also been made in hydrology and oceanography. In addition, beginning in the 1950s the testing of thermonuclear weapons injected large amounts of artificial ^{14}C ("bomb ^{14}C") into the atmosphere, permitting it to be used as a geochemical tracer.

Basis of the Method

Carbon (C) has three naturally occurring isotopes. Both ^{12}C and ^{13}C are stable, but ^{14}C decays by very weak beta decay (electron emission) to nitrogen-14 (^{14}N) with a half-life of approximately 5700 years. Naturally occurring ^{14}C is produced as a secondary effect of cosmic-ray bombardment of the upper atmosphere (**Fig. 1**). As $^{14}CO_2$, it is distributed on a worldwide basis into various atmospheric, biospheric, and hydrospheric reservoirs on a time scale much shorter than its half-life. Metabolic processes in living organisms and relatively rapid turnover of carbonates in surface ocean waters maintain ^{14}C levels at approximately constant levels in most of the biosphere. The natural ^{14}C activity in the geologically recent contemporary "prebomb" biosphere was approximately 13.5 disintegrations per minute per gram of carbon. *See* COSMOGENIC NUCLIDE; ISOTOPE.

To the degree that ^{14}C production has proceeded long enough without significant variation to produce an equilibrium or steady-state condition, ^{14}C levels
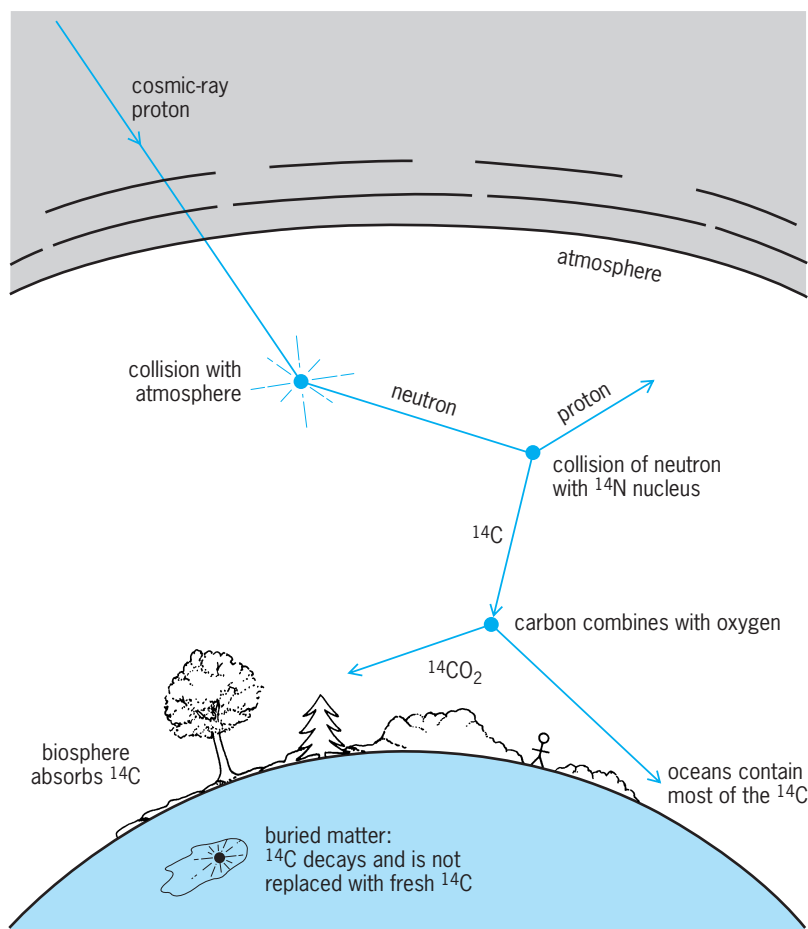
Fig. 1. Generation, distribution, and decay of ^{14}C.

observed in contemporary materials may be used to characterize the original ^{14}C activity in the corresponding carbon reservoirs. Once a sample has been removed from exchange with its reservoir, as at the death of an organism, the amount of ^{14}C begins to decrease as a function of its half-life. A ^{14}C age determination is based on a measurement of the residual ^{14}C activity in a sample compared to the activity of a sample of assumed zero age (a contemporary standard) from the same reservoir. The relationship between the ^{14}C age and the ^{14}C activity of a sample is given by the equation below, where t is radiocar-

$$t = \frac{1}{\lambda} \ln \frac{A_o}{A_s}$$

bon years B.P. (before the present), λ is the decay constant of ^{14}C (related to the half-life $t_{1/2}$ by the expression $t_{1/2} = 0.693/\lambda$), A_o is the activity of the contemporary standards, and A_s is the activity of the unknown age samples. Conventional radiocarbon dates are calculated by using this formula, an internationally agreed half-life value of 5568 ± 30 years, and a specific contemporary standard. Most laboratories define the contemporary standard value by using one of the standards prepared by the U.S. National Bureau of Standards [NBS; now known as the U.S. National Institute of Standards and Technology (NIST)], or a standard with a known relationship to the NBS/NIST oxalic acid preparations.

Measurement of Radiocarbon

The naturally occurring isotopes of carbon occur in the proportion of approximately 98.9% ^{12}C, 1.1% ^{13}C, and 10^{-10}% ^{14}C. The extremely small amount of radiocarbon in natural materials was one reason why ^{14}C was one of the isotopes which had been produced artificially in the laboratory before being detected in natural concentrations. The routine development of the radiocarbon method was made possible by the development by Libby of a practical method of low-level counting. To detect the very weak beta-decay characteristic of ^{14}C, a means had to be devised to introduce the sample directly into the sensitive volume of a detector. In all of Libby's early work, the sample was converted to solid carbon (amorphous elemental carbon) and deposited on a sleeve which fitted inside a screen-wall type of Geiger counter. The counting rate of an unshielded screen-wall counter was on the order of 500 counts per minute.

Since the activity from ^{14}C decay of a modern sample was expected to be about six or seven counts per minute, the total background counting rate had to be radically reduced. This was accomplished initially by placing the instrument in an iron shield with 20-cm (8-in.) walls. This reduced the activity in the detector to 120 counts per minute, still unacceptably high. The final reduction was made possible by enclosing the sample counter in a ring of smaller Geiger counters. The sample counter and the outer guard ring were connected together electronically so that any pulse from any of the outer Geiger tubes would inactivate the sample counter for about 10^{-3} s. This anticoincidence system reduced the background
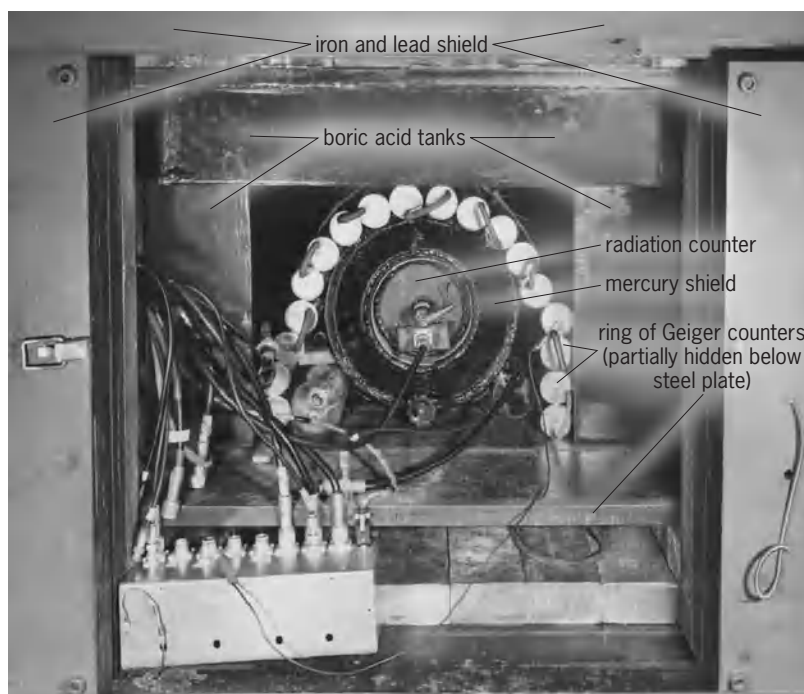

Fig. 2. Large-volume proportional counter, which is used for carbon-14 measurements. The outer shield is closed by rolling doors. The sample is introduced into the radiation counter in the form of carbon dioxide gas. (*Geochemical Laboratory, Lamont-Doherty Geological Observatory, Columbia University*)

in the center detector to about five counts per minute.

With this system, the maximum age that could be measured was about 23,000 years and required the use of 10–12 g (0.35–0.42 oz) of carbon from sample materials. Because of self-absorption of the weak betas in the sample, the efficiency of the detector was only about 5%. Because of this and the susceptibility of the carbon black to contamination from airborne radioactive fallout, the solid carbon technique was replaced by either gas counters or liquid scintillation systems. *See* LOW-LEVEL COUNTING; RADIOACTIVITY.

Gas counters. In the early 1950s, both proportional (**Fig. 2**) and Geiger gas counters were employed in ^{14}C work, using carbon dioxide, carbon disulfide, acetylene, methane, or ethane as counting gases. As in the case of the solid carbon system, the center counter containing the sample was surrounded by individual Geiger tubes or an annular or continuous ring guard, all housed within an iron or lead shield assembly. Efforts to reduce the background values in gas detectors have resulted in various types of experimental arrangements, including the location of counters in underground vaults. In such underground facilities, the contribution of the meson flux, the major contributor to the background rate, can be significantly reduced. Because of the 90–95% efficiency in most gas detector systems, the typical maximum age limits were extended to 40,000–60,000 years, depending on the experimental configuration, including the volume of the detectors and the level of the background count rates in specific detectors. Isotopic enrichment of sample gases permits the maximum age attainable to be extended several additional half-lives. In general, sample-size requirements with gas detectors were reduced from that required with the solid carbon method—special systems being designed to permit the measurement of a sample with as little as 0.1 g (3.5×10^{-3} oz) of carbon. *See* GEIGER-MÜLLER COUNTER; IONIZATION CHAMBER; MESON.

Liquid scintillation systems. Current liquid scintillation systems involve the conversion of samples to benzene. The addition of a scintillator chemical allows beta-decay events to be monitored by photomultiplier tubes. Earlier liquid scintillation systems used for ^{14}C measurements generally required larger samples. However, currently the amounts required are comparable to gas counting systems. Continuing developments in liquid scintillation technology for low-level measurement have provided the ability to monitor counter performance in much greater detail than typically is possible in gas counting, and have also resulted in reduced background values. In these systems, the maximum ages that can be measured can be extended beyond that possible with typical gas systems.

Direct detection. Both of the conventional decay counting methods share a common problem in that they employ an inherently inefficient means of monitoring ^{14}C concentrations in samples. In 1 g of modern carbon, for every decay per minute there are about 4×10^9 atoms of ^{14}C. Counting methods have
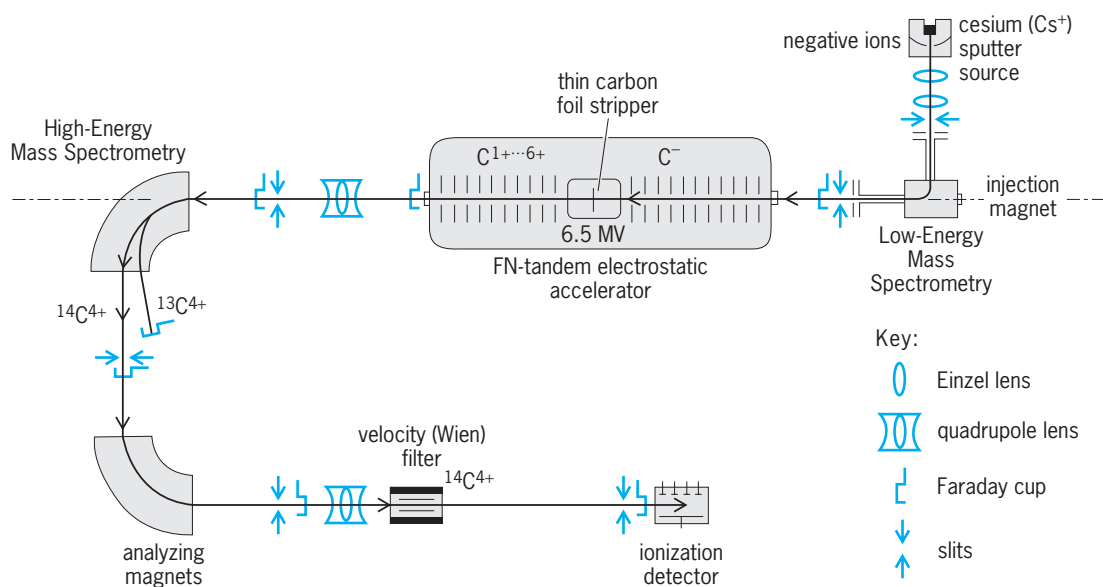
Fig. 3. Accelerator mass spectrometry system for the direct detection of ^{14}C atoms. (*Center for Mass Spectrometry, Lawrence Livermore National Laboratory*)

been developed which employ particle accelerators as very sensitive mass spectrometers, thus counting the ^{14}C atoms directly.

It has long been recognized that if the ^{14}C atoms could be detected directly, rather than by waiting for their decay, smaller samples could be used for dating and older dates could be measured. A simple hypothetical example to illustrate this point is a sample containing only one atom of ^{14}C. To measure the age (that is, the abundance of ^{14}C), the sample can be placed into a mass spectrometer and that atom counted, or the sample can be placed into a Geiger counter and counted, requiring a wait on the average of 8000 years (the mean life of ^{14}C) for the decay. In practice, neither the atoms nor the decays can be counted with 100% efficiency, but the huge advantage for atom counting remains.

Until 1977, attempts at direct atom counting for ^{14}C and other natural radioisotopes failed, because of the extremely low concentration of ^{14}C. Ordinary mass spectrometers could not see the tiny ^{14}C signal in the background of other atoms and molecules in the sample. Even trace amounts of nitrogen would swamp the ^{14}C signal, since ^{14}N forms ions with nearly identical mass and identical charge to that of the ^{14}C atoms.

The technique for detecting ^{14}C atoms [and other natural radioisotopes such as tritium, beryllium-10 (^{10}Be), and chlorine-36 (^{36}Cl)] is a combination of mass spectrometry and accelerator technology, called accelerator mass spectrometry (AMS). The approach was first demonstrated using a cyclotron. This accelerator, which sent particles along a spiral trajectory, was used as an ultrasensitive mass spectrometer to distinguish ionized carbon isotopes by their charge-to-mass ratio. Detecting ^{14}C by this means was possible, but consistent results proved difficult to achieve despite years of effort. *See* MASS SPECTROMETRY.

Another type of AMS technology uses a tandem electrostatic accelerator (**Fig. 3**). The device employs two stages. First, a negative ion beam is accelerated and passed through a stripper, which removes the electrons, converting the beam to positive ions. Then the particles are further accelerated. The stripping process breaks up molecules of mass 14, which would otherwise interfere with the detection of ^{14}C, and the negative ion beam eliminates ^{14}N since there are no known stable negative nitrogen ions. Almost all AMS ^{14}C applications currently use tandem accelerators to accomplish routine measurements. *See* PARTICLE ACCELERATOR.

The advent of AMS technology in the late 1970s brought about an enormous boost in detection efficiency that promised three important advantages for ^{14}C dating. First the amount of carbon required was reduced from grams to milligrams. Second, counting times were reduced from days, weeks, or even months to minutes. Finally, it was initially thought that the detection sensitivity would increase so that the maximum age datable with ^{14}C might be extended to 100,000 years. However, sensitivity is limited by very small amounts of contamination introduced during sample preparation. Much of this contamination stems from the requirement in most laboratories that samples be converted to graphitic carbon for measurement. In routine operation, current AMS technology can measure between 40,000 and 50,000 years, and rarely, 60,000 years.

The ability to use milligram (rather than gram) samples is very important for dating. Certain irreplaceable objects (for example, parchments, cloth, and chips of wood) would have had to be destroyed in order to extract the gram of carbon required for a date; for the accelerator method, only a small piece of the artifact is required. In addition, the ability to date by using only milligrams of carbon allows careful selection of the sample used. Any part of the object

which may have been contaminated by modern carbon can be ignored; small seeds trapped in the object, or even specific amino acid compounds which are less likely to come from modern carbon contamination, can be selected.

Despite the sensitivity of the accelerator technique, decay dating will probably continue to be used for ^{14}C dating when gram amounts of carbon are available. However, the technology of atomic mass spectrometry continues to be developed and will increasingly be used for routine ^{14}C analysis. The accuracy of ^{14}C values based on atomic mass spectrometry has become essentially comparable to that obtained with decay counting.

Accuracy of Radiocarbon Determinations

A measurement of the ^{14}C content of an organic sample will provide an accurate determination of the sample's age if it is assumed that (1) the production of ^{14}C by cosmic rays has remained essentially constant long enough to establish a steady state in the ^{14}C/^{12}C ratio in the atmosphere, (2) there has been a complete and rapid mixing of ^{14}C throughout the various carbon reservoirs, (3) the carbon isotope ratio in the sample has not been altered except by ^{14}C decay, and (4) the total amount of carbon in any reservoir has not been altered. In addition, the half-life of ^{14}C must be known with sufficient accuracy, and it must be possible to measure natural levels of ^{14}C to appropriate levels of accuracy and precision. Studies have shown that the primary assumptions on which the method rests have been violated both systematically and to varying degrees for particular sample types. Several approaches have been developed to provide calibration and corrections of conventional ^{14}C values. The basis of the calibration and correction procedures will be discussed in the context of a brief review of the assumptions of the method.

Constancy in radiocarbon production rates. Carbon-14 determinations on known-age samples have revealed systematic discrepancies in the ^{14}C time scale. The first hint of such anomalies came from early ^{14}C measurements on Egyptian archeological materials. Samples which, on historical grounds, should have dated to the early part of the third millennium B.C. yielded ^{14}C values some 700–800 years too young. Carbon-14 determinations carried out on dendrochronologically dated wood, periglacial varves, and lake sediments confirmed the fact that there have been systematic variations in ^{14}C values over time. The data which first contributed most directly to the study of these anomalies was the dendrochronological time scale provided by the bristlecone pine (*Pinus longaeva*), from the White Mountains of east-central California, developed by C. W. Fergusson. His data provide an unbroken tree-ring series back to almost 6700 B.C. An independently developed bristlecone tree-ring chronology from a different locality in the southern portion of the White Mountains, developed by V. C. La Marche and T. P. Harlan, supports the accuracy of the Fergusson chronology at least as far back as about 3500 B.C. Carbon-14 determinations on bristlecone pine as well as tree-ring-dated samples from the sequoia (*Sequoia gigantea*) and European oaks (*Quercus* spp.) have been undertaken by a number of laboratories, and ^{14}C determinations on these samples provide data over the last 10,000 years. *See* DENDROCHRONOLOGY; VARVE.

Main trend. Upon examination of the data in Fig. 4 and other similar plots, it becomes apparent that radiocarbon years and calendar years are not necessarily equivalent. If such had been the case, all of the data points plotted on **Fig. 4** would lie along the horizontal 0 line. In fact, some of the points lie above the line, indicating that ^{14}C values in these periods are too old. Conversely, those below the 0 line are too young when compared to the tree-ring data. This plot indicates that there are two major components to the deviations. The first is a general main-trend secular variation phenomenon (the curved line) exhibiting during the Holocene, a sine-wave function with an apparent period of about 8500–9000 years, with a maximum deviation of about 800 years, approximately 8000 years ago.

The characteristic of the secular variation anomalies in the period before about 10,000 years ago cannot, at present, be documented by tree-ring/^{14}C data. It has been argued, however, that the major part of the effect may be estimated by examining the record of the Earth's dipole geomagnetic field over time. Variations in the intensity of the dipole field modulate the cosmic-ray flux in the vicinity of the Earth. An increase in the field strength, for example, diverts more of the cosmic-ray particles away from the Earth, resulting in a decrease in the production of ^{14}C.

Geophysicists have collected data which document changes in the intensity of the Earth's dipole field for the last few hundred thousand years. Because of the apparent inverse relationship between the intensity of the field and the ^{14}C production rate, it would, in theory, be possible to extrapolate the maximum and minimum secular variation deviations back to the limit of the ^{14}C method. Unfortunately, such data are not yet as precise as might be wished.
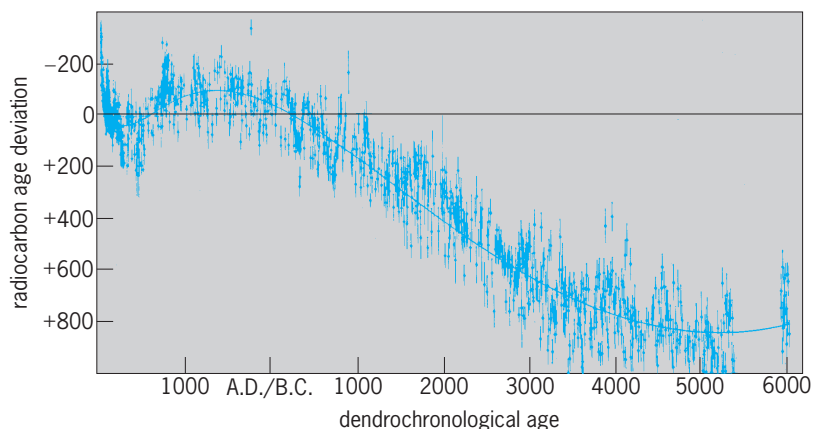

Fig. 4. Secular variation/major trend; relationship between radiocarbon and dendrochronological age of wood samples. (*After J. Klein et al., Calibration of radiocarbon dates, Radiocarbon, 24(2):103–150, 1982*)

However, comparisons of ^{14}C ages with uranium-thorium (U-Th) ages obtained on cores from coral deposits support conclusions based on the ^{14}C/tree-ring data up to the limit of the current dendrochronological data. Uranium-thorium values can be used to continue to examine the ^{14}C deviations over the last 30,000 years. Such data indicate that radiocarbon ages earlier than 10,000 years B.P. continue to be systematically younger than U-Th ages, with a maximum difference of about 3500 years approximately 20,000 years ago. *See* PALEOMAGNETISM; ROCK MAGNETISM.

De Vries effect. In addition to the long-term secular variation phenomenon, the bristlecone pine data have revealed the presence of high-frequency components to the variation in ^{14}C activity. These short-term oscillations or wiggles have sometimes been called the De Vries effect after the pioneering Dutch researcher, Hessel de Vries, who was one of the first to call attention to the existence of systematic anomalies in ^{14}C values. Although almost all investigators concerned with the issue have agreed that the tree-ring/^{14}C data definitely reveal the presence of a number of short-term perturbations, the frequency and magnitude of earlier episodes during the Pleistocene, (that is, > 10,000 years B.P.) have not been resolved. Likewise, there are uncertainties as to the causes of the De Vries effect, although variation in solar activity (heliomagnetic effects) has been seen as an important factor.

Calibration of radiocarbon dates. The existence of main trend and De Vries deviations has important implications in the interpretation of ^{14}C determinations. The long-term variations result in the necessity to calibrate conventional ^{14}C dates in terms of the known variation between radiocarbon time and real or calendar time as documented by the dendrochronological/^{14}C values. The magnitude of the calibration varies depending on from what time period a sample is derived. For the period back to about 1000 B.C., corrections required by virtue of the secular-variation deviations do not exceed about 150 years. Prior to 1000 B.C., the magnitude of the correction steadily increases. By using data such as those presented in Fig. 4, various approaches have been developed to "calibrate" radiocarbon values. In this context, calibration involves taking a ^{14}C age value expressed as a conventional radiocarbon date and adding or subtracting the number of years required to bring the conventional age into conformity with the ^{14}C determinations on known-age tree-ring-dated samples.

The documentation of the presence of the De Vries or short-term anomalies has introduced a second problem in the calibration of ^{14}C values. Periods of rapid change in the ^{14}C content of the atmosphere result in situations where a single ^{14}C value may reflect two or more points in real time. The characteristics of the short-term anomalies are illustrated in **Fig. 5**. During periods of particularly rapid change in ^{14}C activity, it is usually not possible to use ^{14}C data to document temporal intervals in units of less than a few hundred years. Thus the dendrochrono-
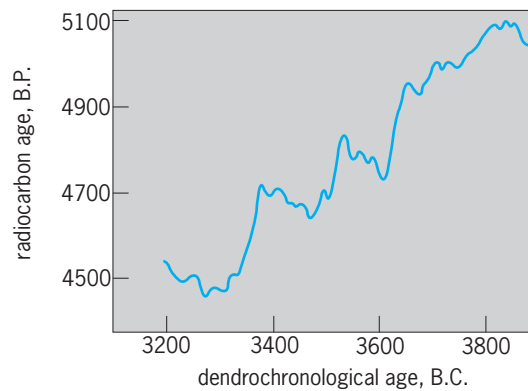

Fig. 5. Secular variation/De Vries effect. Example of short-term variations in ^{14}C activity; detail of De Vries effects using high-precision measurements for period of approximately 4500–5100 ^{14}C years B.P. (*After A. F. M. de Jong and W. G. Mook, Medium-term atmospheric ^{14}C variations, Radiocarbon, 22(2):267–272, 1980*)

logically based calibration data can be used to identify the general degree of deviation of ^{14}C values from real time and also the degree of maximum precision which is possible for specific temporal intervals.

The impact of the De Vries effects on the precision of ^{14}C values, as applied to archeological and historical problems, must be considered. The need to take into account these shorter-term variations have, for example, been demonstrated in the evaluation of ^{14}C values on twelfth- and fourteenth-century European medieval archeological materials. *See* ARCHEOLOGICAL CHRONOLOGY.

Lack of geographical and altitude variations. Concern has been expressed as to whether variation in the ^{14}C content of wood samples taken from a small number of localities in the Northern Hemisphere can be used to document worldwide secular variation effects. This question has been specifically answered as a result of studies of ^{14}C concentrations in tree rings from Patagonia, Canada, and Europe. The maximum deviations noted in contemporaneous woods were between those grown in the Southern and Northern hemispheres. However, even in this case, the variation did not exceed 0.5% or about the equivalent of 40 years.

Studies have also shown the lack of any significant altitude effect on ^{14}C concentrations in wood. A concern had been expressed that solar protons of appropriate energies would interact with nitrogen to form ^{14}C directly in wood samples growing at high elevations. The projected effect would be to inflate the ^{14}C content of a sample so that the ^{14}C age would appear to be significantly younger than its true age. However, calculations indicate that the maximum effect that could be obtained, assuming the most advantageous parameters, would not exceed 40 years. Other evidence suggests the actual effect is much less. The bristlecone pine wood of the White Mountains grows at about the 3350-m (11,000-ft) level. That there is no measurable ^{14}C produced as a function of altitude is indicated by the essential agreement in age between the high-altitude

bristlecone and low-altitude European and American sequoia samples of the same dendrochronological age.

Variability in radiocarbon distribution. An important feature of the ^{14}C method is its potential to provide directly comparable age determinations on a worldwide basis for a wide variety of organic samples. For this potential to be realized, ^{14}C, following its production, has to be mixed rapidly and completely throughout all of the carbon-containing reservoirs on a time scale not exceeding a few tens of years. To the degree that such conditions prevail, the contemporary ^{14}C content of all organic samples will be essentially identical. It was quickly determined that such is not the case. The initial ^{14}C content of samples could be significantly affected as a result of environmental conditions. A classic illustration of the problem was the discovery of living organisms from a fresh-water lake exhibiting ^{14}C ages of approximately 2000 years. In this case a large percentage of the carbon used by the organisms was derived from dissolved CO_2 from the limestone bed of the lake. The fictional age of the modern samples had been produced as a result of the dilution of contemporary ^{14}C activity by "dead" carbon (that is, containing no ^{14}C) from the limestone. Another example is provided from trees growing in active volcanic areas. The CO_2 emitted during volcanic discharges is characteristically depleted of its ^{14}C. Living trees exhibiting apparent ages as much as 1000 years from such environments have been reported.

One effect of the recognition that the geochemical environment of a sample can affect its initial ^{14}C concentration has been to cast doubt on the reliability of particular types of samples. The use of shells in ^{14}C studies has been affected, since a tradition arose that their use should be discouraged. Terrestrial shells (gastropods) from most fresh-water environments generally merit this negative evaluation, since they typically take up carbonate which is not in equilibrium with atmospheric ^{14}C. The reputation of marine shells was adversely affected primarily as a result of early experiences with shells taken from several archeological sites along the Peruvian coast. Marine shell samples were found to have ^{14}C values that exhibited an apparent age as much as 900 years greater than that of charcoal samples assumed to have been deposited contemporaneously. It was therefore assumed that marine shells would consistently yield anomalous values.

Subsequent studies showed that marine shells can yield generally acceptable values if the conventional ^{14}C values can be corrected for upwelling effects. By examining ^{14}C concentrations in shells collected alive in the period before nuclear testing contributed bomb ^{14}C, it was determined that many marine shells exhibited apparent ages ranging as high as 1200 years. Part of the reason has to do with the fact that ocean water depleted of ^{14}C by long residence times in the deeper parts of the ocean is periodically upwelled or brought to the surface and mixed with surface ocean water. The effect is to dilute the contemporary ^{14}C activity of the surface ocean near the westward-facing continental margins, resulting in a spurious apparent age for the organisms utilizing surface-ocean-water carbonates. Shells growing in locations adjacent to the outlets of major river systems whose water is depleted of ^{14}C as a result of exchange with limestone or other carbonate-bearing rocks can also give spurious apparent ages. This is probably the explanation for the false ages exhibited by shells growing in the Gulf of California (Colorado River discharge) and the northern part of the Gulf of Mexico (Mississippi River discharge). *See* CONTINENTAL MARGIN; UPWELLING.

Upwelling and other reservoir effects are highly variable depending on location and specific environmental conditions. For the western coasts of North and South America, for example, the magnitude of the upwelling effects can range from about 80 to 1000 years. It is most severe along the Peruvian coast, contributing to an explanation for the problematical ^{14}C values on marine shells from that region. Unfortunately, it is possible for shells from highly localized regions to exhibit a sizable range in apparent ages. Samples from the Galápagos Islands show a variation of about 350 years. Such a fluctuation in such a relatively small area emphasizes the fact that the magnitude of upwelling effects for any region must be carefully established by multiple sampling of closely spaced areas.

Problems similar to those associated with marine shells arise for a number of sample types and geochemical environments where contemporary samples may not be in equilibrium with the atmosphere. In each case, empirically derived values for the contemporary standard must be obtained for each sample type or locality, or both. In practice, this is accomplished by determining the degree of deviation from whatever contemporary standard is used to define modern or "zero ^{14}C age" samples. For example, specific values are required for marine shells from different oceanic regions, for fresh-water shells in specific terrestrial environments, and for Arctic and Antarctic specimens.

Variability in carbon isotope ratios. For the ^{14}C method, the basic physical measurement used to index time is the ^{14}C/^{12}C ratio. However, carbon has three naturally occurring isotopes. Variation in this ratio can be effected by influences other than the decay of ^{14}C. The most common problem occurs when carbon-containing compounds not indigenous to the original samples are physically or chemically introduced into the sample matrix resulting in the contamination of the sample. Usually less difficult to deal with are fractionation effects in which a variation in the stable carbon ratio translates into a change in the ^{14}C/^{12}C ratio.

Contamination. The sources and effects of the introduction of foreign organics into samples are complex; they depend on the nature and condition of the sample materials, the characteristics of the environment to which the samples were exposed, and the period of time over which the exposure occurred. Precautions exercised to avoid contamination effects are unique to each sample type and source locality.

A series of procedures to remove potential contaminants in samples has been established by research laboratories. Most sample preparation techniques are concerned with completely removing what is assumed to have not been present when the original sample died or was removed from exchange with its carbon reservoir. Samples such as wood and charcoal, which can be subjected to treatment with strong acids and bases to facilitate the removal of absorbed carbonates and soil humic and fulvic acids and other soluble soil organic matter, are preferred. Less desirable are cases where it is difficult to distinguish between contamination and the original sample as with various types of carbonate samples, such as tufa and caliche.

It is usually possible to infer the effect of known contamination effects on a given sample in terms of the direction that the age change will take for a given type of contamination, but the magnitude of the errors can be calculated only if the true age of the original sample, the age of the contaminant, and the percentage contribution of the contaminant are all known. Usually this is difficult to determine. However, with few exceptions, problems of contamination for samples with ages of less than about 10,000 years can be solved, usually by applying standard pretreatment approaches developed by ^{14}C laboratories. For materials with expected ages in excess of 10,000 years, sample contamination problems typically become more serious, and laboratories must exercise even more rigorous care in the pretreatment processes.

Fractionation effects. While all the isotopes of carbon follow the same chemical or physical pathway, the rate at which this occurs varies as a function of their difference in mass. The pioneering studies of Harmon Craig pointed to the need to consider variations in the stable isotope ratio (^{13}C/^{12}C) of samples to obtain precise ^{14}C values. Variations equivalent to up to several hundred years can result if ^{14}C values are not standardized in light of ^{13}C/^{12}C ratios. Fortunately, no significant fractionation effects are usually observed in standard sample materials such as charcoal or wood. Problems arise, however, when it is necessary to compare ^{14}C values from a variety of sample types such as grasses, grains, seeds, succulents, and marine carbonates, as well as standard terrestrial organics. In such cases, it is necessary to use the stable isotope ratios to correct the ^{14}C values onto a common scale.

Variability in amount of carbon. In addition to the variation in production and distribution of ^{14}C over time and within portions of various carbon reservoirs, variations may result from situations where carbon not in equilibrium with the contemporary standard values is added or removed from any reservoir. Two instances are well documented since they occurred within the last century as a result of human intervention in the carbon cycle. The first, beginning in the middle of the nineteenth century, is known as the industrial or Suess effect. The combustion of fossil fuels added enough "dead" ^{14}C to the atmosphere to result in the reduction by about 3% in biospheric

^{14}C activity. In the more recent atomic bomb or Libby effect, relating to the detonation of thermonuclear devices in the atmosphere in the early 1950s, large amounts of artificial ^{14}C were produced, almost doubling the amount of ^{14}C in the terrestrial biosphere. When combined with the late-eighteenth-century De Vries excurses, the Suess effect makes it difficult to distinguish ^{14}C concentrations within the last two centuries. This is one of the reasons why laboratories generally use 100 or 150 years as the minimum age which can be cited. It also explains why laboratories cannot use modern wood as a contemporary reference.

Half-life of radiocarbon. The fundamental constant which permits the conversion of a ^{14}C/^{12}C ratio into an age value is the half-life or decay constant of ^{14}C. Initially in the development of the method, Libby and collaborators used the value 5720 ± 47 as the half-life figure, but soon adopted the weighted average of three independently obtained measurements. The average value was 5568 ± 30 and this became identified as the Libby half-life. In 1962, at the 5th Radiocarbon Dating Conference at Cambridge, it was decided that 5730 ± 40 probably represented a more accurate approximation of the actual half-life. It was agreed, however, that the Libby half-life would be used in the calculation of conventional ^{14}C determinations. The stated reason was that any changes in the value would introduce unneeded confusion in the radiocarbon literature.

The issue of the correct half-life for ^{14}C has lost a considerable amount of its significance because of the discovery and documentation of similiar variation and De Vries effects. The existence of dendrochronologically documented relationships between ^{14}C age and calendar age, for samples up to about 10,000 years old, enables researchers to circumvent the problem of the actual ^{14}C half-life and proceed to calibrate these ^{14}C age values directly.

Statistical and contextual uncertainties. Most ^{14}C determinations are expressed in the form: age value (in ^{14}C years B.P.) $\pm$ statistical uncertainty. The age value is calculated by using the equation previously presented. The measurement uncertainty results from statistical considerations inherent in the random decay process characteristic of all radioactive isotopes. A date, for example, of 5600 ± 80 ^{14}C years B.P. reflects the fact that the count rate of the sample is about 50% of the modern reference standard (that is, it has decayed for a period of about one Libby half-life) and the age value is known to about 1% or 80 years. Statistical uncertainties in ^{14}C work are usually cited in terms of one standard deviation errors. The expression 5600 ± 80 is a shorthand manner of stating that there are two chances out of three that the age equivalent of the counting rate for this sample will be contained within the range 5520 to 5680. An accurate statement of the results of a ^{14}C determination must include a listing of the measurement uncertainty. In addition, some laboratories increase this value to take into consideration changes that affect counting conditions, such as drift of electronic

equipment and changes in barometric pressure. Statistical errors are not cited only when the counting rate of a sample is statistically indistinguishable from the background counting rate for the counter being used. The result is expressed as a minimum or infinite value by stating that the age is "greater than" a limit imposed by the characteristics of the counting system being employed (for example, >40,000).

The processing and counting of a sample to determine its ^{14}C age is a challenging analytical procedure. However, a technically correct value which has been carelessly collected may be scientifically worthless. Often the significance and importance of a ^{14}C determination are only as good as the attention to detail which went into documenting the geological, historical, or archeological context of the sample. It is important to be aware of what a ^{14}C date does and does not indicate. A ^{14}C value provides a temporal index of when the sample was removed from its reservoir. For example, a ^{14}C determination on a piece of charcoal or wood provides an age value for the tree rings which make up the sample. A ^{14}C date taken on a piece of wood taken from a beam excavated from a ruined structure may indicate the time when the building was constructed if the sample was taken from the outside rings of the tree used as the source of the timber and if the timber itself did not happen to be reused from an earlier structure. Problems such as these often confront geologists and archeologists as they attempt to critically interpret the dating evidence provided by the ^{14}C method.

R. E. Taylor

Bibliography. E. Bard et al., Radiocarbon calibration by means of mass spectrometric $^{230}Th/^{234}U$ and ^{14}C ages of corals: An updated base including samples from Barbados, Mururoa and Tahiti, *Radiocarbon*, 40:1085–1092, 1998; D. Polach, *Radiocarbon Dating Literature: The First 21 Years*, *1947–1968*, 1988; M. Stuiver, J. von der Plicht, and A. Long (eds.), Calibration Issue INTCAL 98, *Radiocarbon*, 40:1–1160, 1998; R. E. Taylor, Fifty years of radiocarbon dating, *Amer. Scientist*, 88:60–67, 2000; R. E. Taylor, Radiocarbon dating, in R. E. Taylor and M. J. Aitken (eds.), *Chronometric Dating in Archaeology*, 1997; R. E. Taylor, *Radiocarbon Dating: An Archaeological Perspective*, 1987; R. E. Taylor, R. Kra, and A. Long (eds.), *Radiocarbon After Four Decades: An Interdisciplinary Perspective*, 1992.

Radiochemical laboratory

A laboratory or facility for investigation and handling of radioactive chemicals which provides a safe environment for the worker and the public. It's features can vary depending on the type of radioactive emissions to be handled, the quantity, the half-life, and the physical form (solid, liquid, gas, or powder). Special measures to minimize spread of contaminated material and to dispose of radioactive waste are required. Working surfaces should be smooth and easily washable to permit effective decontamination if necessary. Good ventilation and detectors for monitoring radiation and contamination on surfaces or people are also typical features. Space for clothing change, lockers, and showers are commonly provided.

Investigations utilizing only very small amounts (a few microcuries) of beta or gamma emitters which are not readily dispersed (no powders or volatile liquids) may sometimes be performed without special facilities on the bench top. In this case, precautions such as working on plastic-backed absorbent paper (lightweight disposable diapers are useful) and wearing protective gloves and lab coat may be sufficient. A special bag or can for disposing of the paper and gloves as radioactive waste is required. If the radioactive isotopes are solely alpha-particle emitters, containment and isolation from direct contact are more serious concerns. Due to the limited penetration but high biological toxicity of alpha particles, it is essential to avoid ingestion or inhalation. For very small quantities, work may take place with double rubber gloves in a fume hood with appropriate filter. An air velocity greater than 125 ft/min (38 m/min) into the hood at the opening is necessary for suitable isolation. Generally, an enclosed glove box (**Fig. 1**) is used, situated inside a hood and maintained at negative pressure with respect to the face of the hood and the room. Sensors to monitor proper differential pressure and adequate airflow are usually used to assure containment and to generate an alarm if conditions degrade. Many alpha emitters decay into other radioactive isotopes which subsequently decay into still other radioactive species. Some members of these long decay chains are gamma-emitting or radon isotopes. If the quantity of radioactivity is high, shielding from the penetrating gamma rays and special traps for radon gas emanating from the materials are required.

For work with pure beta emitting isotopes, for example phosphorus-32, confinement within a hood and the use of indirect means of handling the radioactivity is recommended (**Fig. 2**). Long-handled tongs or other tools, sometimes with an attached piece of Plexiglass, approximately 3/16 in. (0.5 cm) thick, are used for higher levels of radioactivity in order to shield the hands. Generally, since most beta emission is also accompanied by penetrating gamma emission, the entire work area must be enclosed in heavily shielded enclosures. *See* ALPHA PARTICLES; BETA PARTICLES; RADIATION CHEMISTRY; RADIOACTIVITY; RADIOCHEMISTRY; RADIOISOTOPE.

Hot laboratories . These special radiochemistry laboratories contain walled enclosures for remotely handling larger quantities of gamma-emitting isotopes. A small enclosure is usually referred to as a cave, while large ones are called hot cells. Hot cells are usually equipped with remote manipulators and thick windows made from high-density lead glass (**Fig. 3**).

Construction. Many different shielding materials have been used for hot cells. High-density concrete is useful because of its relatively low cost. This shielding can be improved with the addition of iron powder to the concrete. When floor space is limited or more effective shielding is required, steel or lead can be
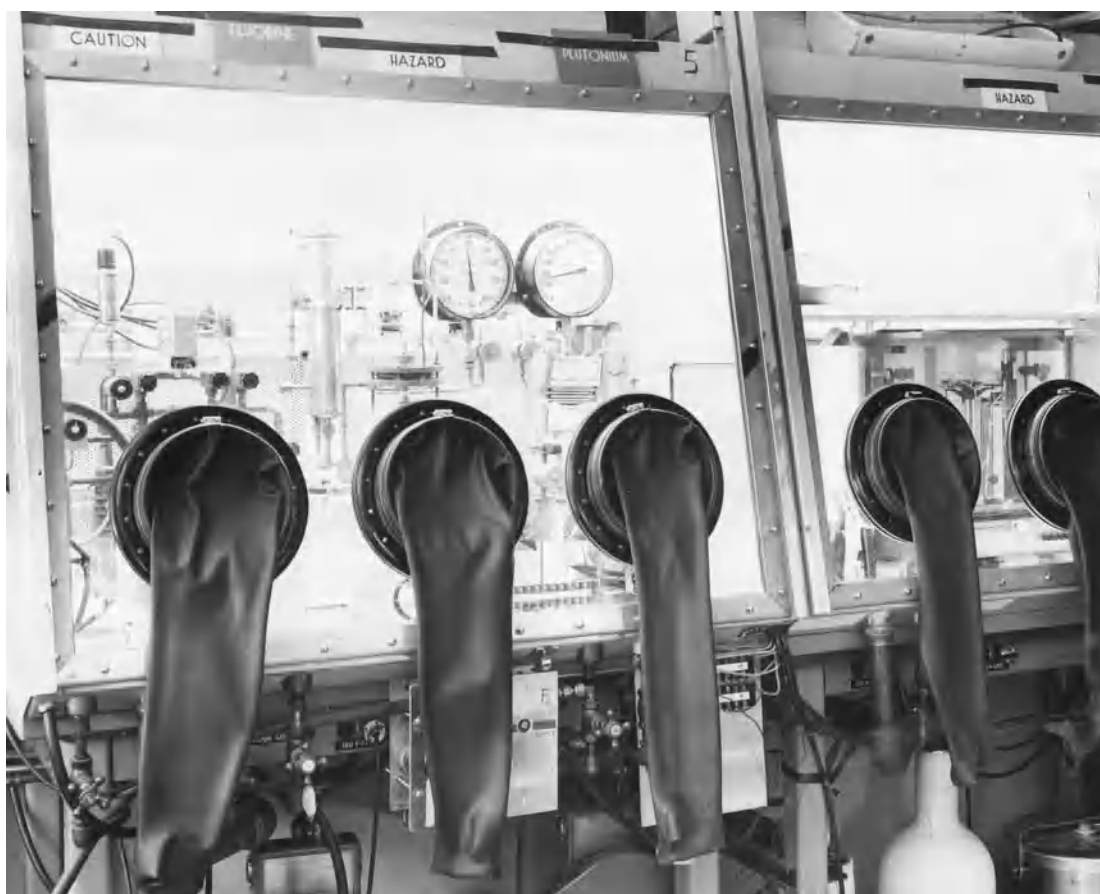
Fig. 1. Radiochemical laboratory, hood-equipped, with rubber-gloved openings. (*Argonne National Laboratory*)

Fig. 2. Fume hood set up with portable shielding and small tongs for working with low-level beta- and gamma-emitting isotopes. (*Brookhaven National Laboratory*)

used for walls and ceilings. Most commercially available hot cells utilize poured lead or interlocking lead brick encased in steel plate. S-shaped penetrations in the walls or ceiling are usually provided for electrical wires, and for tubing to allow water connections and reagent addition. This prevents any straight-line path for radiation to shine through. For disposing of aqueous radioactive waste, small cup sinks with shielded drain lines must be installed. Shielded storage tanks for this waste are usually installed under the floor or in a basement room. Some radiochemistry laboratories specialize in investigations with the short-lived positron-emitting isotopes fluorine-18 (half-life 128 min), carbon-11 (20 min), nitrogen-13 (10 min), and oxygen-15 (2 min). In this case, radioactive waste disposal is significantly simplified by just putting it aside and waiting for decay to turn it into normal waste. Thus, extensive storage areas for either solid or liquid radioactive waste are not necessary.

Hot-cell viewing windows are made from lead glass with a thickness providing equivalent shielding to the walls. Dopants containing cerium are often added to the innermost glass panes to help resist darkening and discoloration caused by high levels of radiation exposure. Since high-density lead glass is yellow in color, bright interior lighting is required for good visibility. It is good design to allow a light bulb to be changed from outside the hot cell. Sometimes closed-circuit television or magnifying

Fig. 3. Exterior view of two hot cells with manipulators. The walls are 7-in.-thick (17.75-cm) lead, encased in 1/4-in.-thick (0.6-cm) steel plate. The lead glass windows are 14 in. (35.5 cm) thick and provide shielding equivalent to the walls. (*Brookhaven National Laboratory*)

periscopes are provided for special inspection needs. Before first use, the shielding integrity of the entire assembly should be checked with a small size, high-level radioactive source. Special attention needs to be devoted to surveying radiation levels around the outside of wall-to-wall joints, conduit or reagent penetrations, and ventilation ducts. *See* RADIATION SHIELDING.

Operations. Hot cells are usually grouped together in order to provide separate processing areas for different isotopes. This minimizes the chances of cross contamination. However, it is efficient to have a single common hot cell to receive and open irradiated targets and to dispose of waste, as well as a single common hot cell with a porthole for dispensing of final product. The materials to be processed can be moved with manipulators through internal connecting portholes between the hot cells, or by conveyor belts or equivalent. In several installations, large-gauge toy electric trains serve this purpose. Most hot cells have small portholes or drawers for sample insertion and removal. In addition, there must be a larger opening, such as a door or a removable wall panel, for major maintenance, cleanup work, and equipment installation. The type of work carried out in hot cells is similar to that in conventional laboratories. This includes materials inspection, wet chemistry, and analytical measurements. The specific equipment must be inserted into the hot cell prior to use with hot sources. Since space is at a premium, selection and location of the equipment (such as balances, remotely operated pipettes, automatic bottle capping device and crimp sealers, and specialized glassware) are important.

A significant difference in operation exists between hot laboratories specializing in preparing radiopharmaceuticals with the short-lived positron emitters compared to those handling longer-lived radioisotopes. The very short lifetimes require very rapid processing and compound labeling. Also, as many such agents are used clinically in human diagnostic nuclear medicine procedures, they must be prepared several times a day, five or six days a week. This preparation burden is therefore very labor intensive and expensive. In response, the preparation of many of the standard, frequently used radiopharmaceuticals, such as F-18 fluorodeoxyglucose, has been automated into "black box" synthesis modules for hot cell use. Solutions are moved internally with small liquid pumps or by vacuum through the synthesis steps. These commercially available systems provide rapid, sterile preparations ready for human use. Alternatively, computer programmable robot arms can control synthesis by operating syringes, moving flasks, and other motions. This approach requires more setup effort than a black box assembly, but also provides more flexibility since the robot arm can be reprogrammed for different motions. *See* NUCLEAR MEDICINE; RADIOACTIVE WASTE MANAGEMENT; ROBOTICS.

Remote manipulators. Many hot-cell operations are performed using remote manipulators. Many varieties are now available in different sizes and capacities. The most generally useful is the mechanical master-slave manipulator. This device features a mechanical arm with a wrist and jaws at the end located inside the hot cell. The movements of this slave device precisely follow the movements of the operator's arm, wrist, and hand. Generally the force applied by the operator is directly transmitted without magnification up to the capacity limit of the device. Manipulators with lift capacity ranging 10–100 lb (4.5–45 kg) are the most common. Many manipulators are mounted such that the slave arm penetrates the ceiling of the hot cell through a rotating steel or lead shield ball. Since a wide range of motion is desirable, it is difficult to completely shield the area at the manipulator penetration. Small leaks through a ceiling area may be tolerable since it is not generally an occupied space. Another style uses a horizontal barrel mounted through the front face of a hot cell near the top, with offset shield plugs at both the inside and outside of the penetration. This style allows the inner slave arm to electrically swing forward and back, or swing left and right without moving the master arm. This extended reach capability is critical in large hot cells because it allows the operator to stay in a good viewing location while still reaching the back or corners. Both styles usually cover the slave arm with a plastic sleeve called a boot to minimize contamination of the gears and pulleys inside the arm. This can be important if repair of the manipulator outside the hot cell becomes necessary. Many manipulators allow the jaws or tongs to be remotely removed and changed. Although different styles can thus be accommodated, it is still usually necessary to design special tools or tool handles in order to

improve the reliability of gripping objects (especially small ones). *See* REMOTE-CONTROL SYSTEM; REMOTE MANIPULATORS.

Ventilation. Good ventilation is critical for radiochemical laboratories, particularly with hot cells, to sweep any airborne radioactivity away from the worker. The radioactivity may be a gas, dust, or fume. Filters are then necessary to trap this material and prevent it from entering the local atmosphere. Because dust is a good vehicle for spread of contamination, filtered inlet air can reduce spread of contamination and prolong the lifetime of outlet filters. The primary outlet filter required in such laboratories is the HEPA (high-efficiency particulate air) filter. HEPA filters must be 99.97% efficient, and are required to be tested annually. Typically a roughing filter is also used upstream to reduce the dust loading on the HEPA. In many hot-cell installations, an on-line differential pressure gauge continuously monitors the performance of the HEPA filter. A low-differential pressure value may indicate a rip in the filter, while too high a value would imply that the filter is clogged.

In a multi-hot-cell installation, proper air balance between the hot cells must be maintained to minimize cross contamination. Adequate airflow and static pressure must also be maintained when portholes or hot-cell doors are opened. This is ordinarily accomplished with either an oversized system, or with variable-speed ventilation fans which speed up in response to a drop in hot-cell airflow rate. In laboratories where use of acids is common, for example for dissolving solid irradiated targets, acid fumes can be sucked off into a separate system with an inverted funnel attached to a flexible hose. These fumes are then passed through a neutralizing buffer before joining the main ventilation duct. Charcoal filters are also required if the emission of gaseous radioactivity, for example iodine-131, is anticipated. Where performance of the ventilation system is crucial for radiation safety or environmental protection, electronic sensors with alarms are used to monitor total system flow rate, individual hood or hot-cell flow rate, static pressure, and filter performance. Backup electric power may be necessary to guarantee negative pressure in these urgent situations. *See* AIR FILTER; VENTILATION.

Safety. In addition to containment, radiation shielding, and good ventilation, several other safety measures are commonly employed. To limit spread of contamination outside the laboratory, radiation monitors are often placed at the doorway. These may be simple hand-held Geiger counters with a large-area detecting probe, called pancake detectors. Or in some radiochemistry laboratories, large step-in detector arrays, called portal monitors, can check the entire body for radioactivity. When high levels of radioactivity are present in hot cells, radiation sensors mounted inside can be interlocked with doors and portholes to prevent access. Area radiation monitors are often installed in the room as well as airborne radiation monitors. Both can initiate audible and visual alarms if trip points are reached.

See INDUSTRIAL HEALTH AND SAFETY; RADIATION BIOLOGY.   Leonard F. Mausner

Bibliography. K. J. Connor and I. S. McLintock, *Radiation Protection Handbook for Laboratory Workers*, 1994; *Manual on Safety Aspects of the Design and Equipment pf Hot Laboratories: 1981*, 1981; *Safe and Secure: Guide to Working Safely with Radiolabelled Compounds*, Amersham International, 1992.

Radiochemistry

A subject which embraces all applications of radioactive isotopes to chemistry. It is not precisely defined and is closely linked to nuclear chemistry. The widespread use of isotopes in chemistry is based on two fundamental properties exhibited by all radioactive substances. The first property is that the disintegration rate of an isotopic sample is directly proportional to the number of radioactive atoms in the sample. Thus, measurement of its disintegration rate (with a Geiger counter, for example) serves to analyze a radioactive compound. With nearly all chemical elements (the most notable exceptions being nitrogen and oxygen, which have no suitable radioactive isotopes), an isotope may be incorporated in a chemical compound, and thereafter, masses of this compound as small as 10^{-6} to 10^{-10} g may be measured with a high precision. Because experimental chemistry depends largely upon analysis, isotopes may be employed in most chemical problems, especially those requiring high analytical sensitivity. The second fundamental property is that the disintegration rate is completely unaffected by the chemical form of the isotope, and conversely, the property of radioactivity does not affect the chemical properties of the isotope. By substituting or labeling a particular atom within a molecule, isotopes can be used to trace the fate of that atom during a chemical reaction. In contrast to physical migration tracer studies, the compounds arising in a reaction must first be isolated in separated pure forms before radioactive assays can be performed. *See* RADIOACTIVE TRACER.

In general, radiochemical studies can be classified according to whether the use of isotopes represents a convenient or a unique solution to a problem.

Convenient applications. These applications usually exploit the high sensitivity of tracer techniques because alternative analytical procedures are slower and often less accurate. The efficiencies of chemical separations, such as those based on selective precipitation, solvent extraction, ion exchange, and electrodeposition reactions, are studied by labeling the desired compound and following the radioactivity during the separations. The rate and extent of adsorption on solid surfaces of either labeled solutions or labeled gases are rapidly determined by assaying periodically the mobile phase, or better, the solid phase. New chemical phenomena, such as the coprecipitation of trace elements and radiocolloid formation, occur at submicro concentrations (10^{-10} g/liter

of solution) and may be studied most conveniently with isotopes. The solubility of an "insoluble" precipitate is measured by saturating a solvent medium with a radioactive solid. Similarly, the vapor pressure of a solid is measured by saturating an evacuated volume with vapor or, for pressures below 10^{-4} mm, by effusing vapor into a cooled target, which is later assayed. Qualitative and quantitative analysis for most trace elements present in parts per million or less in a sample is possible by radioactivation analysis. The sample is irradiated in a flux of neutrons or other suitable particles, and the trace element is identified and determined by its induced activity. Depending upon the element, quantitative determination of masses of 10^{-8} to 10^{-12} g is usually possible. *See* NUCLEAR REACTION.

Unique applications. An understanding of diffusion processes is of considerable importance because the rates of many chemical reactions are governed by the rate at which chemical species can diffuse through a medium to the point of reaction. For example, the rate of many electrode processes depends upon the rate of diffusion of electrolyte to the electrode, and the rate of oxidation of copper is determined by the rate of diffusion of copper ions up to the metal surface. If a layer of radioactive copper is sandwiched between two ordinary copper samples, it is found that at elevated temperatures copper ions will diffuse considerable distances within the metal. The rate of diffusion of copper in copper (that is, the self-diffusion rate of copper) can be observed only by the transfer of radioactivity from the labeled region into unlabeled regions, thin slices being removed from the solid at known distances and then being assayed. The **illustration** shows some experimental points obtained in such an experiment. The distribution curve is that expected from the integrated
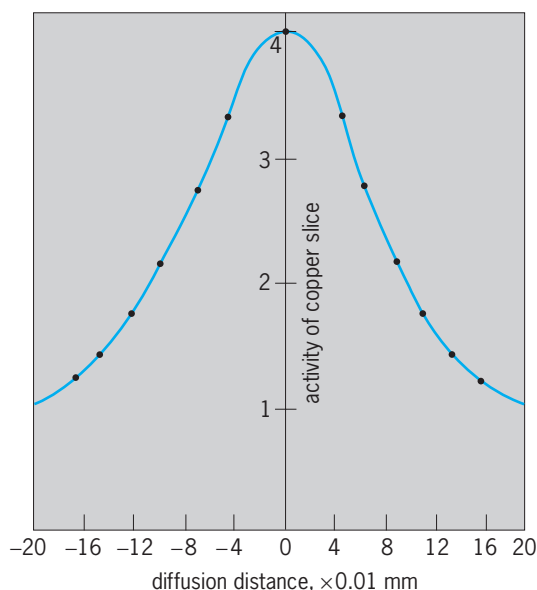
form of Fick's law of diffusion, Eq. (1), where c is

$$\frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2} \qquad (1)$$

the concentration (activity) of the diffusing tracer, t is the diffusion time, x the diffusion distance, and D the self-diffusion coefficient. In addition to solid-state studies with elements, alloys, metallic oxides, and inorganic salts (all of which have important metallurgical implications), self-diffusion experiments are performed with liquids and gases. In all cases, they provide valuable information on the nature of the intermolecular forces which determine the magnitude of D.

Isotopic exchange reactions. When slightly soluble lead chloride crystals are mixed with an aqueous solution of sodium chloride, labeled with chlorine-36, radioactivity rapidly appears in the lead chloride as a result of exchange of chloride ions between the two compounds. Both compounds produce chloride ions on dissociation, and some chloride ions, originating from the sodium chloride, become associated with the lead ions, thereby leading to radioactivity in the lead chloride, as in reaction (2), where s

$$PbCl_2(s) \longrightarrow PbCl_2(soln) \qquad Na\overset{*}{C}l(soln)$$
$$\searrow \qquad \swarrow$$
$$Pb^{2+} + \{2Cl^-, \overset{*}{C}l^-\} + Na^+ \qquad (2)$$
$$PbCl\overset{*}{C}l(s) \longleftarrow PbCl\overset{*}{C}l(soln)$$
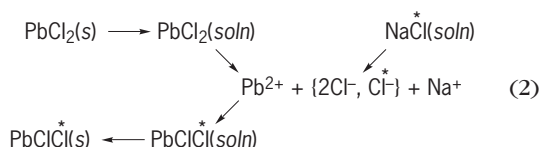
indicates solid and *soln* indicates solution. Exchange processes are proceeding continuously, but they can be detected only with isotopes, hence the term isotopic exchange reactions. Exchange reactions may occur between any two species of molecules having a common atom or group. They may be due to a dissociation process (as above) or to a collision between the two species in which chemical bonds are formed and broken (a bimolecular process), as in the exchange of radioactive sulfur (denoted by an asterisk) between sulfite and thiosulfate ions, reaction (3).

Distribution of radioactive copper after 605 min diffusion at 1742°F (950°C). The full curve corresponds to $D = 8.92 \times 10^{-10}$ cm²/s.

$$\begin{bmatrix} O \\ | \\ O-\overset{*}{S} \\ | \\ O \end{bmatrix}^{2-} + \begin{bmatrix} O \\ | \\ S-S-O \\ | \\ O \end{bmatrix}^{2-} \longrightarrow$$

Initial state

$$\begin{bmatrix} O \quad\quad O \\ | \quad\quad | \\ O-\overset{*}{S}\cdot\cdot S\cdot\cdot S-O \\ | \quad\quad | \\ O \quad\quad O \end{bmatrix}^{4-} \longrightarrow$$

Bond   Bond
forms   breaks

$$\begin{bmatrix} O \\ | \\ O-\overset{*}{S}-S \\ | \\ O \end{bmatrix}^{2-} + \begin{bmatrix} O \\ | \\ S-O \\ | \\ O \end{bmatrix}^{2-} \qquad (3)$$

Final state
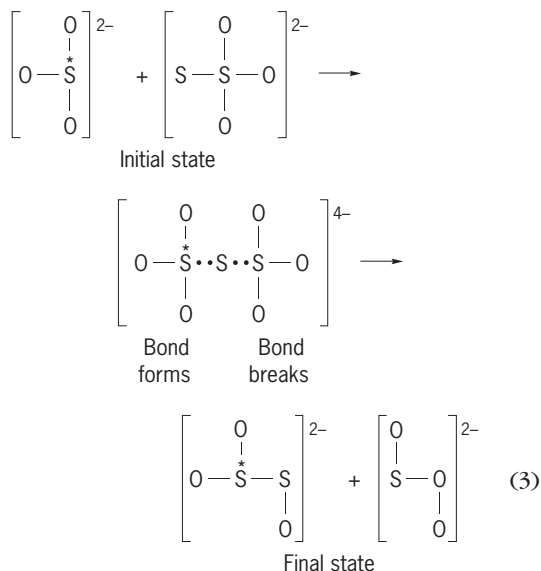
With two oxidation states of an element, exchange occurs by the transfer of an electron (an electron exchange reaction), for example, reaction (4), although

$$Fe^{II} + \overset{*}{F}e^{III} \rightarrow [Fe^{II} \xrightarrow{\ e\ } \overset{*}{F}e^{III}] \rightarrow Fe^{II} + \overset{*}{F}e^{II} \qquad (4)$$

in many cases the electron may actually be transferred by an atom or a group bridging between the reactants. The rates of exchange reactions are measured by separating the two reactants at various times and determining the fraction F of the total radioactivity in each species. For an initially unlabeled species, F increases exponentially with time to an equilibrium value corresponding to equal concentrations (specific activities) of isotope in each species; for an initially labeled species, F will decrease exponentially with time. As in ordinary chemical kinetics, the exchange mechanism is deduced from the dependence of the exchange rate upon reactant concentrations. Rate studies of exchanges by dissociation provide information concerning ionization (including acid-base equilibria) of both solutes and solvents, the reaction of solvents with molecules (solvolysis), the thermal dissociation of gases, and the dissociation of gases on catalyst surfaces. Bimolecular exchanges are exceptionally important in studies of oxidation-reduction reactions and of substitution reactions of coordination complexes of the transitional elements. The rates of exchange by substitution provide a direct measure of the lability of these coordination complexes.

Isotopic tracer studies. Details of reaction mechanisms are provided by labeling a specific atom within a molecule. Thus, when carboxyl-labeled propionic acid is oxidized in acid dichromate, the product, as anticipated, is only radioactive carbon dioxide, reaction (5). However, the oxidation mechanism is

$$CH_3CH_2\overset{*}{C}OOH \rightarrow HOOC{-}COOH + \overset{*}{C}O_2 \qquad (5)$$
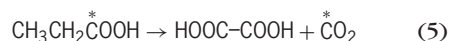
more complex in the case of alkaline permanganate because the isotopic distribution is 75% labeled oxalate and only 25% carbonate. When the masses of the labeling and the normal isotopes differ markedly (notably with hydrogen, carbon, and oxygen), the isotopic molecule will react more slowly than the normal molecule if the reaction mechanism involves significant stretching of the chemical bond to the isotopically substituted atom. Such isotope-effect studies provide mechanistic information.

New elements. The artificially produced transuranium elements and technetium, astatine, and francium are all radioactive, and their chemistry is being elucidated with radiochemical techniques. However, this topic is normally accepted as nuclear chemistry.

Recoil studies. For a discussion of recoil studies *see* NUCLEAR CHEMISTRY; RADIATION CHEMISTRY; RADIOACTIVITY; RADIOCHEMICAL LABORATORY.

Donald R. Stranks

Bibliography. R. A. Faires and G. G. Boswell, *Radioisotope Laboratory, Techniques*, 1980; C. Keller, *Radiochemistry*, 1988; R. Lambrecht, *Nuclear and Radiochemistry*, 1982.

Radioecology

The study of the fate and effects of radioactive materials in the environment. As a hybrid field of scientific endeavor, it is founded upon, and derives its basic principles from, both of its parent disciplines, that is, basic ecology and radiation biology. Following the discovery of ionizing radiation and radioactive particles in biological studies in the 1940s and 1950s, these phenomena were studied under what were usually controlled laboratory conditions. Soon after, however, a need was demonstrated for a better understanding of the fate and effects of radioactive materials that were being released into the environment following the use or testing of nuclear weapons. This need became an important factor in the emergence of radioecology as a scientific discipline in its own right.

Radiation effects. Because of its early association with nuclear weaponry, radioecology was initially focused on radiation effects, usually under the umbrella of concern for survival under postnuclear attack conditions. Mortality induced by radiation exposure has usually been measured by using some form of lethal dose 50. However, the amount of radiation exposure required to produce mortality under free-living conditions in the field (ecological lethality) may often be substantially less than that required to cause physiological lethality when the organism is held under protected conditions in the laboratory. In the case of young bluebirds (*Sialia sialis*), for example, exposure to radiation levels as low as 800–900 roentgens (R) causes the stunting of wing feather growth, which would reduce the bird's ability to fly and escape predators. Therefore mortality in the field would result from these levels of exposure. When birds are hand-raised and protected from predators in the laboratory, however, exposure to as much as 2500 R would be required to produce 50% mortality. *See* LETHAL DOSE 50.

Similar considerations also apply to sublethal radiation effects, such as perturbations of growth rate, reproduction, and behavior. These responses to radiation stress have consequences for both the individual organism and for the population, community, or ecosystem of which it is a part. When populations or individuals of different species differ in their sensitivities to radiation stress, for example, the species composition of the entire biotic community may be altered as the more radiation-sensitive species are removed or reduced in abundance and are replaced in turn by more resistant species. Such changes have been documented by studies in which natural ecological systems, including grasslands, deserts, and forests, were exposed to varying levels of controlled gamma radiation stress (**Fig. 1**). *See* POPULATION ECOLOGY.

Techniques of laboratory toxicology are also available for assessing the responses of free-living animals to exposure to low levels of radioactive contamination in natural environments. This approach uses sentinel animals, which are either tamed, imprinted on the investigator, or equipped with miniature
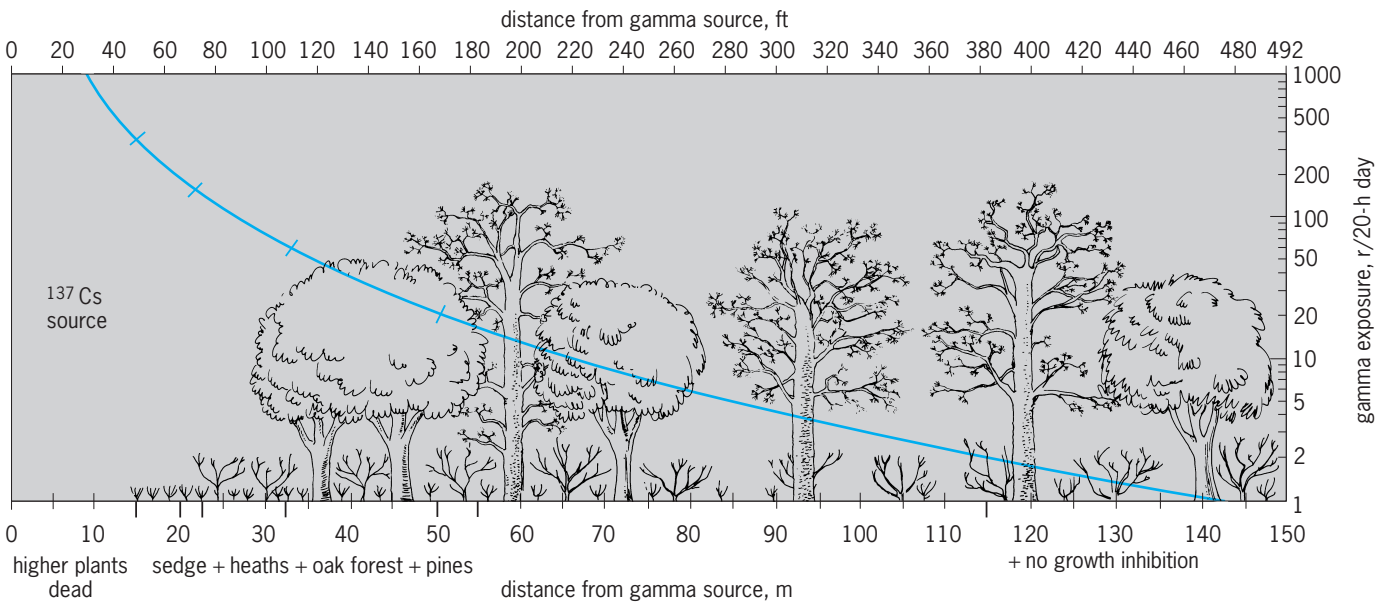
Fig. 1. Gradations in response of the vegetation of an oak-pine forest to continuous exposure to gamma radiation stress. The radiation source was located at the left side of the figure and, as indicated by the curve, exposure levels decreased from left to right with increasing distance from the source. (*After G. M. Woodwell, Radiation and the patterns of nature, Science, 156:461–470, 1967*)

radio transmitters, to permit their periodic relocation and recapture as they forage freely in the food chains of contaminated habitats. When the animals are brought back to the laboratory, their level of radioisotope uptake can be determined and blood or tissue samples taken for analysis. In this way, even subtle changes in deoxyribonucleic acid (DNA) structure can be evaluated over time. These changes may be suggestive of genetic damage by radiation exposure. In some cases, damage caused by a radioactive contaminant may be worsened by the synergestic effects of other forms of environmental contaminants such as heavy metals.

In other cases, assessment of responses to radiation exposures has documented the phenomenon of hormesis, whereby an organism responds to very low-level exposures in the opposite way from which it would respond to higher levels of radiation stress. For example, while growth rates of plants and animals normally decline after high levels of radiation stress, exposure to low levels of radiation may actually result in a stimulation of growth. Moreover, the documentation of such effects may suggest that some physiological processes may be adapted to perform optimally under very low levels of background radiation exposure, such as are provided naturally from cosmic rays and natural sources of radioisotopes in certain rock and soil substrates. The nuclear accidents and other releases of radioactive contaminants at various sites in the former Soviet Union have provided natural laboratories for ecological studies of the effects of radiation contamination.

Radionuclide tracers. Because of the ease with which they can be detected and quantified in living organisms and their tissues, radioactive materials are often used as tracers to study the rates and patterns by which biological processes take place

under natural conditions. The use of such radioactive tracers in medicine is well known. In the same way, radioactive tracers can also be used to trace food chain pathways or determine the rates at which various processes take place in natural ecological systems (**Fig. 2**). Gamma-emitting isotopes are particularly useful in this regard, since their emissions from deep within the body of a plant or animal allow them to be detected and quantified by sensitive counting
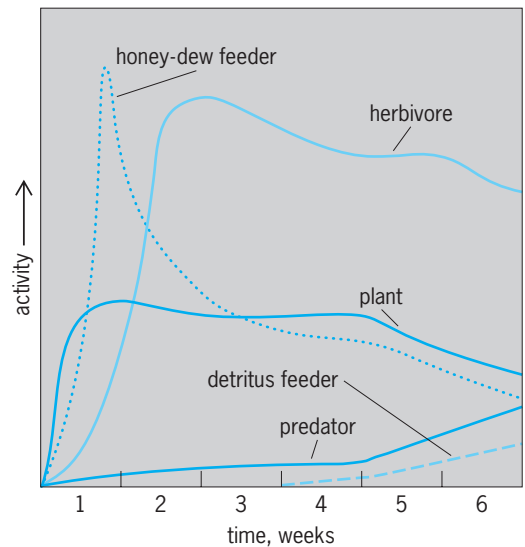
Fig. 2. Varying patterns of uptake of radioactivity by organisms occupying different trophic levels in an old-field ecosystem, following the introduction of radioactive tracers into the plants upon which their food web was based. Animals that suck plant juices and directly consume the vegetation acquired maximum activity levels earlier than did predators and detritus feeders. (*After R. G. Wiegert, E. P. Odum, and J. H. Schnell, Forb-arthropod food chains in a one-year experimental field, Ecology, 48:75–82, 1967*)

equipment without the need for sacrifice and dissection. Other uses of radioactive tracers include tagging adult animals in order to later identify the particular eggs or young they produce, and using radioactive markers to later relocate and recapture animals that are too small to carry even the tiniest of radio transmitters.

Although most of these tracer experiments were performed in the past by deliberately introducing a small amount of radioactive tracer into the organism or ecological system to be studied, they now take advantage of naturally tagged environments where trace amounts of various radioactive contaminants were inadvertently released from operating nuclear facilities such as power or production reactors or waste burial grounds. As long as the amounts of such releases are less than the amounts that could be considered hazardous to health, such low-level contaminated habitats have actually become valuable research sites for conducting studies of functional processes associated with the native flora or fauna of the area. *See* RADIOACTIVE TRACER.

Radioactive contamination. In some cases, the inadvertent release of radioactive materials into the environment creates concern for the health and well-being of humans and other organisms living in the area. An important component of radioecology, and one that is closely related to the study of radioactive tracers, is concerned with the assessment and prediction of the movement and concentration of these radioactive contaminants in the environment in general, and particularly in food chains that may lead to humans. Primary concern is, of course, focused on agricultural pathways through which radionuclides may enter the human food chain through crop plants or meat, milk, or other products of domestic livestock. Often overlooked, however, is the transfer of radioactive contaminants to humans who may consume fish or wild game as food. In many parts of the world, wild game animals such as deer have been found to have consistently higher concentrations of environmental radioactive contaminants than domestic livestock grazing fertilized pastures in the same region. This phenomenon is most noticeable in regions where natural vegetation has difficulty obtaining sufficient nutrients from the soil and must rely more heavily on "fallout" nutrients distributed by atmospheric processes. In high alpine regions and arctic tundra, for example, the concentration of radioactive contaminants in species such as reindeer may be particularly severe, as was observed following the Chernobyl nuclear accident in the Soviet Union. *See* FOOD WEB.

The Chernobyl accident was also important because it demonstrated the potential for global transport of radioactive contaminants by both physical forces, such as atmospheric circulation and meteorological factors, and biological vectors such as waterfowl, which might accumulate contaminants at the site of accidental release and then distribute them elsewhere along their migratory journeys. Studies of the migratory habits of these birds suggest that such biological transport of contaminants may not always follow the same patterns and rates as the distribution of contaminants by physical forces. This, in turn, emphasizes the importance of thoroughly understanding all aspects of the basic ecology and natural history of organisms involved in important issues of radioecological study and concern.

Studies in radioecology, when conducted under field conditions, often produce results that are either unexpected or may reverse the conventional wisdom derived from studies conducted under controlled laboratory conditions alone. An example of the latter is the failure of data from some radioecological field studies to support the popular notion of biomagnification, the systematic increase in levels of contaminant concentration in higher levels of food chains. Although some other environmental contaminants such as pesticides or heavy metals may behave in this fashion in natural food webs, such is certainly not always the case in situations documented for radioactive materials (**Fig. 3**).

Finally, situations where the environmental behavior of radioactive contaminants fails to agree with the predictions of laboratory studies demonstrate how radioecological research can contribute new information about some of the ecological mechanisms responsible for contaminant uptake, cycling, and transport in a more general sense. This information can help to provide a better understanding of the environmental behavior of other forms of contaminants such as pesticides or heavy metals, which, in contrast to radionuclides, are more difficult to detect and measure under natural
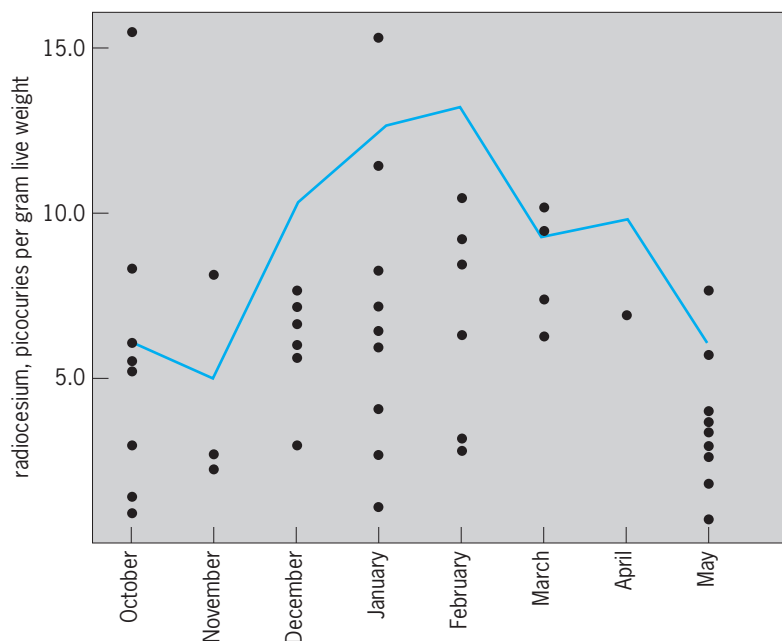

Fig. 3. Monthly changes in levels of radioactive contaminants accumulated by migratory waterfowl wintering on a nuclear-production-reactor cooling reservoir. Contamination levels shown by the strictly herbivorous American coot (*Fulica americana*; curve) nearly always exceeded those shown by other more omnivorous and carnivorous species (solid circles). (*After I. L. Brisbin, Jr., R. A. Geiger, and M. H. Smith, Accumulation and redistribution of radiocaesium by migratory waterfowl inhabiting a reactor cooling reservoir, in Environmental Behavior of Radionuclides Released in the Nuclear Industry, International Atomic Energy Agency, 1973*)

field conditions. *See* ECOLOGY; ENVIRONMENTAL RA-
DIOACTIVITY; ENVIRONMENTAL TOXICOLOGY; RADIA-
TION BIOLOGY.   I. Lehr Brisbin, Jr.

Bibliography. M. Eisenbud, *Environmental Ra-
dioactivity: From Natural, Industrial and Mili-
tary Sources*, 4th ed., 1997; M. Garcia-Leon and
G. Madurga (eds.), *Low Level Measurements and
Their Application to Environmental Radioactivity*,
1988; E. Holm, *Radioecology*, 1994; T. D. Luckey,
Hormesis with Ionizing Radiation, 1980; E. P.
Odum, Feedback between radiation ecology and gen-
eral ecology, *Health Physics*, vol. 11, 1985; V. Schultz
and F. W. Whicker, *Radiological Techniques*, 1982;
F. S. Sterrett (ed.), *Environmental Sciences*, 1987;
F. W. Whicker and V. Schultz, *Radioecology: Nuclear
Energy and the Environment*, 2 vols., 1982.

Radiography

The technique of producing a photographic image
of an opaque specimen by the penetration of radia-
tion such as gamma rays, x-rays, neutrons, or charged
particles. When a beam of radiation is transmitted
through any heterogeneous object, it is differentially
absorbed, depending upon the varying thickness,
density, and chemical composition of this object. The
image registered by the emergent rays on a photo-
graphic film adjacent to the specimen under exam-
ination constitutes a shadowgraph or radiograph of
its interior. Radiography is the general term applied
to this nondestructive film technique of testing the
gross internal structure of any object, whether it be
of the chest of a patient for evidence of tuberculosis,
silicosis, heart pathology, or embedded foreign ob-
jects; of bones in case of fractures or of arthritis or
other bone diseases; or of a weld in a pipe to observe
cracks, inclusions, or voids. Radiography with x-rays
is commonly used in both medical and industrial ap-
plications. Industrial work also involves gamma and
neutron radiography. Radiography with charged par-
ticles is under development. Most of this discussion
will be concerned with radiography with x-rays and
gamma rays. *See* CHARGED PARTICLE BEAMS; GAMMA
RAYS; NEUTRON; X-RAYS.

Industrial radiography enables detection of inter-
nal physical imperfections such as voids, cracks,
flaws, segregations, porosities, and inclusions. It is
frequently used for visualization of inaccessible in-
ternal parts in order to check their location or con-
dition. It is extensively applied wherever internally
sound metallic components are required such as
(1) in the foundry industry to guarantee the sound-
ness of castings; (2) in the welding of pressure ves-
sels, pipelines, ships, and reactor components to
guarantee the soundness of welds; (3) in the man-
ufacture of fuel elements for reactors to guarantee
their size and soundness; (4) in the solid-propellant
and high-explosives industry to guarantee the sound-
ness and physical purity of the material; and (5) in
the automotive, aircraft, nuclear, space, oceanic, and
guided-missile industries, whenever internal sound-
ness is required.

The general term applied to radiation imaging
and inspection is radiology. This includes film and
similar photographic image methods, such as radio-
graphic paper, under the term radiography. In medi-
cal circles, the term roentgenography, derived from
the name of the discoverer of x-rays, W. C. Roent-
gen, is used. The older technique of registering an
image on a fluorescent screen is called fluoroscopy.
The fluoroscopic prompt-response imaging of radi-
ation has largely been replaced by electronic detec-
tion with image intensifiers or sensitive television
cameras. This technique, called radioscopy, is now
widely used in both medical and industrial applica-
tions. *See* LIGHT AMPLIFIER; RADIOLOGY; TELEVISION
CAMERA TUBE.

Other variations of radiation imaging include xe-
roradiography, microradiography, flash radiography,
and computerized tomography. Xeroradiography is
a dry-plate, electrostatic image method similar to
that used in photocopy machines. Microradiography
involves a magnified image to improve spatial res-
olution and the detection of small detail. Flash ra-
diography is the production of an x-ray image in
a very short time, of the order of nanoseconds, in
order to stop fast motion. The computerized tomog-
raphy (CT) method recreates an image that is es-
sentially a slice through the object. Computerized
tomography has made a strong impact on medi-
cal diagnosis and industrial inspection. *See* COMPUT-
ERIZED TOMOGRAPHY; MICRORADIOGRAPHY; PHOTO-
COPYING PROCESSES.

In many cases, radiographic inspection is a satisfac-
tory method, but there is often a need for comple-
mentary, nondestructive inspection methods, such
as ultrasonics (also in medicine), magnetic-particle
inspection, and electromagnetic infrared and micro-
wave test methods. *See* NONDESTRUCTIVE EVALUA-
TION.

Absorption laws. X-radiography is a straightfor-
ward application of the well-known exponential ab-
sorption law given below, where I_0 is the initial

$$I_x = I_0 e^{-\mu x} = I_0 e^{-(\mu/\rho)\rho x}$$

intensity of the x-ray beam; I_x the intensity after pas-
sage through the object of thickness x, ρ the density,
μ the linear absorption coefficient, and μ/ρ the mass
absorption coefficient. The ratio μ/ρ is a function of
the atomic number of the absorbing chemical ele-
ment and of the wavelength of the x-ray beam. The
value of μ/ρ for a compound is equal to the sum
of the weight-proportional values for each of the el-
ements present. The penetrating ability of x-rays in-
creases as the wavelength of the rays decreases (hard
x-rays). The x-rays of larger wavelength (soft x-rays)
are relatively less energetic (less penetration power).
The coefficient μ/ρ is proportional to the cube of the
wavelength. The product ρx is mass per unit area. *See*
ABSORPTION OF ELECTROMAGNETIC RADIATION.

This exponential law and related expressions
quantitatively predict why bones absorb more radia-
tion than soft tissues and thus are clearly delineated,
or why cracks or blowholes in castings absorb to a

TABLE 1. Types of x-ray tubes and power sources used in radiography

Energy of radiation	X-ray tube	Power source
1–50 keV	Conventional	High-tension transformer
50–250 keV	Conventional	Regular or resonance transformer; gas-filled, portable with tube integral
250–500 keV	Conventional	Cascade or resonance transformer
1–3 MeV	Sealed tube with high potential applied over gradient in center of high potential generator	Resonance transformer or Van de Graaff generator, with tube integral
1–20 MeV	Tuned resonant microwave waveguide	Linear accelerator

lesser degree than the solid metal and thus are disclosed on the radiograph.

Equipment. Because of the wide range of materials subjected to radiographic inspection, a wide array of radiation-emitting sources is required.

X-ray equipment. Somewhat arbitrarily the x-ray equipment and operating conditions may be classified according to energy of radiation (**Table 1**). X-ray equipment with an energy up to about 300 keV can be considered portable. However, linear accelerators have been mounted on trailers. *See* PARTICLE ACCELERATOR; X-RAY TUBE.

Gamma-ray sources. Natural radioactive isotopes, such as radium, and many artificial radioactive isotopes are used in industrial radiography (**Table 2**). The artificial isotopes are obtained by bombarding the respective inert isotopes with an intense neutron flux in a nuclear reactor. The most commonly used radioactive sources are iridium-192 and cobalt-60. Medical therapy is also accomplished with radioisotope sources. *See* NUCLEAR MEDICINE; RADIOACTIVITY; RADIOACTIVITY AND RADIATION APPLICATIONS; RADIOISOTOPE (BIOLOGY); RADIOISOTOPE (BIOLOGY).

General principles and techniques. Some of the more important general principles and techniques common to radiography follow.

Exposure techniques. Thickness, physical density, and absorption coefficient of the materials under investigation determine the most suitable radiation energy (**Fig. 1**).

Film radiography. Since radiographs are usually interpreted from film negatives, each object under examination must be subjected to a controlled technique, and properly selected film in terms of contrast, latitude, and sensitivity. Optimum density, D, or blackening is 1.5–2.5, where $D = \log L_0/L$. The photometrically measured light intensities are L_0 before and L after the passage through the film. The normal eye can detect with certainty a minimum blackening dif-

ference between adjacent areas of 0.02. Experienced radiographers can detect density differences as small as 0.005. *See* PHOTOGRAPHY.

Contrast is the difference in density produced by a change in object thickness. Latitude is the extent of object thickness that can be reproduced in the working range of the film density. Sensitivity is the smallest fractional increase in thickness detectable; in industrial radiography such changes are usually in the range of 0.5–3%. Definition is the fidelity (or sharpness) with which a radiograph delineates a sharply defined inhomogeneity.

For better interpretation of the size, shape, and location of discontinuities, two or more radiographic views of the object are obtained under predetermined angles to each other. Stereoradiography is also used for this purpose.

If the thickness of the specimen varies, complete coverage can be obtained by one of three methods: (1) various radiographs may be taken under exposure conditions as determined by the average thickness of the respective section of the object; (2) the multiple film method may be applied by placing a combination of fast and slow films in the same film holder; and (3) compensators may be used by building up the thinnest parts of the object to an average uniform thickness with similar radiographically opaque material.

Target-to-specimen distance. For conventional x-ray tubes, this should be as great as possible, consistent with the fact that radiation intensity decreases inversely as the square of the distance, in order to decrease the unsharpness of images caused by penumbra (**Fig. 2**). If F represents the size of the focal spot of the x-ray tube; d, the distance from the focal spot to the specimen; and t, the distance from defect to film; then $U = Ft/d$ is the additional width of the penumbral shadow (or unsharpness). As discussed below, however, when microfocus x-ray sources are used, the specimen may be placed closer to the x-ray source in order to magnify the image.

Focal spot or source size. This should be as small as possible, to gain emission from (as nearly as possible) a point source, since otherwise there is unsharpness from penumbra (Fig. 2). However, radiation intensity tends to decrease as the focal spot decreases.

Scattered and secondary radiation. Most of this type of radiation is produced by the impact of primary radiation on objects, such as stands, walls, or test specimens. It travels in directions other than the

Table 2. Radioactive sources for industrial radiography

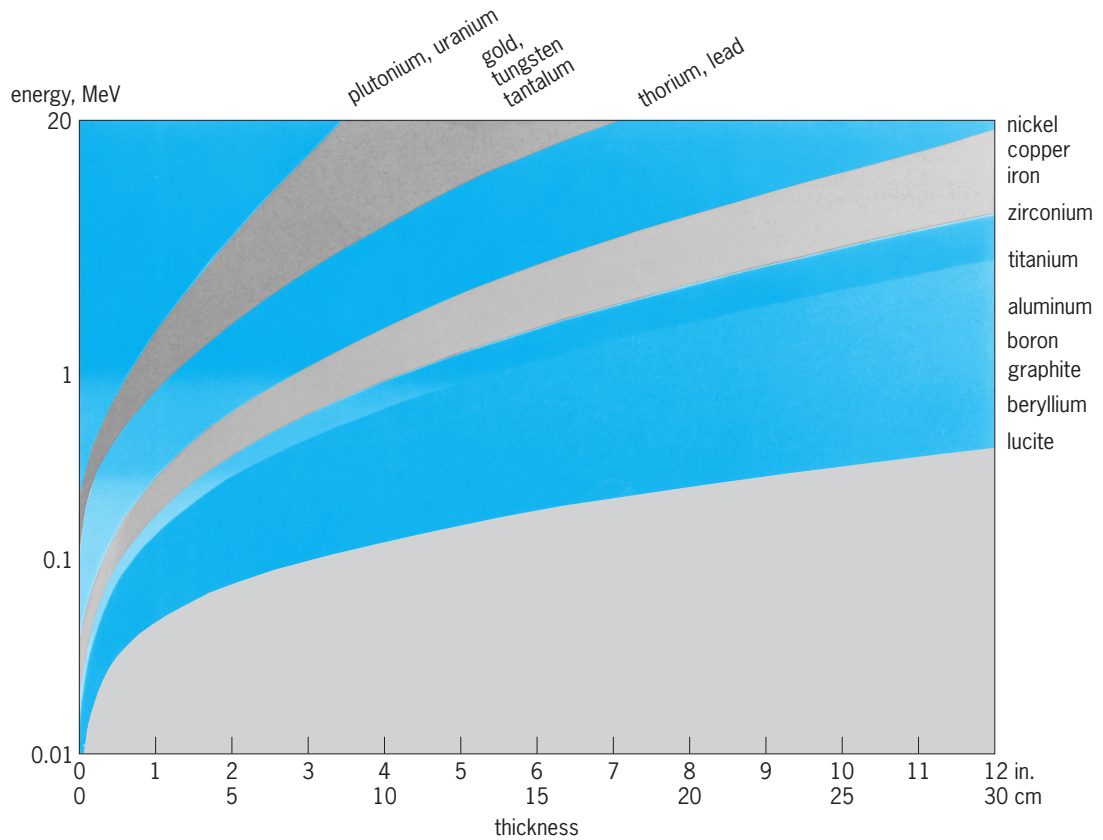| Element or isotope	Half-life	Steel thickness normally inspected
Radium	1580 years	3–8 in. (8–20 cm)
Cesium-137	33 years	1–4 in. (2.5–10 cm)
Cobalt-60	5.3 years	1.5–5 in. (4–13 cm)
Iridium-192	75 days	0.5–2.5 in. (1.3–6 cm)

Fig. 1. Relation of the penetration characteristics of x-rays to the type and thickness of various materials.

primary radiation which originates in the x-ray tube or isotope source. It contributes to film fogging and poor radiographic definition; it can be controlled by lead screens (also simultaneously used as intensifying screens), diaphragms, lead sheets, and increased object-to-film distance. An advantage of microradiography and geometric enlargement, discussed below, is that the detection of scatter from the specimen is reduced because of the increased distance of the specimen from the detector.
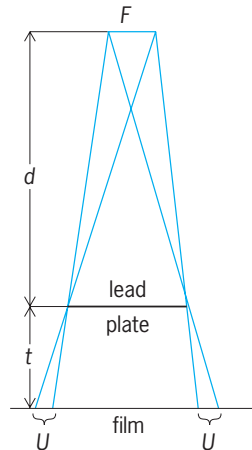
Fig. 2. Geometry of radiographic image formation. F = size of focal spot of x-ray tube; d = distance from focal spot to specimen (lead plate); t = distance from plate to film; U = unsharpness.

Radiographic inspection. In the field of medical radiography, many special techniques are used to visualize the contours of organs and soft tissues. Staining with heavily absorbing chemicals (iodine, barium, and thorium compounds) is a common procedure.

In industrial radiography the adequacy of the radiographic technique is determined by the use of penetrameters or image quality indicators. They usually are plates of the same material as the object under investigation, and their thickness is in definite proportion to the object thickness (normally 2%). They have centrally located holes with diameters of one, two, and four times their thickness and are normally placed on the source side of the object. The visualization of the smallest hole determines the radiographic sensitivity. Radiographic inspection procedures, as accepted by private and governmental industry or technical organizations, normally require the use of such penetrameters (**Fig. 3**).

Stereoscopy. Two exposures are made at tube positions roughly corresponding to interpupillary distances ($^{29}/_{16}$ in. or 57 mm). The two negatives are observed in a stereoscope which fuses the two images into one with the illusion of a third dimension. This permits location of any detail of gross structure (defect in a casting, bullet in a chest, location of cancer) below the surface. Electronic, real-time stereoscopy has also been accomplished by changing the x-ray image perspective in time intervals compatible with the $^{1}/_{30}$-s television frame time. This can be
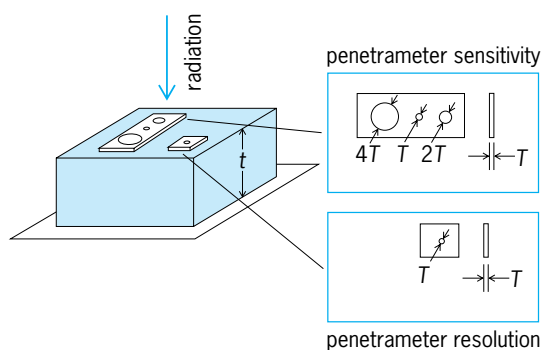
Fig. 3. Penetrameter. The windows show plan views and cross sections of the penetrameter plates. *T* is plate thickness; *t* is object thickness.

done, for example, by rotating the inspection object in the x-ray imaging beam or by electronically moving the x-ray focal spot in a microfocus x-ray tube set up to image with geometric magnification. *See* STEREOSCOPY.

Tomography. Of the three components—tube, subject, and film—two are moved during exposure synchronously so that it is possible to register the radiographic image of one plane in the object while images outside this slice have a continuous relative displacement and are blurred. A series in various planes enables three-dimensional exploration called serioscopy.

Tomographic information can also be obtained by computerized transverse axial scanning. In this technique, use is made of collimated x-ray beams, electronic detectors, and computer reconstruction of images. The radiation beam is made to pass through the object in many directions by rotating the collimated source around the object, or rotating the object. Each ray passing through the object is detected in a detector array. The attenuation in the many radiation paths is analyzed to reconstruct a density image of the inspected object. A computerized tomography image of a tree, for example, would show the growth rings as though the tree had been cut in that location to expose a visual image.

One system, which has been extensively used in medical diagnosis, utilizes 160 linear measurements across a slice of the object, these being taken at 180 one-degree intervals around the object. These 28,800 (180 × 160) measurements are used to reconstruct computer images. The high sensitivity of this method (0.5% variations in density) has enhanced the diagnosis of brain and head malfunctions.

Industrial applications of computerized tomography include inspections of rocket motors and missiles to detect inhomogeneities, voids, cracks, delaminations, and so forth, and of turbine blades to detect flaws and to measure internal passages and details. This use of computerized tomography as a dimensional measurement approach for internal details shows great promise. A computerized tomography system to inspect automobile engine block castings has been placed in operation for dimensional measurements of internal surfaces.

Kymography. Internal motion in an object (heartbeat, breathing of lungs) can be recorded by interposing between object and film a lead diaphragm with a 0.04-in. (1-mm) slit. The film is moved at a speed which is compatible with the motion, and records a sawtooth image as a time record of such motion.

Flash radiography. With an intense x-ray beam from field-emission pulse discharge, it is possible to make flash exposures in 2×10^{-7} s so that a rapidly moving object such as a bullet or shell appears to stand still. This has been of great significance in studies of ballistics, moving parts in engines, and other objects. A series of high-intensity exposures leads to cineradiography—moving pictures. Exposures of 0.01–0.001 s are conveniently made with rotating-target x-ray tubes, which permit passage of electron currents of 500–1000 milliamperes without damage to the target. Thus, images of heart and lungs or moving machinery may be registered between motions (**Fig. 4**).

X-ray generators (linear accelerators) capable of discharging x-ray bursts have been built. With an intensity of up to 60 amperes, at an energy of up to 20 MeV and a pulse duration of 2×10^{-7} s, such instruments are capable of penetrating very dense objects for the purpose of observing fast-moving shock waves through such specimens. *See* SHOCK WAVE.

Monochromatic radiography. By proper techniques of filtration, or reflection from crystal faces or mirrors, it is possible to produce monoenergetic beams,

Fig. 4. Chest radiograph of foundry worker made with intense beam from rotating-target x-ray tube, showing nodules in lungs which are caused by silicosis, and shadows of skeleton, heart, and stomach.

especially below 50 keV. The radiographs of thin or small objects (especially when magnified) are greatly superior to those from beams with a whole spectrum of rays each with a different absorption in the specimen.

Electronic radiography or radioscopy. The basis for this technique is the application of direct image converter tubes or the use of television pickup or electronic scanning. The resultant signals are amplified and presented for viewing on a television monitor. This inspection method permits remote or daylight viewing using x-ray energies up to 15 MeV to inspect steel or other materials. For example, some large rocket motors (about 8 ft or 2.4 m in diameter) are inspected by high-energy x-rays and television detection systems. The electronic method offers technical advantages because of the increased brightness of the radiographic image and because the electronic image signals can be processed to enhance contrast or sharpen image detail.

Digital radiologic imaging. Electronic system images can be digitized for easier processing, storage, and retrieval. Solid-state cameras, such as charge-coupled-device (CCD) cameras, can provide excellent sensitivity and images with extremely wide dynamic range ($10^3 - 10^4$), leading to x-ray image systems with 12-bit (4096 levels of gray) capability or more. *See* CHARGE-COUPLED DEVICES.

Radiographic film images can also be digitized. Large x-ray films (14×17 in. or 35.6×43.2 cm) can be scanned in less than a minute, in a 12-bit image acquisition system to handle films having a density range of 3.5 or more. Digitizing spatial resolution is typically of the order of 35–100 micrometers.

These digital radiologic methods make it possible to store, retrieve, process, and transmit x-ray images produced by film or filmlike methods and by electronic techniques. Optical disk storage of such images makes it much easier to recall and view x-ray images for both medical and industrial inspection. In medicine, a hardware-software-protocol approach known as the Picture Archiving and Communication System (PACS) has been established. PACS units collect image information from a wide variety of diagnostic techniques, including x-radiography and radioscopy, computed tomography, ultrasonic imaging, and magnetic resonance imaging (MRI), making possible improved medical diagnosis by bringing together many test records. In industry, nondestructive testing workstations are assembled. These again combine many test results, but through a neutral image format, so that new methods can easily be entered into the system. *See* COMPUTER STORAGE TECHNOLOGY; MEDICAL IMAGING; OPTICAL RECORDING.

Teleradiology permits the transmission of x-ray image information over telephone lines, via satellite, or by hardwire within an organization such as a hospital. This has been used mostly in medicine to aid in diagnosis in remote communities or to obtain assistance from specialists.

One-step radiography. As in one-step photography, this technique includes a one-step developing process through which a nearly dry, paper-supported radiographic image is obtained within about 2 min after exposure, without the use of a conventional darkroom or wet processing. This process is best suited to low-energy x-radiography but has also been used for neutron and proton radiography. The short density range of the paper is most applicable to x-radiography of high-contrast specimens, such as metal-plastic assemblies.

Paper radiography. Radiographic paper is an alternative to the more expensive x-ray film. Paper costs are typically about one-third those of film. Like one-step prints, the radiographic paper image is viewed in reflection rather than in transmission. A practical range of reflection densities is 0.4–1.0. Paper images are wet-processed in special solutions. These images are not as permanent as those made on film. After about 3 months' storage, the paper images tend to darken, but extended life can be obtained if the paper is chemically fixed like a film and dried. Paper radiographic images for aluminum in the x-ray energy range of 45–160 kV typically show penetrameter sensitivities of 2% or better.

Photothermographic radiography. This technique makes use of silver halide emulsions which can be processed by dry heat. The x-ray image is recorded with light-emitting rare-earth phosphor screens in contact with the film. The dry thermal developing process produces a silver image and stabilizes the translucent film. The dry film processing takes only 10 s and can easily be done in a small portable dark enclosure wherever there is electric power for the processor. These films can be viewed as either a reflection or transmission image. The light-sensitive emulsions display good spatial resolution, to 200 lines per inch (8 per millimeter), and hold the image in storage for periods up to 7 years.

Photostimulable phosphor radiography. Screens made of phosphors such as barium fluorohalide can store x-ray image information for extended periods before the energy is released by exposure to red light. The incident x-ray modify the europium activation ions in the phosphor crystals so that extra electrons are absorbed by the conduction band in a stable configuration. When the exposed phosphor is stimulated by red light, typically a scanning laser beam, the stored energy is released in the form of light in the ultraviolet-blue part of the spectrum. This light is detected, for example by a photomultiplier tube, with the resulting electronic signal converted to a digital image. The phosphor screen images show a very wide dynamic range, as large as 10^5, and a spatial resolution as good as 125 line pairs per inch (5 per millimeter). Images can be displayed on a television monitor or printed on a transparent base to produce a film. Advantages of this technique include excellent sensitivity, the quantitative results inherent in the digital nature of the signal, and the possibility of adjusting the final image processing to compensate for x-ray exposure variations.

Microradiography. Finer definition for thin specimens can be obtained by using microradiography.

Objects can be radiographed in close contact with fine-grain film, and the film then enlarged. Detail as small as 40 microinches (1 μm) can be brought out by this photographic-enlargement process.

An improved approach makes use of a microfocus x-ray tube, a tube having an extremely small source of x-radiation (typically of the order of 10 μm). With such a small focal spot, the inspection object can be placed close to the x-ray tube (instead of the detector), and the resultant image can be magnified with the diverging x-ray beam. This method can provide definition some 10 times smaller than the photographic method. Geometric magnification of 10 to 20$\times$ is common in industrial applications, and geometric magnification of 100$\times$ has been demonstrated.

Neutron radiography. Neutron beams, obtained from nuclear reactors, accelerators, or radioactive sources, can penetrate matter with relative ease since they are not electrically charged. Slow thermal neutrons (those in thermal equilibrium with their surroundings at or near room temperature) cover a broad band of wavelengths similar to the general, or "white," radiation from an x-ray tube. *See* THERMAL NEUTRONS.

The attenuation of neutrons by most materials is relatively small because the neutron carries no electric charge and consequently is neither attracted nor repelled by the charged particles in the nucleus, nor by the electron clouds associated with the atoms of the material through which the neutron passes. Such attenuation as does occur arises from the occasional capture of a neutron by a nucleus to form another nucleus of different mass number, or from scattering by the nucleus. Whereas in the case of x-radiation the mass absorption coefficient increases in a regular fashion with increasing atomic number of the absorber, the mass absorption coefficients (cross sections) for neutrons are randomly distributed. There are low attenuations of thermal neutrons in most of the heavy elements; therefore, large thicknesses of heavy materials can be inspected by neutrons in less time than would be required by x-radiography or gamma radiography. On the other hand, the neutron absorption coefficients of some elements with low atomic numbers are high; hydrogen, lithium, and boron are particularly attenuating. This reversal of attenuation properties between neutrons and x-rays leads to complementary properties for the two radiographic methods. With neutrons, it is possible to visualize materials such as liquids, adhesives, rubber, plastic, or explosives even when they are in metal assemblies.

Neutron radiography offers three major advantages. First is the described sensitivity to light material. Second, neutrons are sensitive to particular isotopes rather than elements. Therefore, a neutron radiograph can show differences between isotopes of the same element, for example, between normal uranium (essentially ^{238}U) and ^{235}U, the isotope of uranium that fissions more easily. A third advantage is that highly radioactive material can be radiographed without problems of fogging the radiographic film. In one method for detecting neutron images, the transfer technique, the image is detected by a thin sheet of potentially radioactive material, such as indium or dysprosium. The neutrons are transmitted through the object to the detector. The radioactive image on the detector is made visible by later placing it on film, remote from the neutron beam and object, and allowing the radioactive decay radiation to expose the film. This transfer method is extensively used to inspect highly radioactive nuclear reactor fuel.

Other methods for detecting neutron images are similar to those used for x-radiography. X-ray film, one-step film, and television methods are all used, but the conversion screens are different. For film the common screen is thin gadolinium metal; this emits internal conversion electrons as a result of gamma radiation promptly excited by neutron capture. Screens that emit light are also used for both film and television systems. These employ materials such as gadolinium, lithium, or boron. Gadolinium emits electrons, as noted above; lithium and boron emit alpha particles. In each case the radiation excites the emission of light from a phosphor material.

Neutron radiography is used to inspect nuclear reactor fuel both before and after irradiation in the reactor, explosive and adhesive-bonded components, as well as a wide variety of other materials and assemblies. The inspection of small, explosive devices, to confirm the presence of explosive, and investment-cast turbine blades to detect the residual ceramic core have been major areas of application (**Fig. 5**). (Ceramic remaining in critical turbine blades can
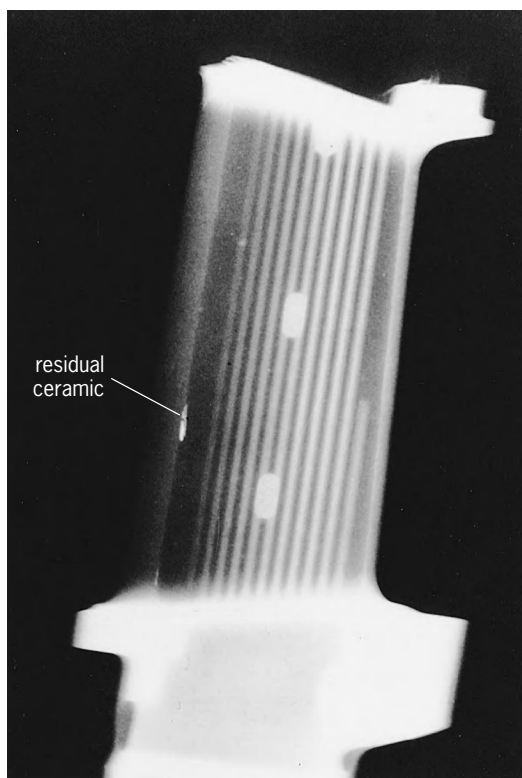

residual
ceramic

Fig. 5. Thermal neutron radiograph of an investment-cast turbine blade, showing image of residual ceramic. (*Industrial Quality, Inc.*)

lead to overheating and blade failure in hot areas of an engine because of the heat-blocking property of the residual ceramic. This residual ceramic can be removed by additional leaching.)

Proton radiography. This method employs beams of protons. A rapidly moving, high-energy proton or other charged particle moves through material with little attenuation until it slows sufficiently for the charges on the particle and the material to interact. A monoenergetic charged particle travels a well-defined distance, called the range, in a given material before it is stopped. Since most of the attenuation of the charged particles occurs near the end of the range, a very small change in material thickness will result in a large change in radiation transmission. Therefore, the sensitivity of this method to small changes in object thickness is very great, if the total path for the radiation approximates the range. This is a major advantage of proton radiography. Changes in object thickness as small as 0.05% have been imaged with one-step film.

Gerold H. Tenney; George L. Clark; Harold Berger

Bibliography. H. Berger (ed.), *Practical Applications of Neutron Radiography and Gaging*, American Society for Testing and Materials, ASTM STP 586, 1976; L. Bryant and P. McIntire (eds.), *Radiography and Radiation Testing*, vol. 3: *Nondestructive Testing Handbook*, 2d ed., 1985; D. C. De Vos, *Basic Principles of Radiographic Exposure*, 2d ed., 1995; S. J. Dwyer III and R. G. Jost (eds.), *Medical Imaging IV: PACS Systems Design and Evaluation*, SPIE vol. 1234, International Society for Optical Engineering, 1990; R. Halmshaw, *Industrial Radiology: Theory and Practice*, 2d ed., 1995; H. V. Lemke et al. (eds.), *Proceedings of International Symposium on Computer Assisted Radiology*, biannually; V. J. Orphan (ed.), *X-Ray Detector Physics and Applications*, SPIE vol. 2009, International Society for Optical Engineering, 1993; U.S. Nuclear Regulatory Commission, *Review and Evaluation of Technology, Equipment, Codes and Standards for Digitization of Industrial Radiographic Film*, 1992.

Radioimmunoassay

A general method employing the reaction of antigen with specific antibody, permitting measurement of the concentration of virtually any substance of biologic interest, often with unparalleled sensitivity. The basis of the method is summarized in the competing reactions shown in **Fig. 1**. The unknown concentration of the antigenic substance in a sample is obtained by comparing its inhibitory effect on the binding of radioactively labeled antigen to a limited amount of specific antibody with the inhibitory effect of known standards.

Method. A typical radioimmunoassay is performed by the simultaneous preparation of a series of standard and unknown mixtures in test tubes, each containing identical concentrations of labeled antigen and specific antibody, as well as variable amounts of standards or the unknown sample. It is generally advisable to have 10 to 20 standards whose concentrations cover the range to be assayed. With the use of automatic pipetting devices, the number of unknown samples that can be handled conveniently in a single assay may be 1000 or more. After an appropriate reaction time, which may be hours or days depending on the association constant for reaction with the particular antiserum, the antibody-bound (B) and free (F) fractions of the labeled antigen are separated by one of a variety of techniques, such as adsorption of free antigen to solid-phase material, precipitation of antigen-antibody complexes, and adsorption or complexing of antibody to solid-phase material.

Under the usual conditions for an assay, the antigen-antibody complexes do not spontaneously precipitate, even if they would do so at much higher concentrations of reactants. The B/F ratios in the standards are plotted as a function of the concentration of unlabeled antigen (standard curve), and the unknown concentration of antigen is determined by comparing the observed B/F ratio with the standard curve (**Fig. 2**).

Substances measured. The specific antibody is produced in a suitable animal, usually a guinea pig or a rabbit, by multiple immunizations with the substance to be measured or another substance with similar immunologic properties. Since a peptide hormone as small as vasopressin (∼1000 mol wt) is antigenic in at least some animal species, and since virtually any substance of biologic interest can be rendered antigenic by coupling it chemically to a large polypeptide or protein, the method has been applied to the measurement of thousands of hormonal and nonhormonal substances.

Competitive radioassay. The radioimmunoassay principle is not limited to immune systems. It has been extended to systems in which, in place of the specific antibody, there is a specific reactor (that is, binding substance), which may be, for instance, a specific binding protein in plasma, an enzyme, or a tissue receptor site. The more general term "competitive radioassay" has been used to describe such applications. However, this term is not completely inclusive, since the principle is applicable to the use
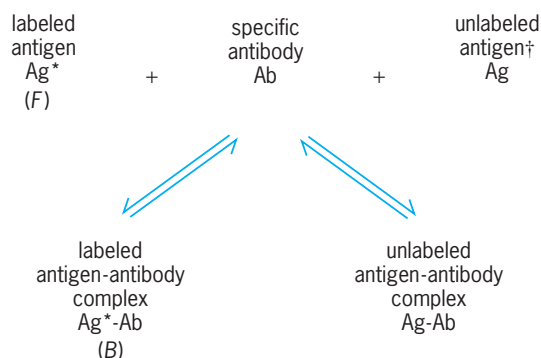

Fig. 1. Competing reactions that form the basis of radioimmunoassay. Dagger indicates antigen in known standard solutions or unknown samples. B indicates bound antigen; F, free antigen.
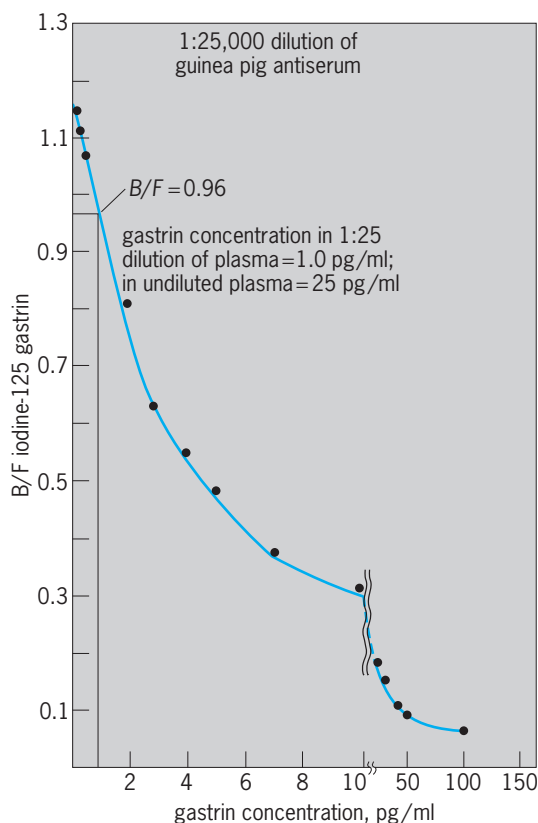
Fig. 2. Standard curve for radioimmunoassay of gastrin. Hormone concentration in unknown plasma sample is determined as shown.

of markers other than radioisotopic tracers (modified phages, enzymes, and so on).

Sensitivity considerations. Radioimmunoassay has found wide application because of its potential for high sensitivity and specificity, and because the methodology is easily adaptable to automation and computerization. Each of the diverse applications does not require all of these attributes. The requirements for sensitivity depend on the minimal quantity that must be measured. For peptide hormones in the nonstimulated state, the concentrations may be as low as 10^{-12} to 10^{-10} M: the concentrations of thyroidal and steroidal hormones are higher by a factor of about 10^5; the concentrations of drugs may be even higher. Theoretical analysis of the radioimmunoassay principle demonstrates that the inherent sensitivity of the system is limited by the equilibrium constant that characterizes the reaction of antigen with the predominating antibodies in the antiserum. Since animals, even of the same species, vary considerably in the concentration and potential sensitivity of the antibodies they produce, the probability of obtaining a satisfactory antiserum increases with the number of animals immunized. Since the presence of other immunological reactions does not interfere with the reaction between labeled antigen and its specific antibody, immunization with several unrelated antigens can be performed simultaneously, affording an advantage in reducing the number of animals to be immunized and bled by a factor equal to

the number of antigens employed. Monoclonal antibodies, which can be produced in quantity and remain absolutely uniform, offer a great improvement in radioimmunoassay technology.

To achieve the maximal sensitivity obtainable with a particular antiserum, the chemical amount of antigen in the radioactively labeled antigen should be low enough that further reduction in its size does not increase the B/F ratio in the absence of added unlabeled antigen. This requirement determines the choice of radioisotope used for labeling. Because of the low concentrations of peptide hormones in plasma, high-specific-activity tracers are required. Iodine-125 (half-life = 60 days) is the radiochemical of choice and can readily be substituted on a tyrosyl or histidyl residue of the antigen using a variety of chemical procedures. For assays of nonpeptidal hormones and drugs, ^{3}H-labeled (half-life = 12 years) antigens prepared and purified in commercial laboratories are generally employed. Long-lived isotopes such as ^{14}C (half-life = 5800 years) are of limited value except when the concentrations of antigen to be measured are quite high, that is, concentrations of 10^{-6} M or higher. *See* RADIOISOTOPE (BIOLOGY).

Nonspecific interference. Radioimmunoassay differs from traditional bioassay in that it is an immunochemical procedure which is not affected by biologic variability of the test system. However, nonspecific factors can interfere with chemical reactions, and the possible presence of cross-reacting related antigens in the samples to be assayed must be considered. Some nonspecific factors that may interfere in the immune reaction include extremes of alkalinity or acidity or high concentrations of salt, anticoagulants, enzyme inhibitors, and so on. It is therefore important that the contents of the incubation tubes be identical for the standards and unknowns except for the differing amounts of antigen. Differences between concentrations of peptide hormones as determined by bioassay and immunoassay may be caused by the presence of precursor hormones (prohormones), molecular fragments, or related but not identical hormones. Requirements for specificity merits consideration in assays for other substances as well. For instance, if the clinical problem relates to the toxicity of a particular drug, the question of whether the assay measures the biologically active form of the drug is relevant. If the question relates simply to whether or not a drug of abuse (morphine, LSD, and so forth) has been used, the presence of metabolites or the exact form of the drug may be irrelevant. *See* BIOASSAY.

Applications. The radioimmunoassay principle has found wide application in the measurement of a large and diverse group of substances in a variety of problems of clinical and biological interest. It is therefore not unexpected that there are differences in the specific methods employed for the assay of a particular substance. It seems that virtually any substance of biologic interest can be measured using radioimmunoassay, modified according to the characteristics of the particular substance. In recent years there has been a tendency for laboratories to try as far as

possible to replace radioimmunoassays with non-isotopic techniques. Radioisotopes are highly regulated substances and are increasingly expensive and difficult to obtain; disposal is cumbersome and represents a potential hazard to the environment. The radioisotope is replaced by a fluorescent marker, for example, in the fluorescence polarization immunoassay that many large clinical laboratories prefer. Enzyme-linked immunosorbent assays (ELISA) have also replaced many radioimmunoassays. Nevertheless, radioimmunoassay remains one of the few techniques available to measure amounts of small (monovalent) molecules in the usual research laboratory. *See* ANTIBODY; ANTIGEN; IMMUNOASSAY; CLINICAL IMMUNOLOGY.   Rosalyn S. Yalow; J. John Cohen

Bibliography. S. A. Berson and R. S. Yalow (eds.), *Methods in Investigative and Diagnostic Endocrinology*, vol. 2: *Peptide Hormones*, 1974; D. S. Hage, Immunoassays, *Anal. Chem.*, 71:294R–304R, 1999; F. Laszlo and T. Janaky (eds.), *Clinical Application of Radioimmunoassay*, 1992; R. Yalow, Nobel Prize Lecture, Dec. 8, 1977: *Radioimmunoassay: A Probe for Fine Structure of Biological Systems*, in T. Charc, *An Introduction to Radioimmunoassay and Related Techniques*, 4th ed., 1991.

Radioisotope

A radioactive isotope (as distinguished from a stable isotope) of an element. Atomic nuclei are of two types, unstable and stable. Those in the former category are said to be radioactive and eventually are transformed, by radioactive decay, into the latter. One of the three types of particles or radiation (alpha particles, beta particles, and gamma rays) is emitted during each stage of the decay. *See* ISOTOPE; RADIOACTIVITY.

The term radioisotope is also loosely used to refer to any radioactive atomic species. Whereas approximately a dozen radioisotopes are found in nature in appreciable amounts, hundreds of different radioisotopes have been artificially produced by bombarding stable nuclei with various atomic projectiles.

The production, separation, purification, and shipment of radioisotopes are usually considered primary production; secondary processing of the primary materials into compounds, special shapes, radiation sources, and medical radiopharmaceuticals is usually required before use in hundreds of applications that have been developed. Radioisotopes are produced by irradiation of the elements with high-energy particles (protons, deuterons) in accelerators and in the high-intensity neutron fluxes of relatively small research-type reactors.

A few radioactive elements occur in nature, for example, uranium, radium, and thorium, which were produced when the Earth was formed. Some of these naturally occurring radioisotopes (primary parents of series) have half-lives (time in which one-half of the atoms decay) greater than 10^8 years, and therefore have not had time to disappear. During the early years of work with radioactive materials, only natural radioactivity was available; The first radioactive tracer experiment was in 1934 with radiolead (^{212}Pb) obtained from thorium decay products.

The first artificially induced radioactivity was also produced in 1934 by irradiating aluminum foil with alpha particles, thus initiating the use of radioactivity on a wide scale in scientific work. A very wide variety of radioisotopes are produced in particle accelerators, such as the cyclotron. Charged particles, such as deuterons (D^+) and protons (H^+), are accelerated to great speeds by high-voltage electrical fields and allowed to strike targets in which nuclear reactions take place; for example, proton in, neutron out (p,n), increasing the target-atom atomic number by one without changing the atomic mass; and deuteron in, proton out (d,p), increasing the atomic mass by one without changing the atomic number. The target elements become radioactive because the nuclei of the atoms are unbalanced, having an excess or deficit of neutrons or protons. Although the particle-accelerating machines are most versatile in producing radioisotopes, the amount of radioactive material that can be produced is relatively smaller than that made in a nuclear reactor [less than curie amounts; a curie (abbreviated Ci) is that quantity of a radioisotope required to supply 3.7×10^{10} disintegrations per second or 3.7×10^{10} becquerels (Bq)]. For large-scale production, nuclear reactors with neutron fluxes of 1×10^{10} to 5×10^{15} neutrons per square centimeter per second are required. *See* NUCLEAR REACTION; NUCLEAR REACTOR; PARTICLE ACCELERATOR; REACTOR PHYSICS; UNITS OF MEASUREMENT.

Production. Radioisotope production proceeds in a series of steps. First, target materials containing the desired element are selected in the most stable heat- and radiation-resistant form (for example, metal oxides) and the highest chemical purity, determined by chemical, spectrographic, or neutron activation analysis. The target material is encapsulated in a low-neutron-absorbing material (such as aluminum or magnesium) and sealed by heli-arc welding or an equivalent process, and the capsule is tested for leakage. Next, the capsule is irradiated for the optimum time period, determined by the half-life of the desired radioisotope, the absorption cross section of the target atom, and the growth of possible undesired impurities. The irradiated capsule is then removed to a remote-control hot cell for chemical processing, and the preparation is analyzed and assayed by the most advanced methods that are amenable to remote-control operations. Aliquots are dispensed as needed for users or secondary processors, and the radioisotope is packaged, inspected for radioactive contamination, and shipped by common carriers. Shipment may be subject to local, state, federal, or international shipping regulations, depending upon the nature, size, and destination of the radioactive material.

Reactor production. Most radioisotopes produced in quantity are made in the nuclear reactor by one of four reactions (**Table 1**): neutron-gamma, neutron-proton, neutron-alpha, and neutron-fission. The neutron-gamma reaction is the most common,

TABLE 1. Reactor radioisotope production reactions	
Reaction	Examples
(1) Neutron-gamma (n, γ)	$^{59}_{27}Co + {}_{0}n^1 \rightarrow {}^{60}_{27}Co + \gamma$
(2) Neutron-proton (n,p)	$^{32}_{16}S + {}_{0}n^1 \rightarrow {}^{32}_{15}P + {}^{1}_{1}p$
(3) Neutron-alpha (n,α)	$^{35}_{17}Cl + {}_{0}n^1 \rightarrow {}^{32}_{15}P + {}^{4}_{2}\gamma$
(4) Neutron-fission (n,f)	$^{235}_{92}U + {}_{0}n^1 \rightarrow$
	$^{131}_{52}Te + {}^{102}_{42}Mo + \sim 2{}_{0}n^1$

because many elements have a good neutron-capture cross section (relative ability for capturing neutrons). By simple neutron capture, important radioisotopes, such as ^{24}Na, ^{59}Fe, ^{60}Co, and ^{198}Au, are made. The procedure is quite simple in this case. Highly purified materials (to prevent contamination by neutron-capturing impurities) in amounts ranging from a few milligrams to several hundred grams are sealed in pure aluminum cans and placed in the reactor. In the higher-flux reactors (flux greater than 10^{14} neutrons per square centimeter per second) the highly purified target material is sealed in quartz ampules which are in turn weld-sealed in small aluminum tubes, well designed for heat removal (**Fig. 1**). Pneumatic or hydraulic tubes are used for passing samples in and out of the reactor. Also, reactors may have special magnesium, beryllium, or aluminum holders for insertion into the reactor lattice. Care must be taken to avoid putting maerials that decompose easily, such as organic compounds, into the reactor because gas pressure may be produced in the container. In general, thermally stable materials are also fairly stable under neutron irradiation. Metals or stable oxides of elements are usually used as target materials.

The production rate for a radioisotope in a reactor depends upon the neutron flux, amount of target atom, the half-life of the radioisotope, and the neutron-activation cross section of the target atom. The formula written below can be used for calculating the amount of radioisotope produced, where

$$A = N\phi\sigma(1 - e^{-\lambda t})$$

lating the amount of radioisotope produced, where A is the activity in disintegrations per second, N is the number of target atoms, ϕ is the neutron flux in neutrons per square centimeter per second, σ is the activation cross section for the reaction in square centimeters per atom, and the expression $(1 - e^{-\lambda t})$ is the saturation factor. The irradiation time t and the decay constant λ are in compatible units ($\lambda = 0.693/$half-life). The activity A, in disintegrations per second, can be converted to millicuries by dividing by 3.70×10^7 disintegrations per second per millicurie.

Fig. 1. Aluminum capsule target for high-neutron-flux reactor.

Target materials of some elements can be obtained quite pure; when such elements have only one isotope that has a high activation cross section, radioisotopes are easy to produce and little chemical purification is required after irradiation. In other cases there is multiple production in the target material, sometimes four or five different radioactive species in one target, which must be chemically separated from the main product. Virtually every known chemical-separation procedure is used in this kind of work, which usually must be done by remote methods because of the high radiation levels involved. *See* NUCLEAR CHEMISTRY.

When radioisotopes are produced by neutron-proton or neutron-alpha reactions or by a neutron-gamma reaction followed by beta decay, the radioelement is of a different chemical species from the target element and can be separated from it chemically to produce carrier-free radioisotopes.

Very often concentrations of radioelements are too low for them to be precipitated directly, so they are carried on the surface of a flocculent precipitate, such as $Fe(OH)_3$; similarly, others are coprecipitated, where isomorphous compounds are brought down together, for example, $^{140}BaSO_4$ with $PbSO_4$. Many methods are more adaptable to work with low concentrations of material or amenable to remote operation, such as solvent extraction and ion exchange. A preferred purification method is gasification, because if the radioisotope goes through various stages in the gaseous phase, very high purity usually results. Distillation is sometimes used, as in purification of ^{131}I (distilled as the element) and $^{103\text{-}106}$Ru (distilled as the tetroxide). Ion exchange is practically the only method for fractionating the rare earths in the low concentrations usually found in radioisotope work.

Fission-product separation. The fission products, fragments of the fissioned uranium or plutonium atom, are an extremely important group of radioisotopes, ranging from zinc (atomic number 30) to samarium (atomic number 62). The fission-product fragments occur in two groups: light atoms with atomic masses between 72 and 110 units and heavy atoms with masses between 110 and 162 units. Those fission products with high yields occur at one of the peaks, and the companion fission fragment occurs with the same high yield on the other peak; for example, mass 140 (Xe through Ba) has companion fragments of about mass 95 (Y through Mo). Fission-product yields are given as the percent of fissions that result in fragments of a certain mass; there are two fragments for each fission, so the overall yield arithmetically totals up to 200%. *See* NUCLEAR FISSION.

The fission products can be separated and purified for use as radioisotopes and fall into two groups (**Table 2**).

With the exception of ^{131}I, the short-lived fission products were first separated almost entirely for research purposes. The extraordinary growth in the use of ^{99}Mo to produce the daughter ^{99m}Tc for medical use has made ^{99}Mo the most important commercial radioisotope. The parent radioisotope, 67-h ^{99}Mo, is also made by neutron-gamma

TABLE 2. Isotopes ordinarily separated from fission products

Radioisotope	Half-life	Fission yield, %
Principal long-lived fission products*		
^{85}Kr	10.27	1.3
^{90}Sr	28	5.8
^{99}Tc	212,000	6.1
^{106}Ru	1	0.4
^{137}Cs	30	5.9
^{144}Ce	0.8	5.7
^{147}Pm	2.6	2.4
Principal short-lived fission products†		
^{89}Sr	54	4.8
^{91}Y	59.5	5.8
^{95}Zr-^{95}Nb	65–35	6.3
^{99}Mo	2.8	6.2
^{103}Ru	39.8	3.0
^{131}I	8.08	2.9
^{133}Xe	5.3	6.5
^{140}Ba	12.8	6.4
^{141}Ce	32.5	6.0
^{143}Pr	13.9	6.0
^{147}Nd	11.1	2.4

*Half-life in years.
†Half-life in days.

the processing of spent fuel from power reactors. Krypton-85 is also separated from the waste gases at the fuel reprocessing plants. The long-lived fission products have potential as sources of radiation and heat.

Accelerator production. The production of radioisotopes using high-energy particles is more versatile than reactor production, especially for radioisotopes that are neutron-deficient (most reactor radioisotopes have an excess of neutrons in the nucleus). Many of the neutron-deficient radioisotopes decay by *K*-shell electron capture with the emission of soft x-rays, or by positron emission with accompanying annihilation radiation (the collision of positive and negative electrons, e^- and e^+, with each electron mass converted to 0.51-MeV gamma radiation). Particularly in nuclear medicine, short-lived radioisotopes are needed that have little beta radiation and thus cause minimum tissue damage, or radioisotopes with directional annihilation radiation to aid location in scanning.

Small cyclotrons have been developed that will accelerate protons, deuterons, and alpha particles. A few of these are now devoted to radioisotope production.

Attachments have been put onto a few large research cyclotrons to get protons of extremely high energy to produce radioisotopes by spallation. This occurs when the particle hitting the target atom is so energetic that "pieces fly off." Many of the particles making up the nucleus are lost and a variety of radioisotopes of lower atomic number and weight are produced. *See* SPALLATION REACTION.

Shipping. The packaging and shipment of radioisotopes present some unusual problems—the penetrating radiation must be shielded and leakage must be prevented because of the extreme toxicity of many radioisotopes. Packaging must be done by remote control. Most radioisotopes are dispensed as water solutions by pipetting from glass storage bottles to glass shipping bottles. Tempered-glass bottles are commonly used, with closures lined with polyethylene inserts. For some special preparations, such as carrier-free ^{45}Ca or ^{32}P, polyethylene storage and shipping bottles are used to cut down losses by

reaction on natural molybdenum, or, in some cases, enriched ^{98}Mo target material is irradiated. The very high specific activity fission product ^{99}Mo does have some advantages for making ^{99}Tc generators. However, because of the high-radiation-level, highly technical separation process, very few commercial radioisotope producers have been interested in taking it over.

The short-lived fission products are made by irradiating a small piece of ^{235}U-Al alloy in the reactor for a few weeks. The target is promptly dissolved in nitric acid upon discharge from the reactor, the radioiodine and xenon separated from the gases are discharged during dissolution, and the uranium-aluminum nitric acid solution is processed by solvent extraction and ion exchange, principally to obtain ^{99}Mo, ^{95}Zr-^{95}Nb, ^{89}Sr, and the rare earths.

The long-lived fission products, such as ^{90}Sr and ^{137}Cs, are separated from long-cooled waste from

Fig. 2. Radioisotope shipping containers. (*a*) Lead shield, stainless-steel encased for gamma sources, such as ^{60}Co. (*b*) Lead shield, stainless-steel encased, in wooden block for pile units and small sources. (*c*) Gas shipping container with internal sealed cylinder for ^{85}Kr and tritium gas. (*d*) Nonreturnable lead shield for processed radioisotope bottles. (*e*) Solid stainless-steel container for bottles of processed ^{90}Sr. (*f*) Lead shield, stainless-steel encased, on a shipping pallet for distributing weight in aircraft, for pile units and small gamma sources. (*Oak Ridge National Laboratory*)

Fig. 3. Nonreturnable shipping container for radioisotopes. (*Oak Ridge National Laboratory*)

adsorption on the walls of the container. High standards of cleanliness must be maintained, although no attempt is made to ensure sterility of solutions except for the special medical radioisotope preparations shipped by pharmaceutical suppliers.

A few preparations have sufficient mass per unit of radioactivity and weak enough radioactivity to permit handling and packaging as dry solids. An example is ^{14}C, which is dispensed as barium carbonate. Materials that are not chemically processed, such as reactor irradiation units or metallic cobalt, are sent out as solids in metal containers.

Two general kinds of shipping container are used; returnable, on which a deposit is paid; and nonreturnable, which the receiver may keep. Returnable containers are usually more massive and expensive, especially those used for shipping large amounts of ^{60}Co or fission products. Pure beta emitters, such as ^{32}P, require little shielding other than the absorbent packing material. Gases are shipped in special glass or metal containers (**Figs. 2** and **3**).

Shipment of radioactive materials by rail or truck is covered by federal government regulations (Department of Transportation) and international regulations, coordinated through the International Atomic Energy Agency. *See* ISOTOPE SEPARATION; RADIOACTIVE TRACER.   Arthur F. Rupp

Bibliography. *Accelerator-Produced Nuclides*, Brookhaven National Laboratory, BNL-50448, 3 vols., 1975, 1978, 1983; F. Helus (ed.), *Radionu-clides Production*, 2 vols., 1983; International Atomic Energy Agency, *Radioisotope Production and Quality Control*, Tech. Rep. Ser. 128, 1971; International Atomic Energy Agency, *Radiopharmaceuticals and Labelled Compounds*, vol. 1, IAEA-SM-171/35, 1973; F. F. Knapp and T. A. Butler (eds.), *Radionuclide Generators: New Systems for Nuclear Medicine Applications*, 1984; U.S. Department of Energy, *Isotope Production and Distribution*, 1994.

Radioisotope (biology)

A radioactive isotope used in studying living systems. This article discusses the use of radioactive isotopes in the investigation of metabolic processes.

The usefulness of radioisotopes as tracers arises chiefly from three properties: (1) At the molecular level the physical and chemical behavior of a radioisotope is practically identical with that of the stable isotopes of the same element. (2) Radioisotopes are detectable in extremely minute concentrations. (3) Analysis for radioisotopic content often can be achieved without alteration of the sample or system. In some applications radioisotopes are not essential but are used because of the advantages of sensitivity or convenience. In other applications, however, principally those in which reaction rates and transfer rates are studied, isotopes, particularly radioisotopes, have unique advantages as tracers, and their use in studies of the kinetics of steady-state systems, self-diffusion, and metabolic pathways has opened numerous fields previously thought to be inaccessible. *See* ISOTOPE; RADIOISOTOPE.

Methods. Radioisotope methods include preparing the labeled material, introducing it into the system to be studied, detecting its presence, and interpreting the results.

Preparation of labeled compounds is one of the most important and most difficult steps in tracer methodology. In synthesis with radioactive reagents, the chemical procedures used must often be modified for remote handling, and other special precautions must be taken to prevent contamination of laboratories and personnel with radioactive material and to minimize irradiation of the personnel. Containment, shielding, and distance are the main items in protecting personnel from radiation. Some compounds labeled by biosynthesis, derived from bacteria or plants grown on media containing radioactive substrates, are commercially available. Before a labeled product is used, it should be tested for radiochemical purity, which means that ordinarily there should be only one radioactive species present and that all of it should be in the same chemical state. Labeling of an element present in two valence states has caused confusion. The chemical stability of the label (that is, the radioactive element) should also be established. It is unwise to use material labeled in easily exchangeable positions if the fate of the material itself (rather than the exchange process) is the

point of interest. *See* RADIOCHEMICAL LABORATORY; VALENCE.

The amount of isotope to be used and the path by which it is introduced into the system are governed by many factors. Sufficient tracer to be detectable must be used, but the amounts of material which are introduced must be small enough not to disturb the system by their mass, pharmacological effects, or the effects of radiation. The mass of 1 curie, the unit of disintegration rate, depends inversely upon the half-life and directly upon the atomic weight of the particular radioisotope; it is 1 gram for ^{226}Ra (half-life 1620 years), but only 8 micrograms for ^{131}I (half-life 8.0 days). In tracer experiments with small animals, microcurie quantities are usually adequate. It is sometimes necessary to choose between administering the isotope in a single injection or using a continuous perfusion, for example, to maintain a constant concentration in the arterial blood. Usually the single injection is easier to achieve, and the resulting data are easier to analyze.

There are many methods for detecting the presence of radioactive material. The Geiger counter has largely been displaced by thallium-activated sodium iodide scintillation crystals for counting gamma rays, but Geiger counters and proportional counters are still useful for counting alpha and beta particles. A well-type scintillation counter may have more than 25% efficiency for ^{131}I. In a scintillation crystal the passage of a gamma ray through the crystal produces a flash of light which is detected by a photomultiplier tube connected to the crystal. The signal from the photomultiplier tube may be fed into a scaler or rate meter, as with the signal from Geiger tubes, or after amplification it may be used as the input signal for a pulse-height analyzer in a method called gamma-ray spectroscopy. The pulse-height analyzer makes it possible to record pulses of a single height, corresponding to single gamma-ray energies, so that one isotope may be counted selectively in the presence of others. With multichannel pulse-height analyzers, simultaneous determinations may be made of a number of gamma-emitting isotopes in a mixture. Beta-particle spectroscopy is also possible but much more difficult both to achieve and to interpret, in part because beta particles are emitted with a spectrum of energies, whereas gamma rays are typically monoenergetic. For very low-energy β^- particles (negative electrons), particularly those emitted by ^{14}C and tritium (^{3}H), liquid scintillation counting is useful. In this method a solution of the sample and a phosphor take the place of the scintillation crystal. Low-energy β^- emitters can also be counted by introduction, in the gas phase, directly into proportional counters. In histological and cytological studies the method of autoradiography, in which photographic film is exposed through contact with the specimen, is very useful. The autoradiographic method is also used extensively in conjunction with paper or column chromatography, particularly in studies of metabolic pathways. *See* AUTORADIOGRAPHY; CHROMATOGRAPHY; GEIGER-MÜLLER COUNTER; PARTICLE DETECTOR; PHOTOMULTIPLIER; SPECTROSCOPY.

In many applications the interpretation of the data obtained with radioisotopes is quite simple. The volume of dilution and total exchangeable mass, for example, are inversely related to the final concentration. Following a metabolic pathway or determining the uptake of radioisotopes by various organs or tissues is analogous to following a flock of sheep by belling the leader. In studies of rate processes, however, particularly when material is moving in opposing directions simultaneously, extensive use of mathematics is needed for correct interpretation of the data. The mathematical basis for such interpretation is outlined in the following paragraphs.

Interpretation of tracer kinetics. For mathematical analysis it is convenient to designate the distinct phases or volumes of a system as compartments. In steady-state systems it is assumed that, except for the tracer, the content of a compartment does not change; the rates of flow of material into and out of the compartment are equal. It is customary, but not always realistic, to assume that within a given compartment mixing of the labeled and unlabeled material is instantaneous and complete; the specific activity (the ratio of the amount of the labeled form of a substance to the total substance present) within a compartment is assumed to be uniform in space, although it may be a function of time.

Given a compartmental model regarded as representing the system of interest (often greatly simplified), the expected behavior of the tracer is described with a set of differential equations in terms of the invariants of the system. The solutions of these equations contain constants which would be determined experimentally from observations on an ideal system. Although prediction of tracer behavior from information about the system has many applications, the reverse problem, that of using the experimentally determinable constants describing the behavior of the tracer to deduce the invariants characteristic of the system, is the usual goal of the tracer experiment.

The passage of a tracer through a chain of steady-state compartments with one-way flow, as in the system $\rightarrow$ A $\rightarrow$ B $\rightarrow$ C $\rightarrow$ D $\rightarrow$, is described with equations identical in form to the Bateman equations, which describe radioisotopic decay chains. The solution equation for the nth compartment will include n exponential terms, with each exponential constant being equal to one of the n turnover-rate constants (the ratio of the rate of entrance or exit of the substance of interest to the amount present in a given compartment). For example, the disappearance of a tracer from the simplest model, $\rightarrow$ A $\rightarrow$, assuming that, after time zero, no tracer enters, is described by Eq. (1), where x is the specific activity in the com-

$$x = x_0 e^{-\lambda t} \tag{1}$$

partment; x_0 is the value at time zero; t is time; e is the base of the natural logarithms, 2.718 . . . ; and λ is the fractional disappearance constant. If, instead, there is no tracer in the compartment at time zero but there is a constant specific activity x_∞ in the inflow from time zero on, the buildup of activity in the

compartment is described by Eq. (2).

$$x = x_\infty \left(1 - e^{-\lambda t}\right) \qquad (2)$$

In both cases λ is equal to the turnover-rate constant k. Because all the λ's appear in the equation for the last compartment of a chain, all the turnover-rate constants may, in principle, be determined by sampling the last compartment only, but this is insufficient for determining the sequence in the chain. *See* NUCLEAR REACTION; RADIOACTIVITY.

When there is flow in both directions, the simple relationship between the k's and λ's is lost. Equations (3) and (4) are differential equations for the

$$\frac{dx_A}{dt} = -k_A(x_A - x_B) \qquad (3)$$

$$\frac{dx_B}{dt} = k_B(x_A - x_B) \qquad (4)$$

two-compartment, steady-state, closed system A $\rightleftharpoons$ B, for which, if at time zero $x_A = x_{A_0}$ and $x_B = 0$, the solutions are expressed by Eqs. (5) and (6), where

$$\frac{x_A}{x_{A_0}} = C_E + C_1 e^{-\lambda t} \qquad (5)$$

$$\frac{x_B}{x_{A_0}} = C_E - C_E e^{-\lambda t} \qquad (6)$$

the C's are constants determined by the amounts of material present. C_E is the equilibrium value, and C_1 is the difference between the initial value and the equilibrium value of the material present. In this example $\lambda = k_A + k_B$ rather than being equal to one or the other turnover-rate constant. Calculation of the individual turnover-rate constants from the experimental data as represented by Eqs. (5) and (6) can, however, be achieved through use of Eqs. (7).

$$k_{AB} = C_1 \lambda \qquad k_{BA} = C_E \lambda \qquad (7)$$

In the general case of N compartments in the steady state, Eq. (8) describes the behavior of the

$$\frac{dx_J}{dt} = -K_J x_J + \sum_{R=A}^{R=N} k_{JR} x_R \qquad R \neq J \qquad (8)$$

tracer in the Jth compartment. Here k_{JR} is the fractional turnover-rate constant resulting from exchange with the Rth compartment, and K_J is the total turnover-rate constant for compartment J. The solutions have the general form of Eq. (9), with $N\lambda$'s

$$x_J = C_{J1} e^{-\lambda_1 t} + C_{J2} e^{-\lambda_2 t}, \ldots, C_{JN} e^{-\lambda_N t} + C_E \qquad (9)$$

for open systems and $(N-1)\lambda$'s for closed systems. Interpretation of experimental data in terms of the size of and rates of transfer among the various compartments involves fitting Eq. (9) to the data and using the constants (the C's and λ's) to calculate the k's and other characteristics of the system. Complete sets of formulas for such calculations for the three-compartment systems are available. Because of the unwieldy equations encountered, explicit formulas for deriving the k's from the C's and λ's have
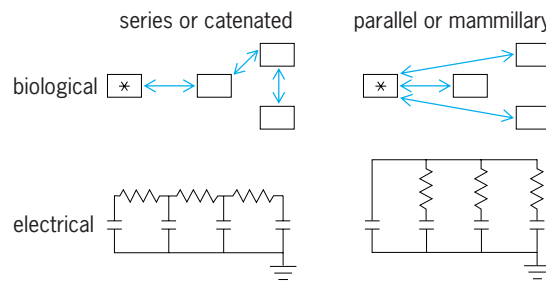
Fig. 1. Simple electrical analog scheme for a four-compartment steady-state system. The asterisk indicates the injected compartment in the biological system, corresponding to an initial charge on the left-hand condenser in the electrical system.

not been published for systems involving more than three compartments, but numerical examples can be worked with matrix algebra using the general relationship $|k| = |C\| - \lambda \|C|^{-1}$, where $|k|$ is a matrix of the K's and k's in Eq. (8), $|C|$ is a matrix of the coefficients in Eq. (9), $|-\lambda|$ is a diagonal matrix of the $-\lambda$'s, and $|C|^{-1}$ is the inverse of $|C|$. This relationship is derived by equating the derivatives of Eq. (9) to those of Eq. (8).

Complicated systems can also be analyzed with electronic analog and digital computers, which eliminate most of the intermediate mathematics involved in translating descriptions of the data into descriptions of the systems. Analog computers have the advantage of directness and simplicity of programming, whereas digital computers have the advantage of numerical accuracy.

Figure 1 depicts an electrical analog for a four-compartment steady-state system which uses a direct analog between the electrical and biological components. The equations which describe the voltage changes in the electrical system are identical in form with those which describe the net tracer movement in the biological system. Adjustment of the electrical parameters therefore provides a direct method for determination of the values of the biological system components. For non-steady-state systems, operational amplifiers and an electrical circuit analogous to the differential equation of the biological model may be used to meet the requirement of unequal flow rates in the two opposing directions. *See* ANALOG COMPUTER; DIGITAL COMPUTER.

Applications. In chemistry the tracer method has been applied to isotopic exchange reactions, chemical kinetics, structural chemistry, self-diffusion studies, and analytical chemistry. Self-diffusion (really isotopic interdiffusion) has been studied in solids, as well as in liquids and gases. In activation analysis a sample is subjected to nuclear bombardment in a reactor or accelerator, and the induced activity provides an extremely sensitive method for assay of the amounts of certain trace constituents.

In biology one of the outstanding achievements in which radioisotopes have played a role has been the use of carbon-14 in the elucidation of the metabolic path of carbon in photosynthesis. The products produced in the first few seconds following exposure
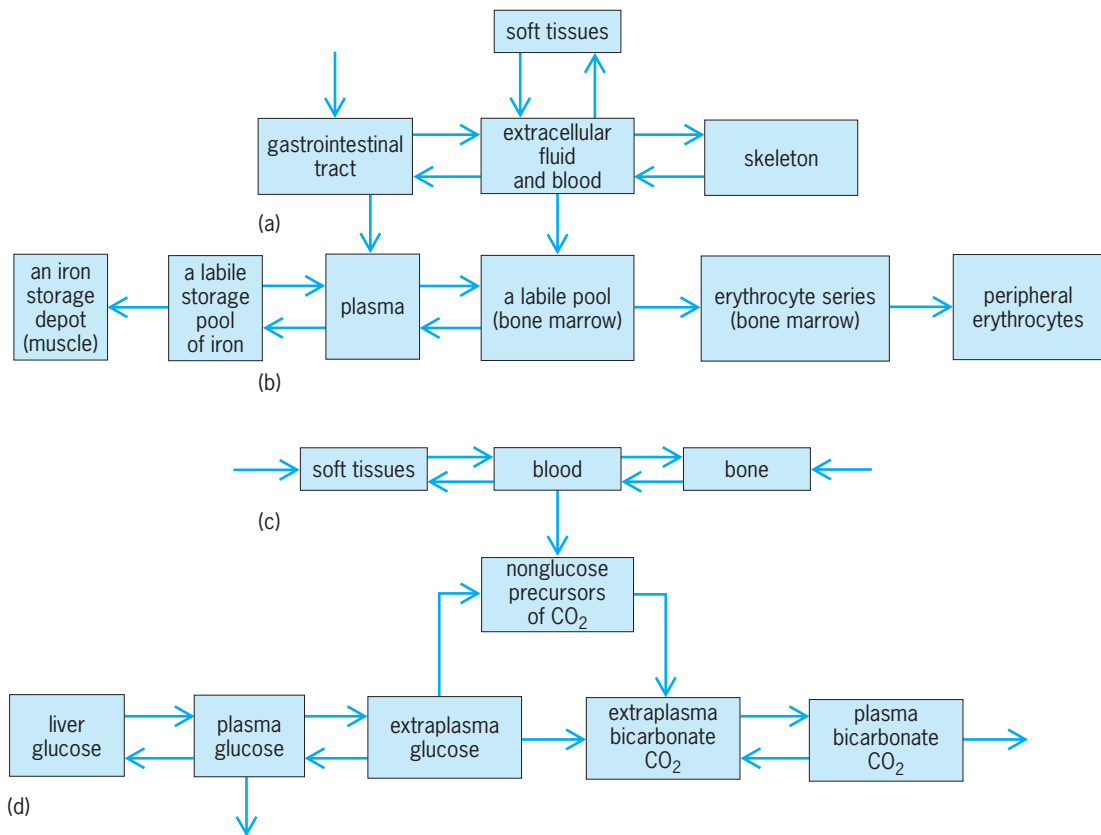
Fig. 2. Models used in analyses of metabolizing systems. (*a*) Calcium. (*b*) Iron. (*c*) Carbonate. (*d*) Glucose. In *b*, the iron returns from the peripheral erythrocytes to the plasma very slowly.
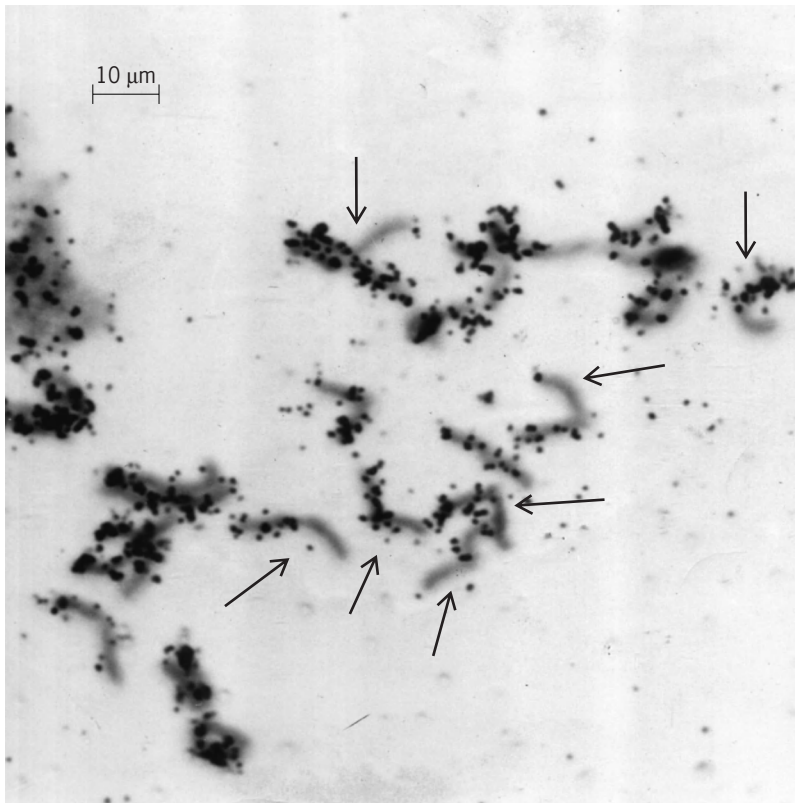

Fig. 3. Photomicrograph of an autoradiograph of *Vicia faba* chromosomes in metaphase labeled with tritiated thymidine. In the chromosome pairs indicated by the arrows, one chromosome is labeled but the other is not. This has been interpreted as evidence that the newly formed deoxyribonucleic acid (DNA) does not exchange with the previously existing DNA. Feulgen stain was used. (*Courtesy of W. L. Hughes*)

to light have been identified by combinations of paper chromatography and autoradiography. Somewhat similarly, the extrathyroidal metabolism of iodine, the path of iodine in the thyroid gland, and other problems of intermediary metabolism have been studied intensively. The concept of the dynamic state of cell constituents is largely attributable to discoveries made with isotopic tracers. At one time it was thought that concentration gradients across cell membranes depended upon their being impermeable, but the use of isotopes has refuted this hypothesis by proving that in many such cases the substances involved are normally transported in both directions across the membrane.

In physiology, radioisotopes have been used in a wide variety of permeability, absorption, and distribution studies. Representative model systems are depicted in **Fig. 2**. As an example, calcium metabolism is described here in some detail. Calcium is absorbed from the intestine into the blood, from which it is distributed to the soft tissues and skeleton or is excreted via the feces, urine, and perspiration. More than 80% of ^{45}Ca injected intravenously leaves the circulation within 5 min, but the serum specific activity is still decreasing 2 months after injection. The initial rapid drop in specific activity is explained by dilution of the ^{45}Ca in the relatively large pool of rapidly exchangeable bone calcium. Although less than 2% of the bone Ca is available for exchange, this is 10–100 times the amount in the serum. Bone also takes up ^{45}Ca by accretion into the

"nonexchangeable" pool. In growing boys the rate of bone-salt formation was found to be two to three times the rate of calcium intake from the intestine, the deficit being made up by resorption. The combination of excretion and slow turnover of nonexchangeable bone accounts for the slow components of the serum-specific activity curve. In adults the processes of accretion and resorption continue at slower rates and are balanced. The distinction between accretion and exchange is possible only with isotopic tracers. Another distinction made possible with ^{45}Ca is the finding, made by comparing fecal contents following ^{45}Ca administration by mouth and by vein, that about 10% of fecal calcium is of endogenous origin, whereas the rest represents a lack of absorption. Comparison of ^{45}Ca absorption and excretion in a boy with that in an adult suggests that absorption depends more on intake than on the need for calcium, and that calcium balance is regulated by excretion in the urine of calcium absorbed in excess of the needs of the body rather than by control of absorption. Depending upon the intake, 15–50% of ingested calcium is excreted in the feces. *See* ABSORPTION (BIOLOGY).

Tritium is especially useful in microautoradiographs because the average beta particle emitted has a range of only 1 micrometer in tissue, giving especially sharp resolution.

As illustrated in **Fig. 3**, chromosomes can be labeled with tritium through the use of tritium-labeled thymidine, a substance which localizes in deoxyribonucleic acid (DNA), which in turn occurs only in the genetic material. After the cells were labeled by being grown in the tritiated thymidine medium, they were transferred to a nonradioactive medium for growth of the second generation. The rate and mechanism of synthesis and degradation of DNA are under intensive study. Several lines of study, such as the labeling of successive generations of bacteria and the absence of label in the second-generation chromosomes marked by the arrows in Fig. 3, indicate that DNA is not normally subject to exchange or turnover processes, but is synthesized only when the cell is preparing to divide. *See* DEOXYRIBONUCLEIC ACID (DNA).

The kinetics of cellular proliferation has provided a rich vein for application of radioisotopic methods. For example, the lifetime of human red blood cells (about 120 days) was established with the use of ^{59}Fe-labeled cells. The life cycles of blood-cell precursors and of white blood cells and platelets have also been studied with a variety of radioisotopes, including tritium.

Some applications, such as the intake of ^{131}I by the thyroid, the measurement of the red-cell mass with ^{51}Cr-labeled red cells, and the absorption of ^{60}Co-labeled vitamin B$_{12}$, are of practical clinical importance in the diagnosis and treatment of disease. Knowledge of the rates of distribution and disposal of a wide variety of radioactive substances is basic to the problem of evaluation of the hazard from fallout radiation. Determination of induced activities, particularly ^{24}Na, provides a basis for calculating the neutron exposure in reactor accidents and in therapeutic exposures to neutrons.   James S. Robertson

Bibliography. L. G. Colombetti (ed.), *Biological Transport of Radiotracers,* 1982; J. S. Robertson, Theory and use of tracer in determining transfer rates in biological systems, *Physiol. Rev.*, 37:133–154, 1957; R. J. Slater (ed.), *Radioisotopes in Biology: A Practical Approach*, 1990; F. Stohlman, Jr. (ed.), *Hemopoietic Cellular Proliferation,* 1970.

Radioisotope (geochemistry)

A branch of environmental geochemistry and isotope geology concerned with the occurrence of radioactive nuclides in sediment, water, air, biological tissues, and rocks. The nuclides have relatively short half-lives ranging from a few days to about 10^6 years, and occur only because they are being produced by natural or anthropogenic nuclear reactions or because they are the intermediate unstable daughters of long-lived naturally occurring radioactive isotopes of uranium and thorium. The nuclear radiation, consisting of alpha particles, beta particles, and gamma rays, emitted by these nuclides constitutes a potential health hazard to humans. However, their presence also provides opportunities for measurements of the rates of natural processes in the atmosphere and on the surface of the Earth. *See* ALPHA PARTICLES; BETA PARTICLES; GAMMA RAYS.

The unstable daughters of uranium and thorium consist of a group of 43 radioactive isotopes of 13 chemical elements, including all of the naturally occurring isotopes of the chemical elements radium, radon, polonium, and several others. Several nuclides in this group are used for measuring the rates of deposition of sediment in lakes and in the oceans and for dating calcium carbonate precipitated inorganically or in the skeletons of corals, mollusks, and other organisms. In addition, the radioactive daughters of long-lived uranium-238 (uranium-234, thorium-230, radium-226, and lead-210) have been used to study the origin of lava flows on active volcanoes, such as Mount Vesuvius near Naples, Italy, and Kilauea on the island of Hawaii.

A second group of radionuclides is produced by the interaction of cosmic rays with the chemical elements of the Earth's surface and atmosphere. This group includes hydrogen-3 (tritium), beryllium-10, carbon-14, aluminum-26, silicon-32, chlorine-36, iron-55, and others. The half-lives of these nuclides range from 2.8 years for iron-55 to 1.5×10^6 years for beryllium-10, which makes them useful for measurements of the rates of a wide range of geological, geochemical, and biochemical processes. The carbon-14 method of dating is especially important in archeology and the geology of the Quaternary Period. This group also includes short-lived radionuclides such as beryllium-7, phosphorus-32, phosphorus-33, and sulfur-35, whose half-lives are measured in days and which find application in the study of atmospheric processes. *See* COSMOGENIC NUCLIDE; DATING METHODS.

A third group of radionuclides is produced artificially by the explosion of nuclear devices, by the operation of nuclear reactors, and by various particle accelerators used for research in nuclear physics. The radioactive fission products and transuranium elements (neptunium, plutonium, and americium) that accumulate in the fuel rods of nuclear reactors must be isolated in underground repositories because of their intense radiation which is harmful to humans and all other forms of life. The dispersion of these radionuclides as a result of accidental explosions of nuclear reactors or during the testing of nuclear weapons creates a health hazard, especially when the radionuclides occur in food, drinking water, and air. The release of these radionuclides into the atmosphere has resulted in their dispersion around the Earth. Some of the radionuclides produced in nuclear reactors decay sufficiently slowly to be useful for geochemical research, including strontium-90, cesium-137, iodine-129, and isotopes of plutonium. The explosion of nuclear devices in the atmosphere has also contributed to the abundances of certain radionuclides that are produced by cosmic rays such as tritium and carbon-14. *See* RADIOACTIVE FALLOUT; TRANSURANIUM ELEMENTS.

Ionium. Thorium-230 (ionium) is a daughter of long-lived uranium-238. It is radioactive and decays to radium-226 with a half-life of 75,200 years. Ionium is used to measure the rate of deposition of sediment in the oceans. The method is based on the fact that atoms of thorium are concentrated with the rock and mineral particles deposited in the ocean, whereas atoms of uranium can remain in seawater for long periods of time. Ionium is thereby separated from its parent, and decays with its characteristic half-life as time passes and sediment continues to accumulate. For practical reasons, the rate of decay of unsupported ionium in the sediment is expressed as the ratio of the activity of ionium divided by the activity of thorium-232, which is the long-lived naturally occurring isotope of this element. The natural logarithm (ln) of the activity ratio (R) of ionium to thorium-232 in the sediment at some depth (h) below the bottom of the ocean is related to the rate (a) of sediment deposition by the equation $\ln R = \ln R_0 - \lambda h/a$, where R_0 is the activity ratio of recently deposited sediment and λ is the decay constant of ionium. The sedimentation rate is determined by analyzing a series of samples taken at intervals along a piston core collected from an oceanographic ship at sea. The activity ratios expressed as $\ln R$ are then plotted versus depth in the core (h) and a straight line is fitted to the data points. The slope (m) of this line yields the sedimentation rate (a) from the equation $m = -\lambda/a$. The results of such measurements indicate that the rate of accumulation of sediment in the ocean basins is of the order of a few millimeters per thousand years or a few meters per million years. The thickness of sediment in the deep basins of the oceans is generally less than 500 m (1600 ft), which indicates that they are quite young when viewed in the perspective of the history of the Earth. This conclusion is compatible with the theory of plate tectonics and sea-floor spreading. *See* MARINE SEDIMENTS; PLATE TECTONICS; THORIUM.

Thorium-230/uranium-234. The calcium carbonate that is precipitated by organisms such as corals and mollusks excludes thorium-230 but admits uranium-234, which is the parent of thorium-230 in the decay chain of uranium-238. Therefore, calcium carbonate, precipitated in the oceans or in saline lakes either by organisms or by inorganic chemical reactions, does not contain thorium-230 at the time of deposition. However, as the calcium carbonate deposit (composed of either calcite or aragonite) ages, thorium-230 is produced by decay of uranium-234. Therefore, the radioactivity due to the presence of thorium-230 increases with time until radioactive equilibrium between uranium-234 and thorium-230 is reestablished. When this condition of secular radioactive equilibrium is achieved, thorium-230 decays at exactly the same rate at which it is produced by decay of uranium-234. However, before that happens the activity ratio of thorium-230 to uranium-234 increases with time. After the thorium-230/uranium-234 activity ratio of a calcium carbonate samples has been measured, its age is determined. This method can be used to date calcium carbonate samples that are less than about 250,000 years old. It was used successfully to determine the ages of coastal terraces on the island of Barbados in the Lesser Antilles. *See* URANIUM; THORIUM.

Lead-210. Radon-222, a radioactive daughter of uranium-238, tends to escape from uranium-bearing minerals into the atmosphere, where it decays rapidly and eventually produces lead-210, which is removed from the atmosphere by rain and snow and is deposited on the surface of the Earth. Lead-210 decays to bismuth-210 with a half-life of 22.3 years, and is therefore useful for dating sediment deposited in lakes and in the coastal ocean during the twentieth century. Sediment deposited in the recent past reflects the effects of pollution of lakes, rivers, and the atmosphere by discharge of industrial and domestic waste. Lead-210 is also used to measure the rate of deposition of snow on the ice sheets of Antarctica and Greenland. The rate of deposition is obtained from the slope of a straight line fitted to measurements of the disintegration rate of lead-210 in snow collected at increasing depths below the surface. An interesting result obtained by this method is that the average annual rate of precipitation at the South Pole is only 2.4 in. (6 cm) per year. Evidently the interior of Antarctica is a cold desert where an ice sheet more than 1.8 mi (3 km) thick has formed by accumulation of snow and ice. The glaciation of Antarctica probably began more than 2.5×10^7 years ago, and has exerted a strong influence on the climate of the Earth for a long time. *See* LEAD ISOTOPES (GEOCHEMISTRY); RADON.

Beryllium-10. Cosmic rays cause nuclear reactions in the atmosphere that result in the formation of beryllium-10, which is rapidly brought down by meteoric precipitation and accumulates in sediment at the bottom of the oceans and in soil on the continents. Beryllium-10 is also produced by nuclear

spallation reactions of high-energy cosmic-ray neutrons in quartz (SiO_2) when this mineral is exposed at the surface. The amount of "cosmogenic" beryllium-10 (and aluminum-26) in quartz exposed to cosmic rays increases with time at a rate that depends on the intensity of the cosmic-ray flux, which increases with elevation above sea level and the geomagnetic latitude of the site. Therefore, the concentrations of beryllium-10 (and aluminum-26) in quartz that has been exposed to cosmic rays can be used to determine the duration of the exposure or its so-called exposure age. This method of dating has been used to determine exposure ages of quartz sandstone on the summits of the Transantarctic Mountains, which are emerging from the East Antarctic ice sheet either because the ice sheet is getting thinner or because the mountains are being uplifted, or both. Cosmic-ray exposure ages can also be measured by means of the concentration of neon-21 in quartz and chlorine-36 in ordinary rocks. These kinds of measurements are possible because of the availability of tandem accelerator mass spectrometers whose detection limit for beryllium-10 and aluminum-26 is about 1 million atoms. *See* BERYLLIUM; COSMIC RAYS; MASS SPECTROSCOPE.

Cesium-137. The testing of nuclear weapons in the atmosphere from about 1945 resulted in the deposition of cesium-137 and other fission products on the surface of the Earth. The injection of this and other fission products into the atmosphere varied from year to year, depending on the frequency and character of the weapon tests. High rates of input of cesium-137 occurred in 1963 and also in 1954. The high radioactivity of sediment deposited in these years is used to correlate sediment deposited in different places and to estimate the rate of deposition. A large quantity of cesium-137 and other radionuclides was discharged into the atmosphere on April 26, 1986, when a nuclear reactor at Chernobyl in the northern Ukraine exploded catastrophically. Fallout from this event was detected in Europe 2 days later, and subsequently spread around the world. *See* CESIUM; ENVIRONMENTAL RADIOACTIVITY; RADIOISOTOPE.

Gunter Faure

Bibliography. N. A. Chapman and G. M. McKinley, *The Geological Disposal of Nuclear Waste*, 1987; G. Faure, *Principles and Applications of Inorganic Geochemistry*, 1998; G. Faure, *Principles of Isotope Geology*, 2d ed., 1986; T. -L. Ku, The uranium-series methods of age determination, *Rev. Earth Planetary Sci.*, vol. 4, 1976; W. W. Nazaroff and A. V. Nero, Jr. (eds.), *Radon and Its Decay Products in Indoor Air*, 1988.

Radiolaria

A group of marine protists, regarded as a subclass of Actinopodea in older classifications, but not recognized as a natural group in some modern systems owing to its heterogeneity. Radiolarians occur almost exclusively in the open ocean as part of the plankton community, and are widely recognized for
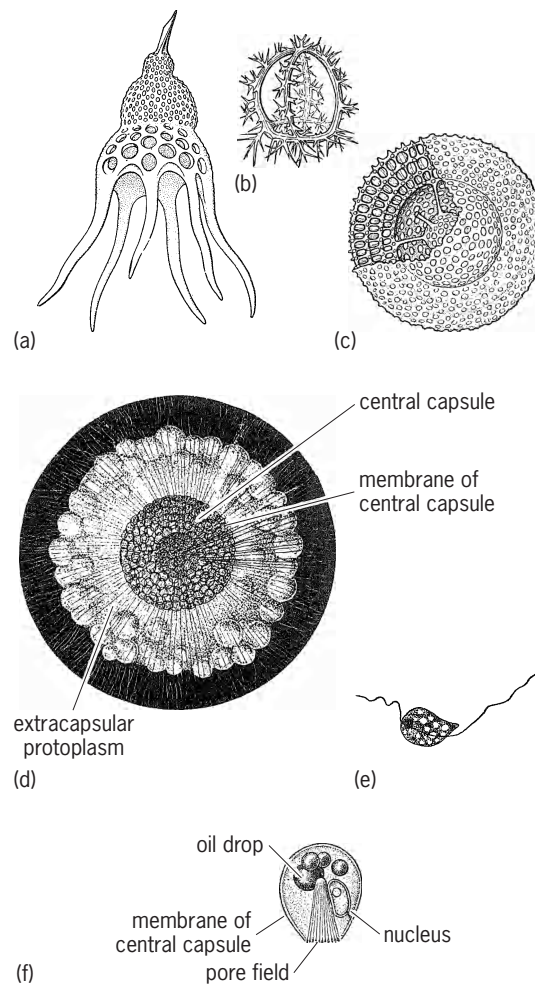
Radiolaria. Skeletons representing (*a*, *b*) certain Nasselarida and (*c*) certain Spumellarida. (*d*) *Thalassicolla*, one of the Spumellarida without skeletal elements. (*e*) Biflagellate gametes. (*f*) Central capsule, one of the Nasselarida showing one group of pores. (*After L. H. Hyman, The Invertebrates, vol. 1, McGraw-Hill, 1940*)

their ornate siliceous skeletons produced by most of the groups (**illus.** *a–c*). Their skeletons occur abundantly in ocean sediments and are used in analyzing the layers of the sedimentary record (biostratigraphy). Although most radiolarians occur only in open ocean locations, a unique skeleton-bearing nasselarian radiolarian (*Lophophaena rioplatensis*) has been identified in a coastal location in the Rio de la Plata estuary in South America.

Phylogeny. In some classification systems, the Radiolaria are subdivided into two classes, Polycystinea and Phaeodarea. However, clarification of the evolutionary relationships (phylogeny) of the Radiolaria using modern molecular genetic analyses indicates that Polycystinea and the Phaeodarea are undoubtedly polyphyletic: that is, they are not derived from closely related phylogenetic ancestors. Therefore, current taxonomic systems separate them.

Molecular genetic analyses have begun to clarify the phylogenetic relationships among some of the major groups of Polycystinea. The colonial and skeletonless species (for example, *Thalassicolla*) form

a lineage that is distinct from the skeleton-bearing species. Also, there is evidence that the Polycystinea and Acantharia (with skeletons composed of strontium sulfate) are phylogenetically closely related. These data may indicate that Acantharia and Polycystinea should be grouped together taxonomically. This is of interest historically since Ernst Haeckel, a prominent nineteenth-century zoologist, originally grouped them together as members of the class Radiolaria. However, we now recognize that much of his system of classification is artificial; that is, it is based on superficial similarities of form that are not indicative of close evolutionary relationships. Much additional work needs to be done using modern techniques of electron microscopy and molecular genetics to more fully elucidate the natural systematic relationships of species included in the Radiolaria.

Morphology. A characteristic feature of the group is the capsule, a central mass of cytoplasm bearing one or more nuclei, food reserves, and metabolic organelles. This is surrounded by a perforated wall and a frothy layer of cytoplasm known as the extracapsulum, where food digestion generally occurs and numerous axopodia (stiffened strands of cytoplasm) and rhizopodia radiate toward the surrounding environment. Radiolarians feed on microplankton captured by the sticky axopodia. Their prey include bacteria, algae, some protozoa, and small arthropod zooplankton such as copepods. Algal symbionts, including dinoflagellates, green algae, and golden-brown pigmented algae, occur profusely in the extracapsulum. The algal symbionts living within the protection of the extracapsulum provide photosynthetically derived food for the radiolarian host.

Polycystinea. In the order Spumellarida (class Polycystinea) the central capsule is perforated by numerous pores distributed evenly on the surface of the capsular wall. These pores, containing strands of cytoplasm, provide continuity between the cytoplasm in the central capsule and the surrounding extracapsulum. The skeletons of the Spumellarida are characteristically developed on a spherical organizational plan, but some are spiral-shaped (resembling snail shells) or produce elongate skeletons composed of numerous chambers built one upon another. In some genera, such as *Thalassicolla* (illus. *d*), there is no skeleton; in others there are rods or spicules, or often a single or multiple concentric latticework skeleton (illus. *c*). The shape of the skeleton is determined by the dynamic action of the radiolarian cytoplasm that secretes the silica within a specialized part of the rhizopodial system called a cytokalymma.

Multicellular aggregates (colonies of spumellaridans), measuring several centimeters in diameter (or even several meters in some rare elongate forms), consist of numerous radiolarian central capsules enclosed within a gelatinous envelope and interconnected by a web of rhizopodia that bears abundant algal symbionts. A thin halo of feeding rhizopodia protrudes from the surface of the colony and is used to capture prey. The kinds of prey observed in solitary and colonial forms of Spumellarida include copepods, small crustacean larvae, and a variety of algae (diatoms, dinoflagellates, and others). Reproduction is poorly understood. In some colonial forms, daughter colonies are produced by asexual reproduction (fission). Flagellated swarmers (illus. *e*) released from mature central capsules of some species are possibly gametes and have a large crystal inclusion of strontium sulfate.

In the Nasselarida (Polycystinea), the central capsule is often ovate and the pores are localized at one pole (illus. *f*). The axopodia and rhizopodia emerge from this pore field and are supplied by a cone-like array of microtubules within the central capsule. The skeleton, when present, is often shaped like a dome or helmet, but some forms may be ornamented with lateral lattices, spines, or other embellishments. Only dinoflagellate algae have been observed in association with the Nasselarida.

Phaeodarea. Radiolarians in the class Phaeodarea possess a central capsule with two types of pore areas, a larger one (astropyle) that serves as a kind of cytopharynx where food is carried into the central capsule, and two accessory openings (parapylae) where smaller strands of cytoplasm emerge. The phaeodarian skeleton exhibits a wide range of shapes, including geodesic-like lattice spheres and small porous clam-shaped shells. No symbiotic algae have been reported in the Phaeodarea.

Given the peculiarities of the two kinds of pores in the central capsule, and their distinct differences from the fusules (cytoplasmic strands) of the Polycystinea, it is very probable (as indicated by some modern molecular genetic analyses) that the Phaeodarea are not closely related to Polycystinea. Similarities in the capsular organization may occur through adaptive convergence; that is, over evolutionary time, these two distinct lineages have adapted to their environment by developing similar morphological features, including siliceous skeletons.

Fossil forms. Radiolarians have a fossil record that extends back to early Paleozoic time, about a half billion years ago. Compared with other groups of skeleton-bearing marine microplankton, they are highly diverse, several hundred species having inhabited the oceans at any given time. Their skeletons of opaline silica are more readily dissolved than are calcareous and organic-walled microfossils; therefore, radiolarians occur in rocks much less commonly than do foraminifera, calcareous nannofossils, and palynomorphs (pollen, spores, and the like). However, the microfossil record of radiolarians in deep-sea sediments and in some rock formations, including cherts and limestones, is very good and has been used by micropaleontologists to elucidate the geological history of the Earth.

In Cenozoic sediments radiolarians usually occur in clays or shales, but older occurrences are often in calcareous nodules or cherts, where they have been protected from the destructive effects of pressure, chemical change, and dissolution. Some notable

land-based occurrences are in late Cenozoic diatomites of California, Japan, and the Mediterranean region; Mesozoic cherts of California and the Alps; and Paleozoic calcareous and phosphatic nodules in England, Australia, and the Canadian Arctic.

Radiolarians have undergone such profound evolutionary changes through time that it is difficult to determine the relationships between Paleozoic, Mesozoic, and Cenozoic forms. Because they are planktonic and have undergone continuous evolutionary change, radiolarians are particularly useful for determining time equivalence (and geological ages) of marine sedimentary deposits at widely separated localities. As with many other groups of organisms, a particularly abrupt change occurred at the Mesozoic-Cenozoic boundary. The Cenozoic record of radiolarians in sediments, particularly on the deep-sea floor, is sufficiently complete to show the course of evolutionary change in considerable detail.

Assemblages of fossil radiolarians also provide clues to oceanic conditions during the geological past. Each of the major oceanic water masses has its characteristic radiolarian fauna, and so changes in the distribution or composition of these assemblages can be interpreted in terms of changes in the pattern of water masses, or in their oceanographic properties. Since some species of Radiolarians inhabit surface water of the oceans in specific climatic regions with narrow temperature requirements for survival, the skeletons of radiolarians in sediments can be used to clarify the ancient climates of the Earth. They also have been used to analyze ancient terranes, since the species of radiolarian fossils isolated from surface rocks can be used to date the origin of the rocks and, in some cases, their original geographic locale.

Some species of radiolarians dwell only in the surface water of the ocean, requiring sufficient biological productivity to feed and survive, while others occur only at great depths. The deep-dwelling species are less dependent on surface productivity. Hence, the proportion of surface-dwelling species to deep-dwelling species deposited as microfossils within a stratum of the sediments can be used as evidence of the amount of productivity of the surface water in the ancient seas. *See* ACANTHARIA; ACTINOPODEA; OCULOSIDA; PROTOZOA; SARCODINA; SARCOMASTIGOPHORA.

O. Roger Anderson; William R. Riedl

Bibliography. O. R. Anderson, The physiological ecology of planktonic sarcodines with applications to paleoecology: Patterns in space and time, *J. Eukaryotic Microbiol.* 43:261–274, 1996; O. R. Anderson, *Radiolaria*, 1983; O. R. Anderson, The trophic role of planktonic foraminifera and radiolaria, *Mar. Microb. Food Webs*, 7:31–51, 1994; D. Boltovskoy et al., First record of a brackish radiolarian (Polycystina): *Lophophaena rioplatensis* n. sp. in the Rio de la Plata estuary, *J. Plankton Res.*, 25:1551–1559, 2003; B. U. Haq and A. Boersma (eds.), *Introduction to Marine Micropaleontology*, 1978; D. Lazarus, A brief review of radiolarian research, *Paläontologische zeitschrifft*, 79:183–200, 2005.

Radiology

The medical science concerned with x-rays, radioactive materials, and other ionizing radiations, and the application of the principles of this science to diagnosis and treatment of disease. Nonionizing radiations of infrared and ultrasound are also used for diagnosis.

Diagnostic radiology. This technology uses radiation, usually x-rays, to study the configuration of anatomical structures or the function of body organ systems. The most common techniques used in diagnostic radiology are discussed below.

Radiography is the formation of an image, usually on a photographic emulsion, by the action of ionizing radiation. The image is the result of the differential absorption of the radiation in its passage through a part of the body. The ionizing radiation is usually x-rays, but protons and fast neutrons have also been used. Contrast media have been developed which help to visualize body structures such as blood vessels, kidney, and gallbladder. Most of these media contain firmly bound iodine, which has a relatively high atomic number. There is a probability, usually quite low, that the patient may have a reaction to the contrast media. Physicians must always be concerned about benefit-cost considerations when obtaining radiographic studies, so as to balance the benefit of the information gained against the monetary, radiation, and possible drug reaction cost. *See* RADIOGRAPHY.

Fluoroscopy is the examination of the structure of a part of the body by passing a beam of x-rays through it and viewing the image formed on a fluorescent screen by the transmitted radiation. This permits visualization of internal organs in motion. In some cases, such as a gastrointestinal examination, contrast is increased by having the patient drink a dense fluid such as a suspension of barium sulfate.

Image intensification of the fluoroscopic image by an electronic device (orthicon, plumbicon, vidicon) permits visualization of the internal organs with x-ray beams of lower intensity than those used in direct fluoroscopy. This permits a decrease in radiation exposure received by the patient. The intensified image may be viewed on a television screen, a practical advantage for the physician and a source of considerable interest to some patients.

There are studies in diagnostic radiology with special names which usually indicate the organs to be examined, such as cholestography (gallbladder), angiocardiography (heart and large blood vessels), and neuroradiology (brain and spinal cord).

Nuclear medicine. Radioactive isotopes are used to obtain images of organ systems and functions. A compound such as radioactive technetium-pertechnetate (^{99m}Tc-pertechnetate) after intravenous injection is

used to locate brain tumors, and thyroid function is studied by the uptake of radioactive iodine (usually 131I) by the thyroid gland after oral administration. The accumulation of isotope in a tumor or an organ such as the thyroid is recorded by a suitable gamma-ray detector such as a gamma-ray-sensitive crystal attached to an electronic amplifier and recording equipment. The image of the radioactivity concentrated in an organ is viewed on a television-type screen and recorded on a photographic print. A large number of body organs and functions can be studied with a variety of commercially available radiopharmaceutical compounds.

Ultrasound. Sound waves of 1–10 MHz/s in pulses of 1.5 μs duration are transmitted from a crystal transducer, and after amplification are displayed as A scan or B scan on an oscilloscope and recorded on a photographic print. The ultrasound pulses demonstrate organ structures such as the heart, liver, and spleen. Although the resolution is less fine than that obtained with x-ray there is an advantage in that the ultrasound is nonionizing radiation. Ultrasound is particularly useful, therefore, in determining the size and degree of development of the human fetus.

Thermography. Infrared radiation from the human body is used to detect tumors such as breast tumors, which are near the body surface. The technique, which is still experimental, is based on the idea that tumors are warmer than the surrounding normal tissue. This increase in temperature is detected by an infrared device, and the "hot spot" scan is displayed on a television-type screen, with permanent records kept on photographic prints.

Ultrasound and thermography are techniques which are used to supplement the information obtained by diagnostic radiology and nuclear medicine studies.

Therapeutic radiology. Radiation therapy deals with the treatment of disease with ionizing radiation. The diseases most commonly treated are cancer and allied diseases. Radiation therapy has been found useful in the management of some diseases such as ringworm of the scalp and bursitis, but because of possible serious complications occurring many years later, the use of ionizing radiation is generally avoided if alternative methods of treatment are available.

In cancer therapy the objective is to destroy a tumor without causing irreparable radiation damage in normal body tissues that must of necessity be irradiated in the process of delivering a lethal dose to the tumor. This applies particularly to important normal structures in the vicinity of the tumor. The relative radiosensitivity of the tumor with respect to these normal structures is the chief factor determining the success of the treatment. The effect of the radiation on the tumor and other tissues becomes apparent gradually, in a matter of weeks. During this latent period a certain amount of recovery or repair may also take place.

The optimal differential between the effect on the tumor and the effect on normal tissues of the patient as a whole results from the proper adjustment of many treatment factors, which in general requires great clinical experience. The time factor is very important. This involves the administration of the therapeutic dose in one of three ways: by one short treatment, by protraction as continuous irradiation over a long time, or by fractionation in small repeated doses. The importance of the distribution of radiation in the patient's body is obvious. Ideally, only the tumor should be irradiated, but this is impossible, because in general there is infiltration of adjacent normal tissues of unknown extent. If the radiosensitivity of the tumor is much higher than that of the surrounding normal tissue, the problem is relatively simple, but in general this is not the case.

Various techniques have been devised to bring about a reasonable distribution of radiation in the region of the tumor. When x-rays are used for the treatment of deep-seated tumors, the beam must traverse a considerable thickness of normal tissues. To minimize absorption in these tissues, higher and higher voltage x-rays have been used. Supervoltage x-ray therapy makes use of x-rays produced at several hundred kilovolts to 2000 kV. Megavoltage x-rays range from a few to 70 MV. Cross-fire treatment involves the use of more than one beam aimed at the tumor, but passing through different skin areas. Rotation therapy involves the use of one beam constantly aimed at the tumor, while either the patient or the source of radiation is rotated about an axis passing through the center of the tumor. When the source is rotated, 1000–2000 Ci of cobalt-60 (^{60}Co) are generally used to provide a well-collimated beam of gamma rays.

Fast neutrons, high-energy electrons, and deuterons, as external beams, have been used for therapeutic purposes. By using radioactive substances, other techniques are possible. In the case of intracavitary therapy, the source, in a suitable container, is placed within a body cavity in which cancer is present. In the case of interstitial therapy, many needlelike sources are distributed throughout the tumor and removed at the completion of the treatment. By using radon, which has a short half-life, the sources may be made into radon seeds, which are very small and may be left in place permanently. Sources of tantalum-182 (^{182}Ta) are sometimes used in a similar way. Attempts to infiltrate a tumor with radioactive material in fluid form have not been very successful, but some good results have been obtained by using colloidal gold-198 in the treatment of cancer of the prostate. The same material has been used for the palliative treatment of ascites tumors in the pleural or abdominal cavity. Phosphorus-32 has been found useful in the treatment of polycythemia vera and leukemia by injection into the bloodstream. Iodine-131, which, when introduced into the body by any route, tends to concentrate almost entirely in the thyroid, is very effective in reducing the size and function of the thyroid in patients suffering from hyperthyroidism. Certain types of cancer of the thyroid and metastases can be treated similarly with radioactive iodine, but the doses must be much larger and the arrest of the tumor process is generally temporary. *See* NUCLEAR RADIATION (BIOLOGY); RADIATION BIOLOGY.   Lee B. Lusted

Radiometry

A branch of science that deals with the measurement or detection of radiant electromagnetic energy. Radiometry is divided according to regions of the spectrum in which the same experimental techniques can be used. Thus, vacuum ultraviolet radiometry, intermediate-infrared radiometry, far-infrared radiometry, and microwave radiometry are considered separate fields, and all of these are to be distinguished from radiometry in the visible spectral region. Curiously, radiometry in the visible is called radiometry, optical radiation measurement science, or photometry, but it is not called visible radiometry. *See* ELECTROMAGNETIC RADIATION; INFRARED RADIATION; LIGHT; MICROWAVE; ULTRAVIOLET RADIATION.

The use of the word photometry to mean radiometry in the visible portion of the spectrum is misleading. Strictly speaking, photometry is the measurement of electromagnetic radiation according to its ability to produce visual sensation. This, of course, is a subjective attribute of the radiation. Radiometry, on the other hand, is concerned with physical attributes of the radiation such as its energy content and spectral distribution. The reason for the confusion between photometry and radiometry is that the meanings of these words have changed considerably over the last 200 years. *See* PHOTOMETRY.

Development and nomenclature. During the 1700s, techniques were developed for measuring light using the human eye as a null detector to compare the relative intensity of light coming from different sources. This was referred to as photometry. During the same period, radiant heat was studied with sensitive liquid-in-glass thermometers, and this activity was referred to as radiometry. Also during this period, the photochemical activity of solar radiation was studied by observing the photoinduced decomposition of silver compounds into metallic silver, and this activity was called actinometry. At that time it was not known that radiant heat and light were the same phenomenon; the demonstrations that light was a wave phenomenon and that mechanical energy and heat were different forms of a more general conception of energy were to come; and the electromagnetic and quantum theories of radiation had not been proposed. Therefore, the emphasis of these studies was not to quantify the radiation, but rather to learn the nature of radiation and its sources.

Following the discovery of infrared radiation in 1800 and of ultraviolet radiation in 1801, a great deal of effort was devoted to studies of their properties. The only practical detectors of ultraviolet radiation at that time were photochemical reactions, and the only detectors of infrared radiation were thermometers, while the eye remained the most sensitive detector of visible radiation. Thus actinometry, photometry, and radiometry became synonymous with studies of radiation in the ultraviolet, visible, and infrared spectral regions, respectively.

It is modern practice to refer to any radiation detector (such as a thermometer) that responds to an increase in temperature caused by the absorption of radiant energy as a thermal detector. Similarly, any detector (such as a photochemical reaction) that responds to the excitation of a bound electron is called a photon or quantum detector. Almost all of the many different types of modern radiation detectors fall into one of these two classes.

By the end of the 1800s, a number of different effects that could be used as the basis for thermal detectors had been discovered, and some of these were sensitive enough for use in the ultraviolet. Similarly, a number of effects that could be used as the basis for photon detectors had been discovered, and some of these responded to infrared radiation. Furthermore, both thermal and quantum detectors could be used quite satisfactorily with visible radiation for certain applications. As a result of these developments, the distinctions between radiometry, actinometry, and photometry in terms of spectral regions became artificial, and the words began to assume new meanings. Radiometry became the more general term, while photometry and actinometry came to denote the visual and photochemical effects of electromagnetic radiation, respectively.

Nevertheless, the older meanings continued to co-exist with the newer. Even now, astronomers refer to radiometry as photometry, and their reference to infrared photometry would be a contradiction in terms according to the modern definition of photometry in terms of the ability to produce visual sensation.

Thermal detectors. Liquid-in-glass thermometers are sluggish and relatively insensitive. The key to developing thermal detectors with better performance than liquid-in-glass thermometers has been to secure a large and rapid rise in temperature associated with a high sensitivity to temperature changes.

The temperature of any thermal detector will increase until the rate of loss of heat to its surroundings is equal to the rate at which heat is absorbed. To secure the largest rise for a given radiation, the detector should absorb the radiation as completely as possible, while the loss of heat, by reradiation, conduction, and convection, must be as small as possible. To secure a rapid rise in temperature, the heat capacity of the detector should be as small as possible. To obtain the least interference from temperature fluctuations, the absorbing surface should be as small as possible.

These criteria lead to detectors which are very small and thin, with thin black coatings housed in vacuum containers with a window to admit the radiation. Special applications often require that this optimized design be sacrificed for improved accuracy, convenience, or robustness. Thus large-area detectors without windows have been used as standards against which the most sensitive detectors are calibrated.

Thermal detectors have been based upon a number of different principles. Radiation thermocouples produce a voltage, bolometers undergo a change in resistance, pyroelectric detectors undergo a change in spontaneous electric polarization, and the gas in pneumatic detectors (Golay cells) and photoacoustic detectors expands in response to incident radiation.
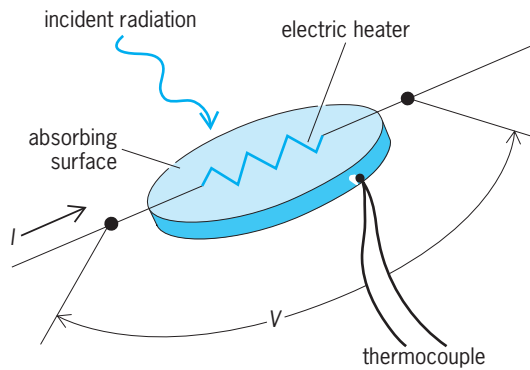
Fig. 1. Absolute radiometer. The electrical power dissipated in the heater is given by the product of the voltage drop *V* across the heater times the current *I* passing through the heater.

The periodic expansion and contraction of the gas in response to high-frequency modulated radiation is detected by a sensitive microphone in the case of the photoacoustic detector. The Golay cell, on the other hand, uses a sensitive photomultiplier and a reference beam of light to detect distortion of a flexible membrane mirror caused by the expansion and contraction of the gas. *See* BOLOMETER; PYROELEC-TRICITY; THERMOCOUPLE.

Thermal detectors with electrical heaters built into the part of the detector that absorbs the radiation can be used as radiation standards. The detector output caused by the unknown radiation is compared with the output caused by a measured quantity of electrical power dissipated in the heater to determine the radiant power content of the radiation. Thermal detectors of this nature are called absolute radiometers (**Fig. 1**).

The main problem with thermal detectors is that they respond not only to electromagnetic radiation but to any source of heat. This makes their design, construction, and use rather difficult, because they must be made sensitive to the radiation of interest while remaining insensitive to all other sources of heat, such as conduction, convection, and background radiation, that are of no interest in the particular measurement.

Photon detectors. These detectors, on the other hand, respond only to photons of electromagnetic radiation that have energies greater than some minimum value determined by the quantum-mechanical properties of the detector material. Since heat radiation from the environment at room temperature consists of infrared photons, photon detectors for use in the visible can be built so that they do not respond to any source of heat except the radiation of interest.

Except for certain measurements of interest in photochemistry and photobiology, chemical reactions that respond to radiation are not very convenient detectors. The key to developing photon detectors that are more convenient than photochemical reactions has been to base them on some electrical principle, so that the most advanced electronic techniques could be used to detect and process the

detector signal. So many different approaches have been used that only the most important will be mentioned here.

After vacuum-tube technology was introduced, its techniques were applied to the external photoelectric effect to produce vacuum photodiodes and photomultipliers. For ultraviolet and short-wavelength visible radiation, photomultipliers are the most sensitive detectors available. By using photomultipliers in properly designed experiments, it is possible to record the absorption of a single photon. This mode of operation is called photon counting. *See* PHOTO-MULTIPLIER.

Following the introduction of planar silicon technology for microelectronics, the same technology was quickly exploited to make planar photodiodes based on the internal photoelectric effect in silicon. In these devices (**Fig. 2**) the separation of a photogenerated electron-hole pair by the built-in field surrounding the p^+n junction induces the flow of one electron in an external short circuit (such as the inputs to an operational amplifier) across the electrodes. The number of electrons flowing in an external short circuit per absorbed photon is called the quantum efficiency. The use of these diodes has grown to the point where they are the most widely used detector for the visible and nearby spectral regions. Their behavior as a radiation detector in the visible is so nearly ideal that they can be used as a standard, their cost is so low that they can be used for the most mundane of applications, and their sensitivity is so high that they can be used to measure all but the weakest radiation (which requires the most sensitive photomultipliers). *See* JUNCTION DIODE; PHOTODIODE; SEMICONDUCTOR DIODE.

Research efforts have been directed at producing photon detectors based on more exotic semiconductors, and more complicated structures to extend the sensitivity, time response, and spectral coverage of the radiometric techniques that have been
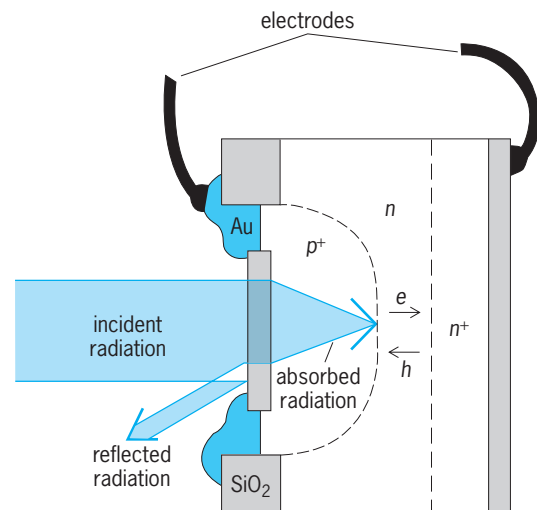
Fig. 2. Silicon photodiode. The separation of a photogenerated electron-hole pair (*e-h*) by the built-in field surrounding the p^+n junction induces the flow of one electron in an external short circuit across the electrodes.

developed around the planar silicon photodiode. *See* OPTICAL DETECTORS.

Radiometric sources. Any source of optical radiation can be used as a radiometric source. Lasers have become important radiometric sources. *See* LASER.

Some sources of radiation have properties that have made them useful as standards at various times. Early standard sources include candles of specified composition and construction, gas mantles, and lamps. The best-known standard source is the blackbody, which emits a unique spectral distribution of radiation dependent only upon its temperature. This distribution is given by Planck's law. No real body can behave as a true blackbody, but cavities surrounded by freezing metals can be used as blackbodies to a very high level of approximation. **Figure 3** illustrates
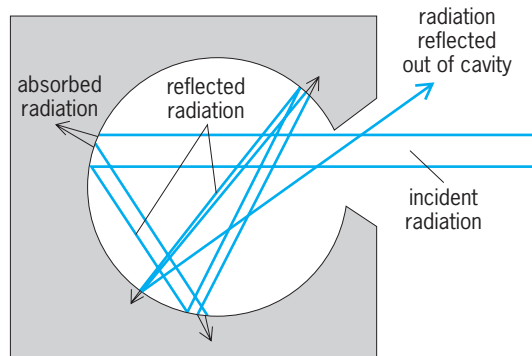
Fig. 3. Cavity approximating a blackbody, that is, an object that absorbs all of the radiation incident on it.

how the cavity geometry reduces the reflectance of the cavity opening below that of the surface of the cavity by multiple reflection and absorption of the radiation before it leaves the interior of the cavity. Thus the cavity appears blacker (more absorptive) than a flat surface made from the same material. Other standard sources in use include the radiation from plasma, and synchrotron radiation. These are used as standards primarily in the vacuum ultraviolet. *See* HEAT RADIATION; SYNCHROTRON RADIATION.

Jon Geist

Bibliography. R. W. Boyd, *Radiometry and the Detection Optical Radiation* 1983; W. Budde, *Physical Detectors of Optical Radiation*, vol. 4 of F. Grum and C. J. Bartleson (eds.), *Optical Radiation Measurements*, 1983; F. Grum and R. J. Becherer, *Radiometry*, vol. 1 of F. Grum (ed.), *Optical Radiation Measurements*, 1979; R. McCluney, *Introduction to Radiometry and Photometry*, 1994; C. L. Wyatt, *Radiometric Calibration: Theory and Methods*, 1978; C. L. Wyatt, *Radiometric Systems Design*, 1987.

Radish

A cool-season annual or biennial crucifer, *Raphanus sativus*, of Chinese origin belonging to the plant order Capparales. The radish is grown for its thickened hypocotyl, which is eaten uncooked as a salad vegetable (see **illus.**). Propagation is by seed.

Radish (*Raphanus sativus*), cultivar Red Boy. (*Joseph Harris Co., Rochester, New York*)

Varieties (cultivars) are classified according to root shape and season or time of maturity. Colors include red, yellow, white, black, pink, and red-white combinations. Popular varieties are short-season (21–25 days), Early Scarlet Globe and Comet; medium-season (30–50 days), Crimson Giant; and long-season or winter varieties (50–70 days), Black Spanish. Commercial production, largely of the round red short-season varieties, is primarily in the field but radishes are also produced commercially in greenhouses. Harvesting by hand or machine begins when roots are approximately $1/2$ to 1 in. (13 to 25 mm) in diameter, often only 21–23 days after planting. *See* CAPPARALES. H. John Carew

Radium

A chemical element, Ra, with atomic number 88. The atomic weight of the most abundant naturally occurring isotope is 226. Radium is a rare radioactive element found in uranium minerals to the extent of 1 part for about every 3×10^6 parts of uranium. Chemically, radium is an alkaline-earth metal having properties quite similar to those of barium. Radium is important because of its radioactive properties and is used primarily in medicine for the treatment of cancer, in atomic energy technology for the preparation of standard sources of radiation, as a source for actinium and protactinium by neutron bombardment,

Physical properties of radium	
Property	Value
Atomic number	88
Atomic weight	226.05
Valence states	0, 2+
Specific gravity	6.0 at 20°C
Melting point	700°C (1290°F)
Boiling point	~1140°C (2080°F)
Ionic radius, Ra^{2+}	0.245 nm (estimated)
Atomic parachor	~140
Decomposition potential	1.718 volt
Heat of formation of oxide	130 kcal/mole
Magnetic susceptibility	Feebly paramagnetic

and in certain metallurgical and mining industries for preparing gamma-ray radiographs. *See* PERIODIC TABLE.

Thirteen isotopes of radium are known; all are radioactive; four occur naturally; the rest are produced synthetically. Only ^{226}Ra is technologically important. It is distributed widely in nature, usually in exceedingly small quantities. The most concentrated source is pitchblende, a uranium mineral containing about 0.014 oz (0.4 g) of radium per ton of uranium.

Biologically, radium behaves as a typical alkaline-earth element, concentrates in bones by replacing calcium, and, as a result of prolonged irradiation, causes anemia and cancerous growths. The tolerance dose for the average human being has been estimated at a total of 1 μg of radium fixed within the body. However, because radiations from radium and its decay products preferentially destroy malignant tissue, radium and radon, the gaseous decay product of radium, have been used to check the growth of cancer.

When first prepared, nearly all radium compounds are white, but they discolor on standing because of intense radiation. Radiation causes a purple or brown coloration in glass on long contact with radium compounds. Eventually the glass crystallizes and becomes crazed. Radium salts ionize the surrounding atmosphere, thereby appearing to emit a blue glow, the spectrum of which consists of the band spectrum of nitrogen. Radium compounds will discharge an electroscope, fog a light-shielded photographic plate, and produce phosphorescence and fluorescence in certain inorganic compounds such as zinc sulfide. The emission spectrum of radium compounds is similar to those of the other alkaline earths; radium halide imparts a carmine color to a flame.

When freshly prepared, radium metal has a brilliant white metallic luster. Some of its physical properties are shown in the **table**. Chemically, the metal is highly reactive. It blackens rapidly on exposure to air because of the formation of a nitride. Radium reacts readily with water, evolving hydrogen and forming a soluble hydroxide. *See* ALKALINE-EARTH METALS; NUCLEAR REACTION; RADIOACTIVE MINERALS; RADIOACTIVITY; RADON. Murrell L. Satutsky

Bibliography. D. R. Lide, *CRC Handbook of Chemistry and Physics*, 85th ed., CRC Press, 2004; G. Van Kaick et al., *Health Effects of Internally Deposited Radionuclides: Emphasis on Radium and Thorium: Proceedings*, World Scientific, 1995; V. M. Vdovenko and Yu. V. Dubasov, *Analytical Chemistry of Radium*, 1975.

Radius of gyration

The radial distance from a rotation axis at which the mass of an object could be concentrated without altering the moment of inertia of the body about that axis. If the mass M of the body were actually concentrated at a distance k from the axis, the moment of inertia about that axis would be Mk^2. If this quantity is equal to the actual moment of inertia I about the rotation axis, then k is the radius of gyration given by $k = \sqrt{I/M}$. The quantity k has dimensions of length and is measured in appropriate units of length such as meters. *See* MOMENT OF INERTIA.

For example, consider a rigid object of arbitrary shape with mass M rotating about the z axis with a moment of inertia I (see **illustration**). The equiva-
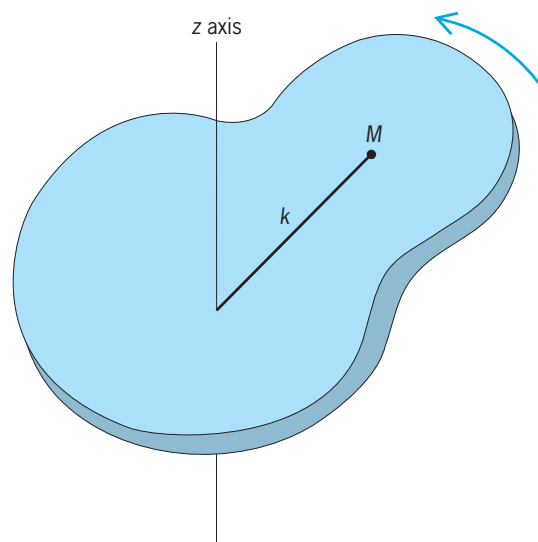
A rigid object of mass *M* rotates about the *z* axis with a moment of inertia *I*. A point mass *M* a distance *k* from the *z* axis rotates about the axis with the same moment of inertia.

lent moment of inertia would be attained if the entire mass of the object could be concentrated at one point a distance k from the rotation axis. That distance k is the radius of gyration. Edwin R. Jones

Bibliography. H. Goldstein, C. Poole, and J. Safko, *Classical Mechanics*, 3d ed., Addison-Wesley, Boston, 2002; J. B. Marion and S. T. Thornton, *Classical Dynamics of Particles and Systems*, 5th ed., Brooks Cole, Belmont, CA, 2004; K. R. Symon, *Mechanics*, 3d ed., Addison-Wesley, Boston, 1971.

Radome

A protective housing for radar, satellite communications, or similar antennas, often spherical but sometimes shaped to satisfy aerodynamic or other requirements of the installation. Radomes are designed to

satisfy both electromagnetic and mechanical requirements.

Radomes protect antennas from harmful environmental conditions. Shielding mechanically scanning antennas from winds, for example, reduces the mechanical power needed to ensure proper motion. Properly designed radomes can lessen the detrimental effects of ice, snow, and rain accumulation. *See* ANTENNA (ELECTROMAGNETISM).

Radomes on aircraft must be shaped for proper flight characteristics. Top-mounted larger antennas (as in airborne early warning radars) are housed in rotating radomes or "rotodomes," turning with the antenna inside; smaller antennas (as in airborne intercept radars or those in commercial aircraft) operate behind shaped stationary radomes at the nose of the aircraft or beneath it.

Spherical radomes for surface installations, surmounting a tower or on a perimeter foundation on the ground, may be tens of meters in diameter, large enough to allow the motion of the antenna inside. They may be constructed of thin dielectric sheets using scores of polygonal pieces, joined at the turned-in edges of the pieces or by attaching them to a skeletal framework of short metal members. Some domes are made of thin flexible material supported by positive relative air pressure maintained on the inside.

Some antennas have protective covers mounted on the antenna itself; the dish-shaped antennas of precision tracking radars and microwave terrestrial relays are often protected this way. Phased-array antennas (usually planar fixed apertures) may have flat dielectric sheets as the outer surface, contributing to both the antenna's impedance matching and its protection from the physical environment.

In general terms, the radome must be simultaneously transparent to the radar or communication signal and structurally sound. The designer has available many materials of varying characteristics to help in meeting these joint requirements such as thin rubberlike skins for air-supported domes and rigid sheets of various resins with reinforcing glass fibers when more strength is required. The dielectric nature of all such materials, causing both absorptive loss in the material itself and surface reflections, is of prime interest to the designer. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; RADAR-ABSORBING MATERIALS; REFLECTION OF ELECTROMAGNETIC RADIATION.

Most materials used are nonmagnetic (relative permeability near unity) but have widely varying permittivity (dielectric constant). Permittivity is a complex value, and the ratio of its imaginary and real values, the "loss tangent," relates to how lossy the material is. Loss tangents in the range of 0.01 to 0.001 are common; absorptive loss in thin sheets can be as little as 0.1 dB and in thicker rigid material still considerably less than 1 dB. *See* PERMITTIVITY.

When strength requires thicker material, materials of low loss tangent will achieve low absorptive loss, leaving surface reflections the major concern. Thickness of the material can be chosen to minimize the combined reflection from the two surfaces,

inside and out; thicknesses of integral multiples of half-wavelengths (in the dielectric) give this favorable resonant condition. Higher values of permittivity sustain good performance over a large range of incidence angles required for large antennas in shaped radomes. Material of dielectric constant 4.2 and loss tangent 0.014 has, for about half-wavelength (in the material) thickness, a transmission efficiency of 94% at $0°$ incidence, but much less for incidence greater than $30°$ or so. Values of dielectric constant 9.0 and loss tangent 0.0004 give greater than 93% transmission efficiency for incidence up to $80°$. The higher-permittivity material, however, requires far tighter manufacturing tolerances and introduces some other detrimental effects (such as polarization distortions). When still thicker material is required for strength, various "sandwich" designs with multiple layers are used. *See* INTERFERENCE OF WAVES; REFRACTION OF WAVES.

Special radomes can be made transparent in just one frequency band and reflective (opaque) at all other frequencies, generally by using conductor-clad dielectric material with slots of frequency-sensitive dimensions etched in the conducting surfaces. Radomes have also been built with artificial dielectric materials (with embedded conductors or such) that can be electronically excited for control of the characteristics. *See* DIELECTRIC MATERIALS.

Not all radar antennas require radomes in their installation. When required, however, the radome becomes a significant component of the radar demanding careful attention to its design. *See* RADAR.

Robert T. Hill

Bibliography. H. E. Schrank et al., Chap. 6, *Reflector antennas*, in M. E. Skolnik (ed.), *Radar Handbook*, 2d ed., McGraw-Hill, 1990.

Radon

A chemical element, Rn, atomic number 86. Radon is produced as a gaseous emanation from the radioactive decay of radium. The element is highly radioactive and decays by the emission of energetic alpha particles. Radon is the heaviest of the noble, or inert, gas group and thus is characterized by chemical inertness. More than 25 isotopes of radon have been identified. All isotopes are radioactive with short half-lives. *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

Radon is found in natural sources only because of its continuous replenishment from the radioactive decay of longer-lived precursors in minerals containing uranium, thorium, or actinium. ^{222}Rn (half-life 3.82 days), ^{220}Rn (thoron; half-life 55 s), and ^{219}Rn (actinon; half-life 4.0 s) occur in nature as members of the uranium (U), thorium (Th), and actinium (Ac) series, respectively. All three decay by the emission of energetic alpha particles. *See* ACTINIUM; ALPHA PARTICLES; RADIOACTIVITY; RADIUM; THORIUM; URANIUM.

Any surface exposed to ^{222}Rn becomes coated with an active deposit which consists of a group of short-lived daughter products. The radiations of this active deposit include energetic alpha particles, beta particles, and gamma rays. The ultimate decay products of radon following the rapid decay of the active deposit to lead-210 include bismuth-210, polonium-210, and, finally, stable lead-206. Radon possesses a particularly stable electronic configuration, which gives it the chemical properties characteristic of noble gas elements. It has a boiling point of $-62°$C ($-80°$F) and a melting point of $-71°$C ($-96°$F). The spectrum of radon has been extensively studied, and resembles that of the other inert gases. Radon is readily adsorbed on charcoal, silica gel, and other adsorbents, and this property can be used to separate the element from gaseous impurities.

Earl K. Hyde

The rocks and soils of the Earth's crust contain approximately 3 parts per million of ^{238}U, the long-lived head of the uranium series; 11 ppm of ^{232}Th, the head of the thorium series; but only about 0.02 ppm of ^{235}U, the long-lived member of the actinium group. The radon isotopes ^{222}Rn and ^{220}Rn are produced in proportion to the amount of the parent present. Some of the newly formed radon atoms which originate in or on the surface of mineral grains escape into the soil gas, where they are free to diffuse within the soil capillaries. Some of the radon atoms eventually find their way to the surface, where they become a part of the atmosphere. Even though thorium (^{232}Th) is generally more abundant than uranium in the Earth's crust, the probability for decay is smaller; hence, the production rate of ^{222}Rn and ^{220}Rn in the soil is roughly the same. Much of the ^{220}Rn decays before reaching the Earth's surface due to its short half-life.

When radon (^{222}Rn or ^{220}Rn) passes from soil to air, it is mixed throughout the lower atmosphere by eddy diffusion and the prevailing winds. Mean radon levels are found to be higher during those times of year when atmospheric stability is the greatest such as may occur during the fall months. Radon and its daughters play an important role in atmospheric electricity. Near the Earth's surface almost half of the ionization of the air is due to ^{220}Rn and ^{222}Rn and their daughter products. The alpha emitters from these chains typically produce about 10^7 ion pairs per second per cubic meter. *See* ATMOSPHERIC CHEMISTRY; ATMOSPHERIC ELECTRICITY.

Radon is readily soluble in water. Since ground and surface waters are in close contact with soil and rocks containing small quantities of radium, it is not surprising to find radon in public water supplies.

The radon isotopes ^{220}Rn and ^{222}Rn are used widely in the study of gaseous transport processes both in the underground environment and in the atmosphere. Radon accumulates to high levels of the order of 4000 becquerels/m^3 or more in caves unless natural or artificial ventilation occurs. Changes in ^{222}Rn concentrations in spring and well water and in soil and rocks have been suggested as a means of predicting earthquakes.

The tendency of the decay products of radon to attach to aerosols means that these nuclides will be inhaled and deposited in the bronchial epithelium and lungs. The daughter products, therefore, make up the major part of the internal radiation dose from radon. Ways of reducing radon levels within homes or workplaces include increased ventilation and sealing of major sources of entry from soil and building materials. Workers in uranium mines may encounter radon and decay product levels of the order of 50,000 Bq/m^3 or more. Ventilation procedures and special filters for the miners must be used. *See* BIOSPHERE; RADIATION INJURY (BIOLOGY); RADIOECOLOGY; RADIOISOTOPE (GEOCHEMISTRY).

M. Wilkening

Bibliography. W. Burkart et al. (eds.), *High Levels of Natural Radiation and Radon Areas: Radiation Dose and Health Effects by International Conference on High Levels of Natural Radiation and Radon*, Elsevier Science, 2002; D. R. Lide, *CRC Handbook of Chemistry and Physics*, 85th ed., CRC Press, 2004; W. W. Nazaroff and A. V. Nero (eds.), *Radon and Its Decay Products in Indoor Air*, Wiley-Interscience, 1988.

Rafflesiales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Rosidae of the class Magnoliopsida (dicotyledons). The order consists of three families and fewer than a hundred species, nearly all tropical or subtropical. The plants are highly specialized, nongreen, rootless parasites which grow from the roots of the host. They have few or solitary, rather large to very large flowers with numerous ovules and a single set of tepals that are commonly united into a conspicuous, corolloid calyx. *Rafflesia* is famous for its gigantic flowers, which in *R. arnoldii* of Sumatra are about 3 ft (1 m) in diameter. No other family has individual flowers even approaching this size. *See* MAGNOLIOPHYTA; MAGNOLIOPSIDA; PLANT KINGDOM; ROSIDAE.

Arthur Cronquist

Railroad control systems

Those devices and systems used to direct or restrain the movement of trains, cars, or locomotives on railroads, rapid-transit lines, and similar guided ground-transportation networks. Such control varies from

the use of simple solenoid valves to fully automatic electronic-electromechanical systems.

A primary function of railroad control systems is to ensure the safe movement of trains. This is generally accomplished by providing train operators and track-side operators with visual indications of equipment status. The simplest form of control consists of track-switch-position indicators combined with track-side manually operated "stop" or "proceed" signals, which the train operator follows. Advanced systems incorporate fully automated train control, subject to human supervisory control and potential intervention when faults occur in automated systems.

Automatic block signaling. Block signaling significantly improves the safety of railroad operations. Automatic block signaling is accomplished by sectionalizing the track into electrical circuits to detect the presence of other trains, engines, or cars. Logic circuits in the control system detect the locations of the trains and the positions of switches, and then set the necessary signals to inform the train operators when to stop, run slowly, or proceed at posted speeds. The control system automatically detects the presence of a leading train, selects the signal to be given, and then sets the signal indications for the following train operators to read so that they may perform accordingly. If a train is within a block, its wheels and axles form a circuit with the rails. This produces a path of very low electrical resistance, and therefore the block signal receives negligible current. This low-current condition sets the signal at "stop" for any following train. In a block that is unoccupied, all track current passes through the signal circuit and sets the signal at "clear," allowing an approaching train to proceed.

Automatic train stop. In conjunction with automatic block signals, many subway rapid-transit lines incorporate automatic trip stops along the tracks to ensure that train operators obey the stop signals. If a train passes certain restrictive signals, such as a red light indication stop, the air brakes on the train are automatically actuated by a trip arm located alongside the tracks. The train cannot be restarted until the mechanism aboard the car is reset and the train operator goes through a safety startup procedure.

Automatic cab signaling. Automatic cab signaling systems display signaling information (traditionally, permitted speeds) on board the train. Coded information is transmitted to the train, generally via the running rails (**Fig. 1**). Antennas and receivers aboard the train pick up, amplify, decode, and distribute the intelligence, which then causes the proper signal aspects to be displayed in the cab. The display is mounted near the normal line of sight of the operator. Automatic cab signaling reduces or eliminates the need for wayside signals and improves the all-weather capability of trains and the train-handling capacity of the track.

Automatic train control. Automatic train control (ATC) subsystems, located wholly on board the train, sense whether or not the train is operating within safe speed limits. If it is not, automatic train control
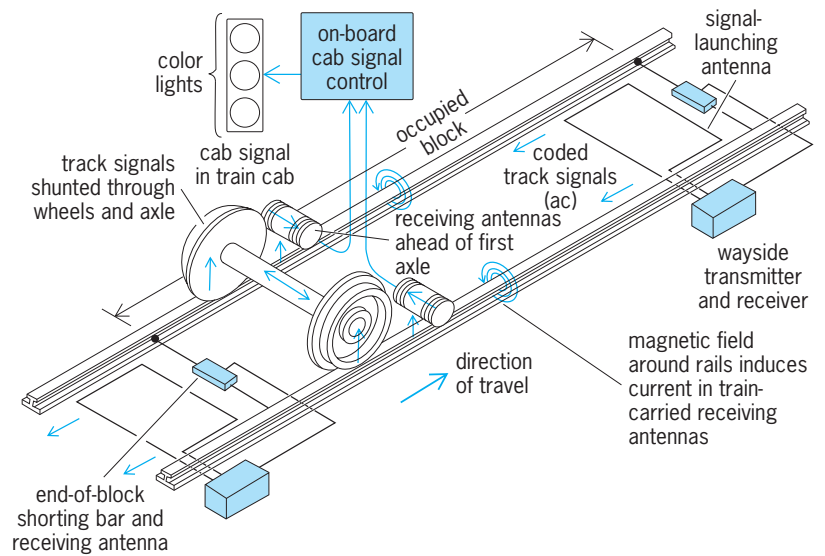
Fig. 1. Track circuit and cab signal using alternating-current track circuits.

sets the brake to bring the train to a stop or to a speed below the allowed speed.

Automatic cab signaling with automatic train control is used on many railroads and several rapid-transit lines in the United Sates and on systems in Europe and Japan.

Automatic train operation. Automatic train operation (ATO) subsystems perform nonvital operating functions such as starting, running at the prescribed speeds, slowing down, and stopping, and on some rapid-transit installations include passenger-door controls (**Fig. 2**). Automatic train operation
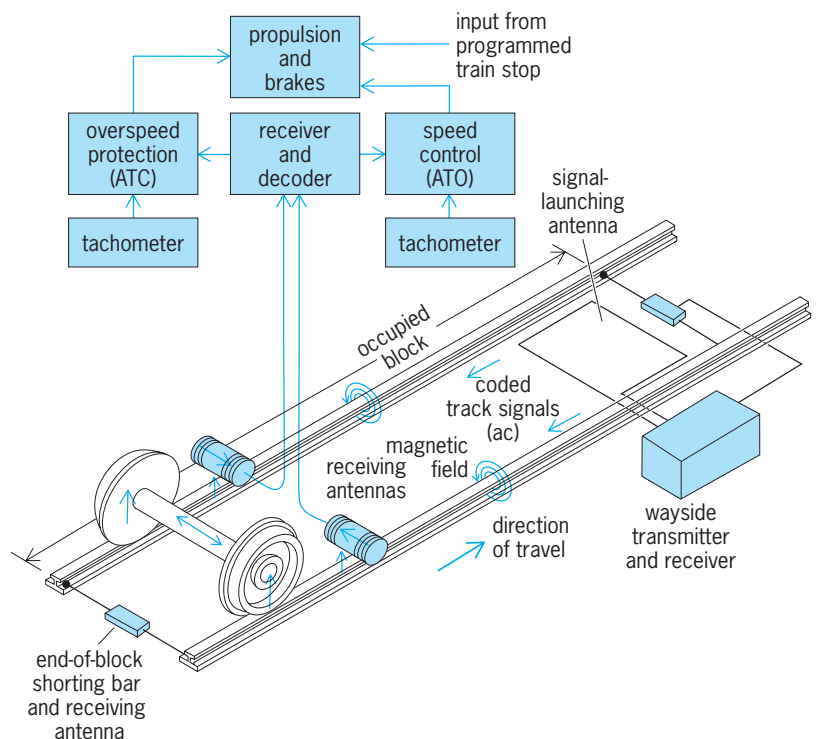
Fig. 2. Automatic train operation with fail-safe or vital overspeed protection and provision for programmed train stopping.
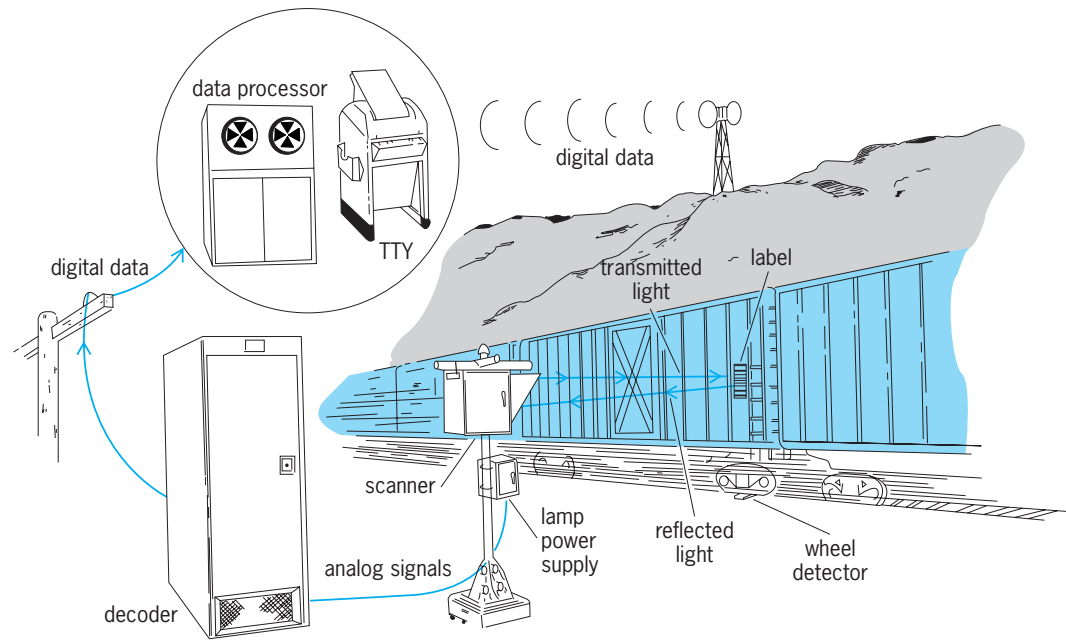
Fig. 3. Automatic car identification system.

builds upon the information transmitted to the train as part of automatic cab signaling, and is a logical next step in automating train operations.

The civil speed limits of the railroad are those imposed by the physical constraints of the railroad right-of-way, such as curves, grades, and the location of stations. Traffic speed restrictions supersede the civil speed limit when another train in the vicinity requires a lower, more restrictive speed than that allowed by the civil limit. Thus, a train approaching a slower train from behind will receive progressively more restrictive speed signals as it closes in. A speed of zero, the most restrictive speed signal, is used to stop the approaching train at a safe distance from the one ahead. In the case of cab signaling with automatic train control, the operator is responsible for obeying speed signals. With automatic train operation, the train automatically slows down, and comes to a complete stop when necessary. When automatic train operation systems are used, automatic train control (with or without cab signaling) is installed as a separate subsystem to act as a fail-safe backup for the automatic train operation.

Programmed train stop. Station stopping presents a special set of requirements for rapid transit, commuter railroads, and mainline railroad passenger operations. Accurate positioning of car doors at the station platform and smooth deceleration at relatively high rates are desirable for passenger comfort and efficient operating performance. A special subsystem of control referred to as programmed train-stop systems (or station-stop systems) are a combination of on-board and wayside electronic and electromechanical equipment that can bring a train to rest within inches of its stopping-point target.

Central line supervision. Efforts by railroads and transit operations to provide more efficient service is increasing demands on train dispatchers and sched-

ulers. These demands are being met by a number of control systems designed to provide information about the trains and cars and to perform some of the less critical functions previously assigned to dispatchers and tower or central operators.

Car identification systems are one example of central line supervision. This system scans and decodes a series of colored and patterned lines placed on the side of each car to identify an individual car (**Fig. 3**). This information is transmitted to the operations area, where a computer system is used to establish routing, determine maintenance schedules, and so on. Dispatchers and central operators can also use computer workstations to obtain information on system status.

Highly automated railroad. The use of advanced technology and automation in railroad operations is growing. A number of systems employ the control subsystems described above. The San Francisco Bay Area Rapid Transit (BART) regional system has a fully automatic train-control system as well as computerized central line supervision. BART trains can operate at 80 mi/h (128 km/h) at 90-s headways. Similar systems have been installed worldwide, including those in São Paulo, Brazil, and Washington, D.C. Japan's Tokaido Bullet Trains, running at 125 mi/h (200 km/h), are too fast for wayside signals to be read. Cab signals are used, and automatic train control prevents overspeeding.

Higher performance standards are written into specifications for each new system and system upgrade, making it clear that the trend is toward complete systemwide automatic train control. Safety, reliability of higher-speed performance, availability of equipment, return on investment, and other factors build a good case for automation.

Terminal control systems. Railroad terminals (points of origin and destination of trains) are

critical to the efficient, cost-effective operation of railroads, so they represent a major focus for automation.

A terminal generally contains three types of yards: a receiving yard, where incoming trains from the main line are temporarily stored; the hump yard, where cars are classified and resorted into new trains; and a departure yard, where trains are assembled and stored for dispatch onto the main line.

Gravity or hump yards. These have evolved as an efficient method of sorting and reforming large trains. Typical trains may contain 100–200 units.

As the train moves into the hump yard to be classified, cars or coupled assemblies of cars (called cuts) are moved to the crest of a small hill, or hump, beyond which they roll by gravity onto the appropriate classification track. The steady, one-direction movement allows a single locomotive to move three to five cars per minute.

The hump yard presents interesting control problems. The car or cut is given an initial velocity by the locomotive. Ideally, the rolling cut should couple with the next car standing on the same classification track at a speed of 2–4 mi/h (3–6 km/h). Too low a speed may prevent the couplers from engaging properly. Too high a speed can damage the lading or rolling stock.

The initial speed given to each cut must reflect the physical characteristics of the cut, the track and geometric conditions, and the distance that the cut must travel. Rolling performance of the cut is affected by a number of physical characteristics, including weight, type, and condition of the bearings, the center of gravity, and the cross-sectional area. The track and geometric conditions that impact rolling performance include curves, track condition, and the grade (slope) of the track. The distance that the car must travel on the selected classification track depends not only on the length of the track but also on the number of cars already stored there.

The speed of cuts within the hump yard is controlled primarily by the use of remotely controlled retarders which act as external breaks on the car wheels to slow the cut to a desired speed. The cuts are directed into the proper classification track by manipulation of remotely controlled track switches.

Hump yard control. Automatic control of the hump operation has many advantages. Operators can maintain a high output with moderate effort and training. More importantly, automation reduces damage claims, derailments, and misclassifications.

Advanced automatic controls for hump yards involve digital process control computers combined with automatic car-identification systems. The control system accepts input information about the cars, such as their destination and characteristics, and inputs from the yard, such as the speeds of the cars, the number of vehicles already on the classification tracks, and the positions of the switches.

The digital process control computer provides major advantages over older systems that employ automatic relay and analog computer systems. The use of digital process control techniques in hump yards opens up the entire terminal to automation, including control of switches and signals in the receiving and departure yards, as well as the handling of terminal operating paperwork and accounting functions. *See* CONTROL SYSTEMS; RAILROAD ENGINEERING; TRAFFIC-CONTROL SYSTEMS.

James Costantino; Donna C. Nelson

Bibliography. American Railway Engineering Association, *AREA Manual of Railway Engineering*, annually; American Railway Engineering Association, *The Signal Manual of American Railroads*, 4 vols., 1984; F. George, How a digital computer controls a class yard, *Railway Sig. Commun.*, 60(12):24–31, 1968: *Railway Age Mag.* pp. 50–51, September 29, 1975.

Railroad engineering

A branch of engineering concerned with the design, construction, maintenance, and operation of railways. Railway engineering includes elements of civil, mechanical, industrial, and electrical engineering. It is unique in being concerned with the interaction between moving vehicles (mechanical engineering) and infrastructure (civil engineering). The employment of both a load-supporting guideway and groups or strings of connected vehicles on flanged wheels for the transport of goods and people sets railways apart from other modes of transport.

Track geometry. Components of track geometry are alignment, curvature, and track gage.

Horizontal alignment and curvature. The plan view of a railroad track is known as the horizontal alignment. It is made up of a series of curves (arcs of simple circles), tangents (straight tracks), and spirals joining the curves and tangents. Deviations from any of the three are flaws. These imperfections are corrected periodically by a technique known as lining the track.

The radii of simple railroad curves are usually very large, so that the centers of the circles of which the curves are an arc are frequently inaccessible from the track. For convenience, measurements of curve severity and regularity are made in terms of midchord offsets (**Fig. 1**). In North America, curves are described in terms of degrees. The degree of curve is the number of degrees of central angle subtended by a 100-ft chord. It is also equal to the midchord offset, in inches, of a 62-ft chord. Outside of North America and on some transit properties, curves are described in terms of their radii.

On mainline tracks the outer rail of curves is superelevated above the inner rail to compensate for centrifugal force. The amount of superelevation depends on train speeds and the degree of curvature.

Vertical alignment. The side or elevation view of track, composed of a series of straight portions and the vertical curves joining them, is known as the vertical alignment. The vertical change in elevation, in feet, over a horizontal distance of 100 ft is the percent grade. Because the friction coefficient of steel wheels on steel rails is low, railroad grades must also be low, with values from zero to 1.5% fairly common.
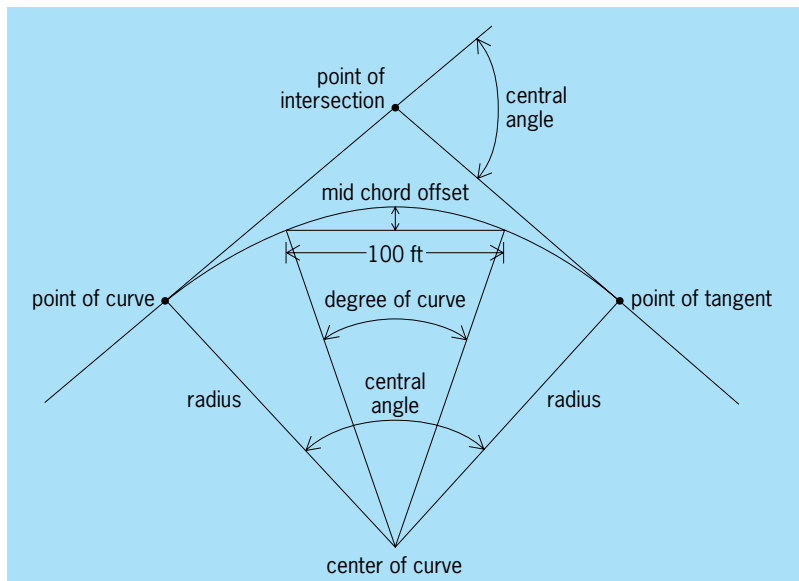
Fig. 1. Simple railroad curve without spirals.

Two-percent grades are severe, usually requiring helper locomotives. Grades that are more severe, up to about 4%, can be surmounted only with considerable extra operating care and at significant additional expense. The macro view of elevation is the profile of the railroad, while the micro view is the surface. Irregularities on the surface are corrected by raising the low spots and packing ballast under the raised track.

Gage. Track gage is the distance between rails. The standard gage throughout the world is 4 ft 8.5 in. Narrow gages of 3 ft and broad gages of 5 and 6 ft have all been tried at various times and in different places. The reasons are diverse and include attempts to reduce construction cost, to intentionally make rolling stock of other countries incompatible, and to hinder attempts at invasion, as in Russia where broad gage prevails.

Track components. These include rails, welded rails, crossties, and ballast.

Rails. The function of rail is to guide wheels and distribute their vertical and lateral loads over a wider area. Neither cast nor wrought iron was ideally suited to this task. The development of the Bessemer furnace to produce steel that was three times harder than wrought iron at reasonable cost made it possible for the weight of vehicles and therefore the productivity of railways to increase. Unfortunately Bessemer steel hardness was only half that of cast iron, causing more frequent renewal, especially on heavy curves. Nonetheless, steel was so vastly superior overall that it quickly supplanted its predecessors. *See* CAST IRON; STEEL; STEEL MANUFACTURE; WROUGHT IRON.

For a given shape the moment of inertia of rail (and its resistance to bending) is proportional to its cross-sectional area. Rail size was rapidly increased from 30 or 50 lb per yard of length (15 or 50 kg/m), to a maximum of 155 lb/yd (77 kg/m) as engineers attempted to reduce the cost of maintaining proper vertical and horizontal alignment of track subject to heavier loads and higher speeds. *See* MOMENT OF INERTIA.

In attempting to find the optimum distribution of material to resist deflection, wear and service failures, and to facilitate economic manufacture, the shape and contour of rail sections have undergone many adjustments.

As competitive pressure heightened, freight-hauling railroads in the United States tended to increase the proportion of 100-ton (90-metric-ton) carrying capacity cars. Some special-purpose noninterchangeable cars, with even higher axle loads, have become common as well. Speeds are also increasing in both passenger and freight traffic. This has led North American railroads to search for stronger and harder steel. Fine-grain pearlitic steel (high-strength steel) with tensile strengths of approximately 150,000 lb/in.^{2} (10 gigapascals) and hardness in the 340–388 Brinell range is being produced to meet this need. *See* HARDNESS SCALES.

Welded rails. Rails are joined end to end by butt welding, whereby continuous rails of over 1000 ft (300 m) in length can be produced. When laid in track, the rails are heavily anchored to restrain movement due to temperature changes. The stresses produced by temperature changes are well below the strength of the steel.

Crossties. Crossties play important roles in the distribution of wheel loads vertically, longitudinally, and laterally. Each tie must withstand loads up to one-half that imposed on the rail by a wheel. The crosstie must then distribute that load to the ballast surrounding it.

Timber crossties, varying in section from 6 in. × 6 in. (15 cm × 15 cm) for the lightest applications to 7 in. × 9 in. (18 cm × 23 cm) for heavy-duty track and in length from 8 ft 6 in. (2.6 m) to 9 ft (2.7 m) in length, are sawed in highly automated saw mills. Ties may be prebored for fastenings; and to prevent splitting, irons, dowels, or nailplates are frequently applied. Air drying or artificial seasoning is necessary before pressure treatment with preservatives, most often creosote–coal tar or creosote–petroleum mixtures. To protect timber crossties from crushing, it is standard practice to employ steel tie plates between rails and ties to further spread train loads beyond that accomplished by the wide rail base.

Well-treated hardwood ties such as red oak, locust, beech, hickory, and fir in well-maintained track may be expected to last 30 years or more. Timber crossties become unserviceable after time because of splitting, decay, insect attack, center cracking, mechanical wear, and crushing. Tie life varies with species, weather, traffic, track condition, and geometry.

In North America, crossties are replaced selectively on a periodic basis. At certain intervals, all the crossties that are considered badly deteriorated, but only those, are renewed. In other parts of the world, all the crossties in a given length of track are replaced along with the rail and ballast, and those components considered to have substantial remaining life are

recycled by installing them in lighter traffic lines, yard tracks, or sidings.

Prestressed concrete monoblock crossties are standard in Great Britain and parts of continental Europe. In North America, they are less extensively used than timber. Concrete crossties are used extensively in the high-speed tracks of Amtrak's Northeast Corridor. Several western railroads employ them, especially in mountainous terrain where heavy curvature requires frequent rail renewal or transposition, the practice of interchanging the inner and outside rail of curves to equalize wear. Rail fasteners used with concrete ties may be removed and reinstalled easily without damage to the crosstie. Tests conducted by the Association of American Railroads (AAR) have shown a significant increase in track stiffness in concrete tie track over timber tie track of otherwise similar construction and material.

Ballast. The granular material that supports crossties vertically and restrains them laterally is known as ballast. Ideal ballast is made up of hard, sharp, angular interlocking pieces that drain well and yet permit adjustments to vertical and horizontal alignment. Materials that crush and abrade, creating fines that block drainage or that cement, should not be used. Soft limestones and gravel, including rounded stones, are examples of poor ballast, while crushed granite, trap rock, and hard slags are superior.

The ballast must distribute crosstie load over an area sufficiently wide so that the underlying subsoil will not be overstressed. Tests conducted by the AAR during 1988–1991 showed that a 31,500-lb (14,200-kg) wheel load generates an average pressure of 37 lb/in.^{2} (260,000 Pa) at the bottom of a tie and that 18 in. (45 cm) of ballast spreads this over sufficient area to reduce its magnitude to about 10 lb/in.^{2} (70,000 Pa). Depending on the nature of the subsoil a depth of 12 to 18 in. (30 to 48 cm) beneath the bottom of the tie is usually adequate. A subballast of a fine-grain, well-draining material such as cinder or sand may be used beneath the stone or slag top ballast.

Track stiffness. A loaded wheel depresses the track it rolls on, and lifts the track ahead and behind. The depression is greatest directly below the wheel. Thus, as the wheel moves it creates a wave action in the track. A characteristic of track representing its stiffness is the modulus of elastic track support, frequently called the track modulus. This quantity is determined mathematically and is used to calculate track depression. Track modulus is more significant than rail stiffness in resistance to track depression under load. Because a rolling wheel is always climbing out of its own depression, track stiffness is an important contributor to train resistance and therefore energy conservation. There is a limit to the amount of stiffness desirable, for as stiffness increases, the attenuation of dynamic load decreases, as well as the spreading of wheel load. A value of track modulus approaching 10,000 lb/(in.)(in.) [70,000 newtons/(m)(m)] is considered close to ideal, given current suspension designs and track geometries.

Train resistance. Railroads were developed to reduce the effort required to move people and goods. However, a train of freight or passenger cars does resist movement. This resistance is the result of a combination of factors.

The most critical factor is gravity. Locating engineers seek the lowest possible grade consistent with the intended route. Practices such as following river valleys and stream beds, and adding length to reduce the percent grade in mountainous territory are common. Grade resistance, estimated at about 20 lb/ton (10 kg/metric ton) of train weight for each percent of grade, means a 10,000-ton (9,000-metric ton) train requires 200,000 lb (90,000 kg) of force just to overcome gravity on a 2% grade.

The second factor is the track deformation. Other factors may be grouped into the category of friction. There is the friction of axles turning in their journals, and of wheel flanges rubbing the sides of rails. The conical shape of wheels is intended to make them roll straight with respect to the track and to be self-correcting when they veer away from that path. However, irregularities of track alignment and wear of wheels and curves make following a true course without flange contact impossible. At certain speeds the wheels follow a zigzag path, striking first one rail and then the other. This is called truck hunting, and it consumes energy.

As a train negotiates curved track, the flanges bear against the outer rail, causing wear of both rails and flanges and consuming additional energy estimated at about 8 lb/ton (4 kg/metric ton) per degree. This form of resistance is sufficiently severe that when sharp curves occur on grades the grade is frequently reduced throughout the curve to compensate for the added resistance. Such grades are known as compensated grades.

Many of the friction factors vary directly with speed. Aerodynamic resistance, the last factor, varies with the square of speed. Up to speeds of 40–50 mi/h (24–30 km/h) it is not critical, but at speeds above these values it assumes considerable importance.

Locomotives. Force must be applied to the leading coupler of a train to overcome the combination of resistances. To supply this force, locomotives convert some form of potential energy to produce useful work.

A locomotive's performance, measured in a number of ways, begins with its tractive effort, and in particular, its starting tractive effort (the force it can apply to a standing train). The locomotive's starting tractive effort is usually equated to its weight on the driving wheels times the coefficient of friction of those wheels on steel rail. This coefficient can vary from about 0.15 to about 0.30, depending on the rail surface condition. The most modern locomotives employ electronic slip detection and control to enhance their starting tractive effort, and almost all locomotives use sand when necessary to increase the coefficient of friction.

Once a locomotive is moving, the tractive force that it can apply to its train once the train is moving is a function of speed, usually decreasing as speed

increases. Tractive force varies with speed for several types of locomotives. The locomotive's power is the product of its tractive force at a given speed and that speed. The difference between the locomotive's power and the train's resistance is the power available for acceleration or for overcoming additional resistance. Generally, train resistance increases with speed, especially when aerodynamic resistance is a factor. Therefore, a speed known as the balancing speed is attained where the locomotive power equals train resistance, and the entire power capability of the locomotive is consumed just to maintain speed. Any increase in resistance will slow the train until a new balance is reached.

Other significant factors in rating locomotive performance are thermal efficiency, reliability, and availability (time available for duty versus time in shops for repair or service). For all of these, and especially availability, the diesel-electric locomotives, which first appeared in the late 1930s, surpassed steam. In addition, since diesel-electric power does not require the proliferation of expensive servicing facilities such as coaling and watering stations, by the early 1960s the transition to diesel was complete.

But even diesel-electric locomotives have limitations. The exhaust gases containing oxides of nitrogen and sulfur have become environmental concerns. A longstanding difficulty has been the expense of maintaining diesel-electric traction motors and especially their commutators. *See* COMMUTATOR; DIESEL ENGINE.

Electric traction motors exert their maximum torque at starting and slow speeds. This characteristic is desirable for railroad applications, since it is in starting that train resistance is greatest (until aerodynamic resistance takes over). Tractive effort curves showing tractive force versus speed are useful in rating locomotives for various types of service and for developing computer-generated simulations of train performance. By matching tractive effort curves against those for train resistance, these programs predict elapsed time, braking distance, and fuel consumption.

The earliest diesel-electric switching locomotives developed about 600 horsepower and road freight units 1350 hp. Single locomotive units of 4000 hp are common. Common practice since dieselization began has been to employ a number of locomotive units coupled together to form a single, more powerful power source.

Locomotive development includes designs using liquefied natural gas as a fuel to reduce environmental pollutants. Locomotives equipped with the necessary power conditioning equipment and squirrel-cage asynchronous motors, made possible by the advent of high-capacity solid-state electronics, exhibit superior adhesion and have no troublesome commutators. They are adept at hauling heavy-tonnage mineral freight, fast passenger trains, or high-speed merchandise trains (freight trains that haul primarily high-value merchandise, as opposed to low-cost raw materials such as coal or grain) [**Fig. 2**]. *See* LOCOMOTIVE.

Freight services. Until World War II, several general-purpose cars such as open hoppers, open gondolas, closed box, and flat cars, which any shipper could load with any commodity, met most needs. However, since then a number of special-purpose cars were designed which are frequently assigned to or even owned by individual shippers or third parties, such as open and covered hopper cars for coal or grain, insulated and refrigerator cars for fresh food and beverages, and bi- and trilevel cars for automotive products.

The high level of productivity of railroads in America largely has been made possible by the semi-automatic coupler, invented to replace the dangerous link-and-pin system, and the Westinghouse air brake. The formation of the Master Carbuilders' Association, a forerunner of the AAR, set standards for railway cars and permitted them to move freely on all lines, thereby obviating the need to transfer lading from one car to another at railroad junctions. The AAR administers the interchange agreement under which cars and car components are tested and certified for free interchange.

Air brakes permitted much longer trains to move safely and eliminated the need for brake personnel to leap from car to car to set manual brakes. Janney couplers facilitated the making up and breaking of long trains. They also introduced slack of about 6 in. (15 cm) per car so that long trains could be started by steam locomotives by overcoming starting resistance one car at a time. *See* AIR BRAKE.

Both of these attributes have come to be viewed as mixed blessings. Free slack together with the delay in propagating pneumatic signals to actuate brakes in very long trains can lead to longitudinal train action and dynamic shock loads. The use of semipermanent solid drawbars, which eliminate longitudinal slack, is preferred in solid trains where all cars have a single origin and destination. Such trains have increased, but there is a tendency for other trains to bypass as many switching yards as is feasible.

The fastest-growing segment of freight business in North America is intermodal traffic: highway trailers mounted on flat cars and containers (both domestic and overseas) mounted either on flat cars or specially

Fig. 2. Locomotive with ac asynchronous traction motors. (*General Motors*)

Fig. 3. Articulated double-stack container train, showing single four-wheel truck under one end each of two adjacent cars, and containers stacked two high. (*Gunderson, Inc.*)

designed container cars. Many of these container cars are actually a series of cars permanently connected by articulated joints in which a single truck supports one end each of two adjacent cars. The platforms of these cars have depressed centers or wells, so that two containers can be stacked on top of one another without violating certain height restrictions (**Fig. 3**).

Passenger services. Rail passenger systems such as the Shinkansen in Japan (1964); TGV (Très Grande Vitesse; 1981) in France; ICE (Inter City Eisenbahn; 1991) in Germany; X-2000 in Sweden (1990); or Britain's several High-Speed Intercity Trains are notable for speed and convenience, as well as magnetic levitation. These systems have developed in context of an awareness of deteriorating highway infrastructures, serious concern with the air pollution generated by automobiles and trucks, increasing traffic congestion in urban areas, and worries over petroleum usage and supply. *See* AIR POLLUTION; MAGNETIC LEVITATION.

Therefore questions have arisen concerning the possibility of high-speed rail passenger service in the United States. The answerers are many, complex, and relate only indirectly to technology. Using the accepted criteria of 125 mi/h (200 km/h) to qualify as high speed, there are no less than 18 such weekday Metroliners operated between New York and Washington. These trains are powered by 5300-continuous-horsepower class AEM-7 rectifier electric locomotives capable of 7600 hp for short periods. The 4000-hp Genesis diesel-electric locomotives (**Fig. 4**) are also in use. In addition, Amtrak successfully tested in revenue service both the Swedish X-2000 tilting train and the German ICE at speeds up to 140 mi/h (220 km/h).

The lack of extensive services in the United States outside the Northeast Corridor can be partly explained by demographic factors. Population centers in both Japan and Europe are typically closer to each other than in North America, making construction costs for dedicated high-speed trackage more feasible and travel times far more competitive with air travel. In addition, high-speed passenger trains require significantly different track configurations (for example, curve superelevation and turnout designs) than do slower freight trains, even those high-valued merchandise or intermodal highway trailers and containers, which frequently travel at speeds of 70 mi/h (110 km/h). These engineering differences are impractical on lines primarily moving heavy mineral freight, where axle loadings and speeds differ even more radically from high-speed passenger trains. The United States freight lines are also frequently burdened with highway grade crossings that are incompatible with passenger trains traveling at speeds well over 100 mi/h (160 km/h). In spite of all these difficulties, advances in the caliber of passenger service continue at a rapid pace in and outside the Northeast Corridor, particularly in the area of suburban commuter rail.

Environmental issues. Railroads are considered to be three to four times more energy efficient than other forms of land transport, with lower air contaminates per ton-mile for diesel-electric powered trains. The goal is to reduce the level of the oxides of nitrogen and sulfur even further. Railroads are the one form of heavy-freight transport that can be electrified by using available technology.

One of the more serious environmental concerns of railroads is posed by potential spills and leaks of hazardous materials. As a result of engineering studies, many steps such as head shields and special nonoverriding couplers on tank cars have already been taken, and further studies are under way to improve the integrity of these cars. *See* RAILROAD CONTROL SYSTEMS; TRANSPORTATION ENGINEERING.

George H. Way

Bibliography. American Railway Engineering Association, *Manual of Recommended Practice*, 1993, supplemented annually; American Railway Engineering Association, *Stresses in Railroad Track— The Talbot Reports*, 1980; H. I. Andrews, *Railway Traction*, 1986; J. H. Armstrong, *The Railroad, What It Is, What It Does*, 3d ed., 1990; W. W. Hay, *Railroad Engineering*, 2d ed., 1982; V. A. Profillidis, *Railway Engineering*, 2d ed., 2000; *Railway Age Magazine*, pp. 48–53, 55–77, January 1994.

Fig. 4. The genesis, 4000-hp passenger locomotive capable of speeds up to 103 mi/h (165 km/h). (*Amtrak*)

Rain shadow

An area of diminished precipitation on the lee side of mountains. There are marked rain shadows, for example, east of the coastal ranges of Washington, Oregon, and California, and over a larger region, much of it arid, east of the Cascade Range and Sierra Nevadas. Precipitation on the northern Oregon coast is around 100 in. (2.5 m) per year; east of the coastal ranges in the Willamette Valley it is about 40 in. (1.0 m); and east of the Cascades it drops to around 10 in. (0.25 m). All mountains decrease precipitation on their lee; but rain shadows, as is shown by annual totals, are sometimes not marked if moist air often comes from different directions, as in the Appalachian region.

The causes of rain shadow are (1) precipitation of much of the moisture when air is forced upward on the windward side of the mountains, (2) deflection or damming of moist air flow, and (3) downward flow on the lee slopes, which warms the air and lowers its relative humidity. *See* CHINOOK; PRECIPITATION (METEOROLOGY).   J. R. Fulks

Rainbow

An optical effect of the sky formed by sunlight falling on the spherical droplets of water associated with a rain shower. The circular arc of colors in the rainbow is seen on the side of the sky away from the Sun. The bright, primary rainbow shows the spectrum of colors running from red, on the outside of the bow, to blue on the inside. Sometimes a fainter, secondary bow is seen outside the primary bow with the colors reversed from their order in the primary bow. The shape of each bow is that of a circle, centered on the antisolar point, a point in the direction exactly opposite to that of the Sun, which is marked by the shadow of the observer's head (see **illus.**).

As a light ray from the Sun strikes the surface of a water drop, some light is reflected and some passes through the surface into the drop. The primary bow results from light that enters the drop, reflects once inside the drop, and then leaves the drop headed toward the observer's eye. Light that is reflected twice inside the drop produces the sec-
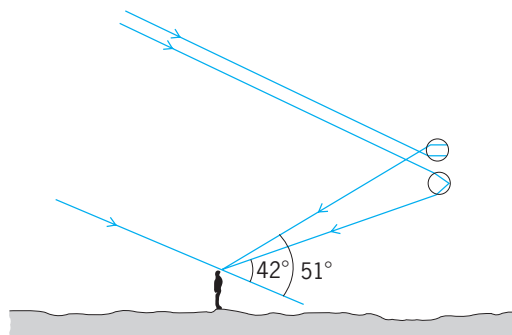
Raindrops contribute to the primary rainbow (with one internal reflection) and to the secondary rainbow (with two internal reflections). (*After R. Greenler, Rainbows, Halos and Glories, Cambridge University Press, 1980*)

ondary bow. The change of direction that occurs when a light ray enters or leaves the waterdrop (refraction) is different for the different colors that make up white sunlight. As a result, the size of the circle is different for each color, thereby separating the colors into the rainbow sequence. *See* METEOROLOGICAL OPTICS.   Robert Greenler

Rainforest

Forests that occur in continually wet climates with no dry season. There are relatively small areas of temperate rainforests in the Americas and Australasia, but most occur in the tropics and subtropics.

Occurrence. The most extensive tropical rainforests are in the Americas. These were originally 1.54×10^6 mi^2 (4×10^6 km^2) in extent, about half the global total, and mainly in the Amazon basin, extending north to the Orinoco basin and the Guyanas. A narrow belt also occurs along the Atlantic coast of Brazil, and a third block lies on the Pacific coast of South America, extending from northern Peru to southern Mexico.

The eastern tropical rainforests (970,000 mi^2 or 2.5×10^6 km^2) represent the second most extensive block. These are centered on the Malay archipelago, extending toward Queensland (Australia), the western Pacific islands, and continental Asia; west to the western Ghats of India; and north to Xishuangbanna in southernmost China (see **illus.**).

Aerial view of a typical rainforest in Windsor tableland, Queensland, Australia.

The African tropical rainforests, originally 700,000 mi^2 (1.8×10^6 km^2), have the least extent. They are situated mainly in the Zaire basin, with outliers in east Africa, and extending along the coast of west Africa, where they have now largely been cleared for farms.

Characteristics. Tropical rainforests have a continuous canopy (commonly 100–120 ft or 30–36 m tall) above which stand huge emergent trees, reaching 200 ft (60 m) or even taller. Within the rainforest canopy are trees of many different sizes, including pygmies, that reach only a few feet. Trees are the main life form and are often, for purposes of description and analysis, divided into strata or layers. Trees form the framework of the forest and support

an abundance of climbers, orchids, and other epiphytes, adapted to the microclimatic conditions of the different zones of the canopy, from shade lovers in the gloomy, humid lower levels, to sun lovers in the brightly lit, hotter, and drier upper levels. Many trees have buttresses, and some bear their flowers directly on the trunk (cauliflory). Many leaves, especially in the lower part, have long extended drip tips. Most trees have evergreen leaves, many of which are pinnate or palmate. These features of forest structure and appearance are found throughout the world's lowland tropical rainforests.

There are other equally distinctive kinds of rainforest in the lower and upper parts of perhumid tropical mountains, and additional types on wetlands that occur either as permanent or periodic swamp forests. Peat swamp rainforests can be found in Asia and in smaller areas elsewhere. These forests and others that exist on limestone and ultramafic rocks form a total of about a dozen distinct types, commonly called formations. Each type is associated with a particular site and is usually, therefore, sharply bounded.

Climate. Rainforests occur where the monthly rainfall exceeds 4 in. (100 mm) for 9–12 months. They merge into other seasonal or monsoon forests where there is a stronger dry season (3 months or more with 2.5 in. or 60 mm of rainfall). The annual mean temperature in the lowlands is approximately 64°F (18°C). There is no season unfavorable for growth.

Primary and secondary communities. Rainforests, like other vegetation, are dynamic; the canopy is in a continual state of flux. As trees die, singly or in groups, gaps develop and new trees grow from seedlings to fill them. Thus, the forest consists of a mosaic of patches at different stages of a growth cycle, and any layering (stratification) of tree crowns is continually changing as trees mature and eventually die. Most species have seedlings that can grow in the shade on the forest floor, filling in the smaller gaps of the canopy. Thus, the species maintain themselves from generation to generation, and although different species may fill any given gap, replacing what was there before, overall the canopy composition remains the same. Such species are often called climax species.

Where a section of rainforest is cleared, or if a natural catastrophe (such as a cyclone, landslide, or volcanic mudflow) occurs, much larger gaps in the canopy can result. However, these gaps are colonized by a different set of tree species that grow up as a single canopy layer from seed after the creation of the gap. These species are adapted to the hotter, drier, brighter conditions of large gaps, and are commonly called pioneers, and they form secondary forest. Their seedlings cannot establish in shade, and pioneers are replaced as they die by climax species, thereby restoring the original or primary forest. The complete succession may take several hundred years.

Primary rainforests are exceedingly rich in species of both plants and animals. There are usually over 100 species of trees 2.5 in. (10 cm) in diameter or bigger per 2.4 acres (1 ha). The richest forests yet known (in Kalimantan, Peru and Ecuador) have over 250 species. There are also numerous species of climbers and epiphytes.

In addition, flowering and fruiting occur year-round, but commonly there is a peak season; animal breeding may be linked to this. Secondary rainforests are much simpler. There are fewer tree species, less variety from location to location, and fewer epiphytes and climbers; the animals are also somewhat different.

Total clearance of rainforest for pasture or for growing food crops is followed by species-poor, structurally simple secondary forest when habitation ceases. However, extraction of timber causes less severe damage by creating gaps in the canopy. Damage can be minimized, and the forest can recover, remaining a rich mixture of climax species. Where damage is heavier and large canopy gaps are formed, secondary forest fills them. Thus, when logging-induced disturbances mimic natural gap creation, less alteration of forest species composition and structure occurs. *See* ECOLOGICAL SUCCESSION.

Useful products. Tropical rainforests are a source of resins, dyes, drugs, latex, wild meat, honey, rattan canes, and innumerable other products essential to rural life and trade. Modern technology for extraction and for processing has given timber of numerous species monetary value, and timber has come to eclipse other forest products in importance. The industrial nations use much tropical hardwood for furniture, construction, and plywood. Rainforest timbers, however, represent only 11% (1×10^{10} ft^3 or 284×10^6 m^3) of world annual industrial wood usage, a proportion that has doubled since 1950. West Africa was the first main modern source, but by the 1960s was eclipsed by Asia, where Indonesia and Malaysia are the main producers of internationally traded tropical hardwoods. Substantial logging has also developed in the neotropics. For example, since the 1980s parts of the Brazilian Amazon have been heavily exploited for the south Brazil market; and there is also an export trade in mahogany. In Asia, second to timber in value are the rattans, climbing palms largely confined to these eastern rainforests. *See* FOREST TIMBER RESOURCES.

Tropical deforestation and climatic change. Tropical deforestation is the source of about one-quarter of the annual increase of atmospheric carbon dioxide, or about 7300 million tons. This net flux to the atmosphere is the sum of carbon dioxide from burning and decay of residues, oxidation of soil humus, and slow oxidation of harvested products, offset by carbon fixation in vegetation that regrows on the site. (Three-quarters of annual carbon dioxide increase, however, comes from burning fossil fuel.) In addition to an increase in the production of carbon dioxide, deforestation adds methane and nitrous oxide to the atmosphere. These three gases together contribute 75% of the annual radiative forcing, the greenhouse effect [the rest comes from synthetic chloro (fluoro) carbons], and 20–25% of the greenhouse effect is attributable to deforestation. *See* GREENHOUSE EFFECT.

Conservation. Tropical forests as a whole are estimated to have disappeared at a rate of 0.8% or 37×10^6 acres (15.4×10^6 ha) per year during the 1980s, and much of this was rainforests. Perhaps half as much rainforest has been logged but left to regenerate. For this reason, there is worldwide concern regarding the destruction of rainforests. Apart from esthetic and moral issues, numerous potentially valuable species are being lost or threatened. Countries with rainforests have set aside 5–10% of this forest for national parks; however, not all are ideally sited to be adequately representative. Given the high species richness, and the very sparse occurrence of some species (for example, carnivorous animals at the top of the food chain), the national park network is inadequate. Another approach has been to skillfully conduct logging and extraction procedures of other useful products in a manner that maintains the forest ecosystems and plant and animal habitats. *See* CONSERVATION OF RESOURCES; FOREST ECOSYSTEM.

T. C. Whitmore

Bibliography. L. A. Bruijnzeel, *Hydrology of Moist Tropical Forests and Effects of Conversion: A State of Knowledge Review*, 1990; A. Gomez-Pompa, T. C. Whitmore, and M. Hadley (eds.), *Rain Forest Regeneration and Management*, 1990; M. D. Swaine and T. C. Whitmore, On the definition of ecological species groups in tropical rain forests, *Vegetatio*, 75:81–86, 1988; J. Terborgh, *Diversity and the Tropical Rain Forest*, 1992; T. C. Whitmore, *An Introduction to Tropical Rain Forests*, 2d ed., 1998; T. C. Whitmore, *Tropical Rain Forests of the Far East*, 2d ed., 1984.

Rajiformes

One of four orders making up the Batoidei. Aside from the few exceptions noted below (mostly in family Rajidae), the combination of characters that distinguish the order from other batoids are a single unbranched rostral process on front of the cranium, extending to the tip of the snout; antorbital (in front of the eye socket) cartilage that does not extend forward to help support the anterior part of the disc; two dorsal fins on the tail; lack of spines on the tail; lack of electric organs; a snout that is not a long flat blade; a head lacking anterior finlike expansions of pectoral fins; and ovoviviparous reproduction. Their food consists of various crustaceans, mollusks, and small fishes.

Order Rajiformes
    Rhinidae (genus *Rhina*)
    Rhynchobatidae (*Rhynchobatus*)
    Rhinobatidae (*Aptychotrema, Rhinobatos, Trygonorrhina, Zapteryx*)
    Rajidae (18 genera)

Rhynchobatidae. The general shape of these batoids is intermediate to typical sharks and typical rays. They are distinguished by a moderately flattened body, and narrow, pointed snout with the length in front of the eyes about equal to the width of the head at the eyes. The caudal fin is bilobed, with both lobes sharp-pointed. The dorsal fins are relatively large; the origin of the first is about over midpoint of the pelvic fins. The rear margin of pectoral fins do not extend posteriorly to the origin of the pelvic fins. The back is sometimes variously marked by small white spots. Their habitat is shallow inshore waters and estuaries.

The family consists of one genus, *Rhynchobatus*, and four species, all of which occur in the western Indian Ocean from the Red Sea to South Africa and east to Malaysia and the Philippines and north to Japan.

Rhinidae. This family, comprised of only one species, *Rhina ancylostoma*, occurs in the tropical Indian Ocean to Malaysia, Southeast Asia, and Australia. It is similar to species of Rhynchobatidae, but differs in having a short, broadly rounded snout, its length in front of the eyes being less than the width of the head at the eyes; and large spines on the head and shoulders. It inhabits coastal areas over substrates of mud and sand, as well as coral reefs.

Rhinobatidae (guitarfishes). In body form, these batoids are between sawfishes and skates, and they vary from a relative flat wedge-shaped body sector (for example, *Aptychotrema* and *Rhinobatos*, see **illustration**) to a broadly rounded body sector and stout tail sector (for example *Zapteryx*) to an almost circular body sector and long narrow tail sector (that is, banjo shaped, such as *Trygonorrhina*). The origin
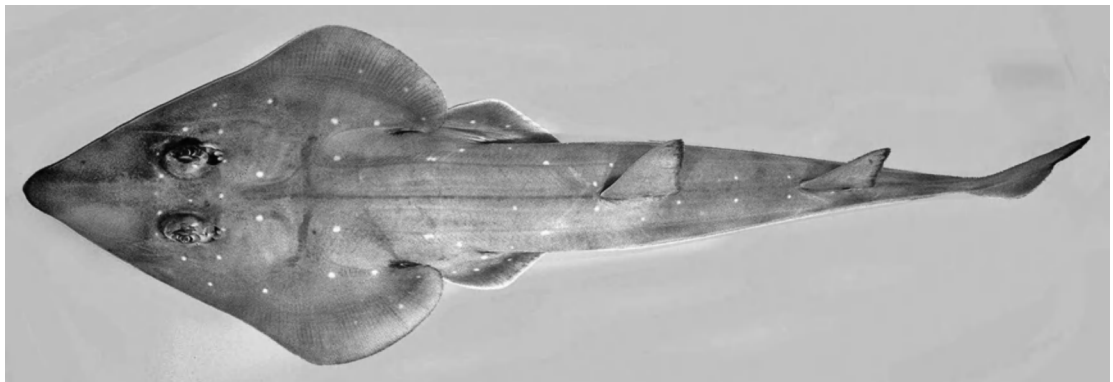
Spotted guitarfish (*Rhinobatos punctifer*). (*Photo* © *John E. Randall*)

of the first dorsal fin is behind the pelvic fins; the caudal fin is without a definite lower lobe; and the rear margin of the pectoral fins extends posteriorly to the origin of the pelvic fins, or beyond. The numerically largest genus, *Rhinobatos*, contains about 36 species, accounting for about 75% of the family.

Guitarfishes swim near the bottom, where they use their robust tail sector for propulsion and their pectoral fins as elevator and rudder. They often lie half buried in the sand or mud with their eyes and spiracles exposed. Occurring primarily in tropical and warm temperate coastal waters of all oceans, they rarely enter estuaries or freshwater.

Rajidae (skates). In body form, skates are more raylike than sharklike. The head and body are strongly depressed; the snout is moderately slender; the pectoral rays are united with the sides of the head, to very near the tip of the snout; the outer margin of the pelvic fins is concave or notched (so deeply notched in *Cruriraja* and *Anacanthobatis* that the upper lobes appear to be little legs); the tail sector is sharply distinct from the body sector; there is a single unbranched rostral process reaching near the tip of the snout (falling short in *Breviraja*, and absent in *Psammobatis*, *Sympterygia*, *Malacoraja*, and *Anacanthobatis*); there are two small dorsal fins near the tip of the tail (one in *Arhynchobatis*, none in *Anacanthobatis*); the caudal fin is reduced to a small membranous fold or is completely absent in some adults; and the upper surface of the disc and tail is variously rough with prickles and thorns (though smooth in *Anacanthobatis*). *Raja* is the typical genus and has the largest number of species; however, there is one genus in addition to those noted above that depart radically from the typical. *Dactylobatus armatus* is unique among batoids in having six or eight of its pectoral rays, at midlength of its pancake-shaped disc, extending beyond the edge as a tongue-shaped lobe.

The family consists of 18 genera and about 200 species. It is represented in all seas of the world, from the tropics to subpolar waters, from shallow depths to at least 1000 m (3300 ft). Like other rajiforms, skates are bottom feeders, relying primarily on their pectoral fins for locomotion. *See* BATOIDEA.

Herbert Bosehung

Bibliography. H. B. Bigelow and W. C. Schroeder, Sawfishes, guitarfishes, skates and rays, in Fishes of the Western North Atlantic, *Mem. Sears Found. Mar. Res. Mem.*, 1(pt. 2):1–514, 1953; J. D. McEachran and T. Miyake, Phylogenetic interrelationships of skates: A working hypothesis (Chondrichthyes, Rahjoidei), pp. 285–304, in H. L. Pratt, Jr., et.al. (eds.), Elasmobranchs as living resources: Advances in the biology, ecology, systematics, and the status of the fisheries, *NOAA Tech. Rep. 90*, 1990; J. S. Nelson, *Fishes of the World*, 3d ed., Wiley, New York, 1994.

Raman effect

A phenomenon observed in the scattering of light as it interacts with a material medium in which the incident light suffers a change in frequency due to internal energy change of the molecular scatterers. Raman scattering differs in this respect from Rayleigh scattering in which the incident and scattered light have the same frequency. Shifts in frequency are determined by the type of molecules in the scattering medium, and spectral analysis of the scattered light can provide a "fingerprint" of the chemical structure of the scatterers. Both incoherent and coherent forms of the Raman effect exist. Spontaneous Raman scattering, the usual (incoherent) form, is very weak, with Raman signals 4–5 orders of magnitude smaller than Rayleigh scattering and about 14 orders of magnitude smaller than fluorescence. In stimulated Raman scattering, a coherent form, the signals may be quite large. In addition, there are nonlinear forms of Raman scattering, including hyper-Raman scattering and coherent anti-Stokes Raman scattering (CARS). *See* SCATTERING OF ELECTROMAGNETIC RADIATION.

Discovery. Raman scattering was experimentally observed in 1928 by the Indian physicists C. V. Raman and K. S. Krishnan, who examined sunlight scattered by a number of liquids. With the help of complementary filters, they found frequencies in the scattered light that were lower than the frequencies in the filtered sunlight. By using light of a single frequency from a mercury arc, they showed that the new frequencies were characteristic of the scattering medium. Within a few months of Raman and Krishnan's announcement of their discovery, the Soviet physicists G. Landsberg and L. Mandelstam communicated their independent discovery of the effect by studying scattering of light in crystals, and in the Russian literature the phenomenon is still often referred to as combination scattering.

Compton scattering is another example of inelastic scattering. In this case, x-ray or gamma-ray photons cause a scattering electron to recoil, transferring energy to it and leading to a reduction in photon frequency. *See* COMPTON EFFECT.

Physical principles. The mechanism of the Raman effect can be understood either by the particle picture of light or by wave theory. Both features emerge in the quantum theory of radiation.

Photon picture. The particle model envisages light quanta or photons traversing a molecular medium undergoing scattering. If the interaction is elastic, the photons scatter with unchanged energy E and momentum, and hence with unchanged frequency ν, using Planck's relationship, $E = h\nu$, where h is Planck's constant. Such a process gives rise to Rayleigh scattering. If, however, the collision is inelastic, the photons can gain or lose energy from the molecules. A change ΔE in the internal energy of the scatterer causes a change $-\Delta E$ in photon energy, hence a corresponding frequency change $\Delta \nu = -\Delta E/h$. Frequency-shifted scattered photons can occur at lower or higher energy relative to the incident photons, depending on whether the scattering molecule is in the ground or excited vibrational state. In the former case, photons lose energy by exciting a vibration and the scattered light appears at a lower frequency, $\nu_s = \nu - \Delta E/h$, called the Stokes
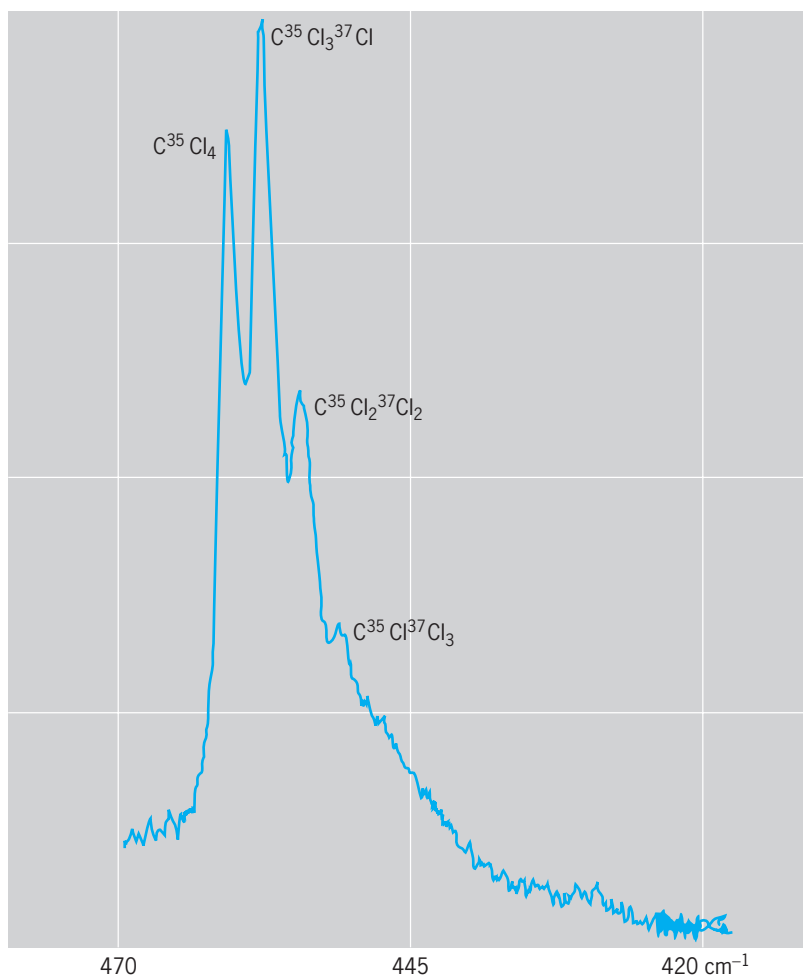
Fig. 1. Raman spectrum of carbon tetrachloride excited by He-Ne laser light at 632.8 nm. The scattered signal is recorded as a function of the frequency shift from the exciting spectral line. These shifts are usually expressed in units of inverse wavelength, $1/\lambda$, with λ measured in centimeters, called wave numbers and denoted by cm^{-1}, rather than actual frequency. (The frequency is $\nu = c/\lambda$, with c the speed of light in vacuum, so the two are proportional.) The Rayleigh-scattered spectral line (peak at 0 cm^{-1}) is 3 orders of magnitude more intense than the Raman lines, and its maximum is therefore far off scale. The Stokes lines appear at negative frequency shifts and the (weaker) anti-Stokes lines at positive frequency shifts. The lower scale gives the wavelengths in nanometers.

ΔE the internal energy of the scatterer. This induced dipole moment interacts with the light wave, producing emission not only at the frequency ν of the incoming field (elastic component), but also at sidebands at the sum and difference frequencies between the incoming light and the molecular vibrations (Stokes and anti-Stokes components ν_S and ν_{aS}). *See* DIPOLE MOMENT.

In general, any scatterer with internal energy structure can undergo the Raman effect by exchanging internal energy with the radiation field. In molecules, rotations, as well as vibrations, can produce Raman effects. Similarly, atoms can give rise to Raman scattering. However, in the context of revealing chemical structure, the Raman effect is of the greatest interest in the spectroscopy of molecules and crystals. *See* MOLECULAR STRUCTURE AND SPECTRA.

Raman spectroscopy. Raman scattering is analyzed by spectroscopic means. When a medium is excited by monochromatic incident light, the collection of new frequencies in the spectrum of scattered light is called its Raman spectrum. As the Raman effect probes vibrations of the molecule, which depend on the kinds of constituent atoms and their bond strengths and arrangements, a Raman spectrum provides a high degree of chemical structural information. **Figure 1** shows a typical Raman spectrum. The kind of information provided by this spectrum is similar to that provided by infrared spectroscopy; in fact, Raman and infrared spectra often provide complementary data about molecular structure. *See* INFRARED SPECTROSCOPY.

The advent of lasers as intense sources of monochromatic light was a milestone in the history of Raman spectroscopy and resulted in dramatically improved scattering signals. Today, intense sources of laser light over a wide range of frequencies from the near-ultraviolet to the near-infrared region are used in Raman scattering studies, enabling selection of optimum excitation conditions for each sample. Tunable dye lasers and, most recently, high-intensity diode lasers have become useful sources of near-infrared light and have made this spectral region attractive for practical applications. In particular, diode lasers provide compact, inexpensive Raman instrumentation. Development of low-noise, high-quantum-efficiency multichannel detectors [charge-coupled device (CCD) arrays], combined with high-throughput single-stage spectrographs used in combination with holographic laser rejection filters (such as Raman "notch" filters), has led to high-sensitivity Raman spectrometers. In addition, by using a microscope, particularly in combination with diaphragms, for focusing the excitation light and collecting the scattered light, Raman signals from femtoliter volumes (about 1 μm^3) can be observed, enabling spatially resolved measurements from very small structures (micro-Raman spectroscopy). Also, using pulsed laser sources, Raman spectra can be measured on the picosecond time scale, providing information about short-lived species. *See* CHARGE-COUPLED DEVICES; LASER; SPECTROGRAPH.

Figure 2 shows the layout of an advanced Raman facility. Tunable excitation provides access to all

frequency. In the latter case, by interacting with a molecule in an excited vibrational state, the photons gain energy from the molecular vibrations and the scattered signal appears at higher frequency $\nu_{aS} = \nu + \Delta E/h$, the anti-Stokes frequency. In general, anti-Stokes Raman scattering results in much smaller scattering signals than Stokes scattering because only a small fraction of molecules is in an excited vibrational state and can contribute to anti-Stokes Raman scattering. *See* LIGHT; QUANTUM MECHANICS; SCATTERING EXPERIMENTS (ATOMS AND MOLECULES).

Wave picture. In the wave picture, light as an electromagnetic wave interacts with the molecules in the medium, inducing an oscillating dipole moment by separating the positively charged nuclei and negatively charged electrons. In Rayleigh scattering (elastic), the oscillation frequency is ν, the same as that of the incident light. However, in the case of Raman scattering (inelastic), scattering from the molecules, vibrating at frequency $\Delta E/h$, gives rise to two frequency components at $\nu \pm \Delta E/h$, with

Fig. 2. Layout of a modern Raman facility.

visible and near-infrared wavelengths for resonance Raman studies. In addition, a microscope, with a three-axis control stage and multiple objectives, enables excitation regions as small as 0.5 μm in diameter to be studied. *See* CONFOCAL MICROSCOPY; OPTICAL MICROSCOPE.

The main advantage of Raman spectroscopy is its capability to provide rich information about the molecular structure of the sample. Sophisticated data analysis techniques based on multivariate analysis, in which spectral information from many Raman spectral lines is used simultaneously, have made it possible to exploit the full information content of Raman spectra, and to elucidate the chemical structure and composition of very complex systems such as biological materials. *See* SPECTROSCOPY.

Incoherent Raman scattering. The above discussion on Raman spectroscopy illustrates the most common form of incoherent Raman spectroscopy, also known as the spontaneous Raman effect. Several other variants of incoherent Raman scattering are also important.

Resonance Raman scattering (RRS). In usual Raman scattering studies, the frequency of the excitation light is not resonant with any actual molecular transition. In resonance Raman scattering, however, the excitation laser frequency is tuned to match a specific absorption peak of the scattering medium. Under these conditions the Raman scattering can become up to 10,000 times stronger, but in most molecules it is a different and much weaker process compared to fluorescence, which is excited in the same way and, in many cases, produces overlapping emission. In addition to the increased size of the Raman signals, resonance Raman scattering has the advantage of higher specificity, since the resonance Raman spectrum is dominated by vibrations related to the part of the molecule responsible for the appropriate resonant electronic transition. This is useful in understanding and assigning Raman spectra of large and complex molecules.

Raman scattering and fluorescence can be distinguished by the characteristics of their emission spectra. The fluorescence emission spectrum, determined by the electronic transition frequencies of the molecule, always occurs at the same frequency, regardless of the frequency of the excitation light. In contrast, the Raman emission spectrum is always shifted from the frequency of the excitation light by the same amount, regardless of its frequency. In many cases, resonance Raman scattering and fluorescence can also be differentiated in time-resolved experiments: Whereas resonance Raman scattering emission occurs at the same time as the excitation laser pulse, fluorescence occurs with some time delay, on the order of picoseconds to nanoseconds. *See* FLUORESCENCE.

Surface-enhanced Raman scattering (SERS). As discussed above, Raman scattering signals are usually exceedingly weak, rendering detection difficult and limiting practical applications. Therefore, about 50 years after the discovery of the Raman effect, a development which showed unexpectedly strong Raman signals from pyridine on a rough silver electrode surface attracted considerable attention. The effect was called surface-enhanced Raman scattering. Later experiments showed that it is not so much a surface effect as it is a nanostructure effect. (A nanostructure is a collection of nanoparticles, regular arrangements or clusters.) The strongly enhanced Raman signals arise from molecules attached to silver and gold nanostructures, such as small colloidal particles in solution, or silver or gold nano-island films. The enhancement of Raman signals in SERS can be as large as 14 orders of magnitude. Such a tremendous enhancement effect is primarily due to the optical properties of metal nanoparticles, based on the collective oscillations of their conduction electrons, known as surface plasmons. For silver and gold nanoparticles, these oscillations can span a relatively broad spectral range, and fall in the frequency range of Raman excitation and scattered signals. Resonances between the excitation light, the Raman scattered light, and the surface plasmons create strongly enhanced and spatially confined local optical fields in the vicinity of the nanoparticles, where Raman scattering takes place,

resulting in strongly increased Raman signals. SERS enhancement factors associated with enhanced electromagnetic fields have been found to be up to 12 orders of magnitude in size. Chemical interactions between the molecule and the metal can further enhance the Raman signal size in ways that are still not completely understood. *See* NANOSTRUCTURE; PLASMON.

In addition to its general scientific interest, since the mid-1990s SERS has gained considerable interest as a spectroscopic tool. Enhancement factors of 14 orders of magnitude increase the sensitivity of SERS to the single-molecule level. At present, SERS is the only way to detect and simultaneously identify the chemical structure of a single molecule based on its strongly enhanced Raman signal. Another important feature of SERS is its ability to precisely locate Raman scattering. Exploiting local optical fields of special metallic nanostructures, SERS can result in probed volumes as narrow as 5 nm or less in lateral dimension.

Hyper-Raman scattering (HRS). At extremely high excitation intensities, scattering signals Raman-shifted relative to twice the frequency of the excitation laser can be observed. This effect, called hyper-Raman scattering, is most easily described in the photon picture: with an intense laser source there is a small probability that two photons can simultaneously interact with the same molecule in a manner similar to that of a single photon of approximately twice the frequency. Due to its two-photon nature, hyper-Raman scattering can probe vibrations (called silent modes) that cannot be studied in either Raman scattering or infrared absorption. Hyper-Raman scattering is a very weak effect and requires extremely strong pulsed laser excitation. However, it can be enhanced in an analogous fashion to normal Raman scattering in SERS, when the two-photon excited Raman process takes place in the vicinity of silver and gold nanoparticles. Hyper-Raman scattering is an example of a nonlinear incoherent Raman effect.

Coherent Raman scattering. In addition to the incoherent Raman processes discussed above, the availability of lasers has also opened the field of study of coherent or stimulated Raman processes.

Stimulated Raman scattering (SRS). Stimulated Raman emission (or scattering) in a Raman transition bears the same relationship to spontaneous Raman emission as does stimulated emission to spontaneous emission in a conventional molecular transition. In stimulated Raman scattering, a sample is pumped by an intense laser source, generating gain at the Stokes frequency. Spontaneous Raman light emitted along the pump direction can then be amplified, producing strong coherent Raman emission along that direction, comprising in most cases one strong Raman line. This intense Stokes light can, in turn, act as pump light for generating a second Stokes field. In this way a comb of intense frequencies can be generated.

Raman lasers and inverse Raman scattering. This Raman gain can also be used to convert intense light from pump to Raman frequencies. In this case, two incident laser fields offset in frequency by a molecular vibration frequency are incident on a Raman-active medium. As the waves propagate through the medium, pump field photons are converted into Stokes photons, so that the Raman field grows. In a closely related configuration, the higher-frequency field is broadband. Absorption can then be observed at the anti-Stokes frequencies, a process known as inverse Raman scattering or Stoicheff absorption.

Coherent anti-Stokes Raman scattering (CARS). Using two excitation lasers offset in frequency by the Stokes shift, stimulated scattering at the anti-Stokes frequency can also be obtained. This method of generating anti-Stokes light is called coherent anti-Stokes Raman scattering (CARS). It is a coherent four-wave mixing interaction, that is, three coherent waves (two pump fields and one Stokes field) interact nonlinearly to produce a fourth. Unlike the other stimulated scattering processes described above, the anti-Stokes light is phase-coherent with the excitation light waves. CARS has been studied for several decades, but the effect has recently become of interest as a spectroscopic method for imaging a specific chemical compound by subjecting the sample to two laser fields offset by the molecular vibrational frequency of the molecule to be imaged. *See* NONLINEAR OPTICS.

Applications. Raman spectroscopy is of considerable value in determining molecular structure and in chemical analysis. It can be used to characterize the chemical and structural composition of nearly all kinds of molecules in liquid, gas phase, and solid samples for both scientific study and industrial monitoring, such as control of products and processes. As with many optical spectroscopy effects, the Raman effect can be used to interrogate a substance without altering it under ambient conditions in almost every environment. Measuring a Raman spectrum does not require special sample preparation techniques, in contrast with infrared absorption. Materials can be
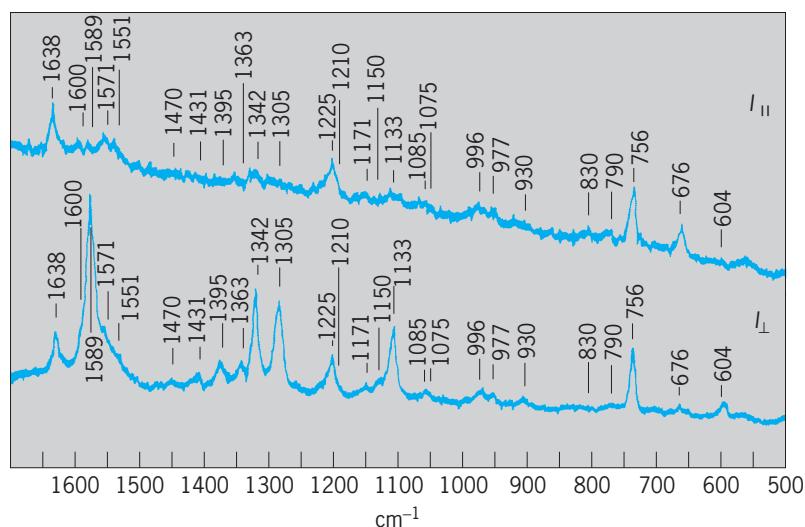
Fig. 3. Resonance Raman spectrum of oxyhemoglobin. (*After T. G. Spiro and T. C. Strekas, Resonance Raman spectra of hemoglobin and cytochrome c, Proc. Nat. Acad. Sci. USA, 69:2622–2626, 1972*)

studied remotely, either with line-of-sight techniques or via optical fiber probes (with appropriate filters). *See* FIBER-OPTIC SENSOR.

Biomedical applications. Raman spectroscopy can be used to characterize the chemical and structural composition of complex systems, such as biological materials. It has the potential to become a central tool for medical diagnosis, based on quantitative extraction of molecular information. The technique can monitor small changes in the chemical composition of a biological system such as human tissue, which can provide very early indications of development of disease. It is also used for noninvasive (that is, noncontact) measurement of glucose and other blood analytes in the body. As an example, (**Fig. 4**) shows Raman spectra of four tissue samples—one from normal tissue, two from benign lesions, and one from a breast cancer—along with stained microscopy sections of biopsy samples used in making the pathology diagnoses. The underlying

composition, obtained from multivariate analysis, is also shown. The spectral differences and concomitant changes in composition observed indicate the power of the technique.

Trace monitoring. With developments in SERS, Raman spectroscopy not only is a tool for structural analysis, but also allows ultrasensitive and trace detection of substances down to the single-molecule level. Moreover, SERS is an analytical technique that can provide information about surface and interface processes. SERS opens up exciting opportunities in laboratory medicine and for basic research in biophysics and biochemistry, where it provides ultrasensitive detection and characterization of biophysically and biomedically relevant molecules and processes, as well as a vibrational spectroscopy with extremely high spatial resolution.

Nanostructure materials. Raman spectroscopy has become an important tool for characterizing nanostructures such as carbon nanotubes and nanowires.
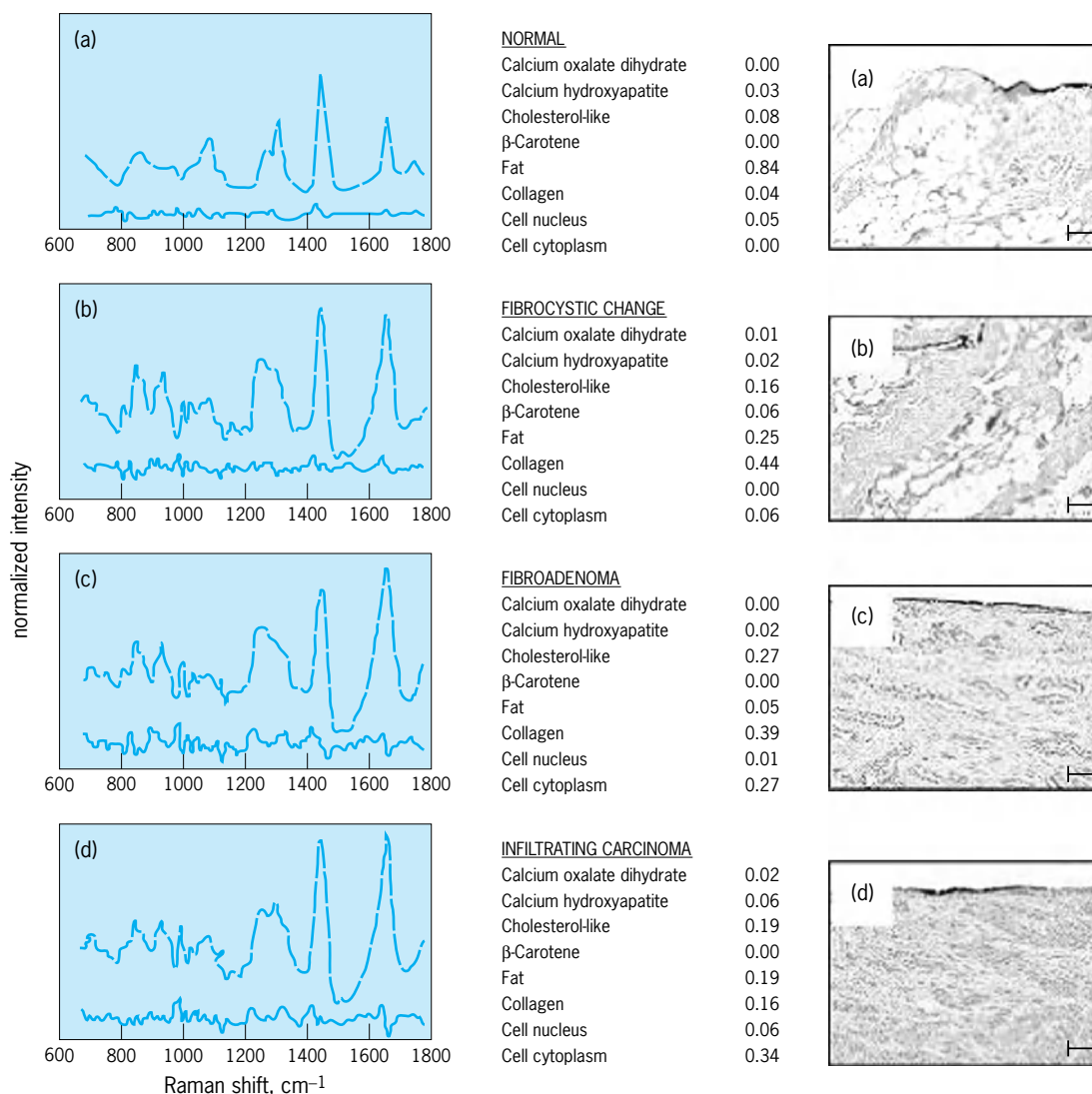
Fig. 4. Raman spectra of four human breast samples. The spectra at the left are from (*a*) normal tissue, (*b*,*c*), benign lesions, and (*d*) a breast cancer. At right, the corresponding pathology sections are shown. The chemical/structural compositions, based on multivariate spectral analysis, are shown in the center. In each panel on the left, the upper traces show the observed Raman spectrum and multivariate fit, and the spectrum below shows the difference between them.

Due to their extremely small size, in the nanometer range, the electronic transitions of nanostructures exhibit very specific and narrow novel features, which lead to interesting new and strong effects in resonance Raman spectroscopy, providing information about both the vibrational and electronic structure.   Katrin Kneipp; Michael S. Feld

Bibliography. N. Bloembergen, *Nonlinear Optics*, 4th ed., World Scientific, 1996; D. B. Chase, *Fourier Transform Raman Spectroscopy*, Academic Press, 1994; R. J. Clark and R. E. Hester (eds.), *Advances in Infrared and Raman Spectroscopy*, vols. 1–18, 1975–1989, *Advances in Spectroscopy*, vols. 19–26, Wiley, 1991–1998; N. B. Colthup, L. H. Daly, and S. E. Wiberly, *Introduction to Infrared and Raman Spectroscopy*, 3d ed., Academic Press, 1990; A. M. K. Enejder et al., Raman spectroscopy for non-invasive glucose measurements, *J. Biomed. Opt.*, 10(3):031114, May/June, 2005; J. R. Ferraro, K. Nakamoto, and C. W. Brown, *Introductory Raman Spectroscopy*, 2d ed., Academic Press, 2003; E. B. Hanlon et al., Prospects for in vivo Raman spectroscopy, *Phys. Med. Biol.*, 45(2):R1–R59, 2000; K. Kneipp et al., Ultrasensitive chemical analysis by Raman spectroscopy, *Chem. Rev.*, 10:2957–2975, 1999.

Fig. 1. Ramie (*Boehmeria nivea*), height 3–6 ft (1–2 m).

Ramie

The plant *Boehmeria nivea*, a stingless member of the nettle family; the only member of the family used commercially for fiber. This herbaceous perennial is erect, usually nonbranching, and 3 to 6 ft (1 to 2 m) tall at maturity (**Fig. 1**). The fiber, which comes from the inner bark, is exceptionally strong and has uses similar to those for fiber flax.

Cultivation. Ramie is grown most in the tropics and subtropics (to about 30–35° latitude) of the Far East and Brazil. It is planted from vegetative parts, usually root (rhizome) cuttings, but sometimes from stem cuttings. Because of the extreme variability of seedlings, seed is used only for variety development.

The rhizomes are placed in rows at a depth of about 5 cm (2 in.) on fertile, well-drained soil that has been worked two or more times to kill weeds that have sprouted. During the first year, while the plants spread and establish solid stand, weeds must be kept out, and the ramie must be kept cut often enough to prevent production of seed. Fertilizers expediate the establishment of the stand and are required in considerable amounts in subsequent years to replace the large quantities of nutrients removed by the crop.

Fiber extraction. Harvesting is usually done about every 60 days during the growing season. Three to as many as six crops are harvested annually. Plants are usually defoliated, either chemically or mechanically, before they are cut and brought to a central location for decortication (separation of the inner bark containing the fiber from the remainder of the stem).

On very small plantings, the fiber is usually extracted by hand-stripping and scraping methods. Elsewhere, hand methods have generally given way to mechanized decortication, most frequently a small hand-fed raspidor, driven by electricity or more often by a small internal combustion engine. Automatic decorticators similar to the ones used for sisal are used frequently on plantations. Ramie leaves are highly nutritious, and the tops or an entire young crop is sometimes used as livestock feed.

Diseases. Few diseases affect ramie and are usually limited to minor outbreaks. Probably the most serious is the white fungus, *Rosellina necatris*, reported in the Far East. The fungus attacks the roots and, by the time it is detected, may kill the plant. If infected planting stock is suspected, rhizome cuttings should be soaked in a mercuric chloride solution. *Rhizoctonia solani* (**Fig. 2**), a fungus that destroys young plantings and causes broken stems in older plants,
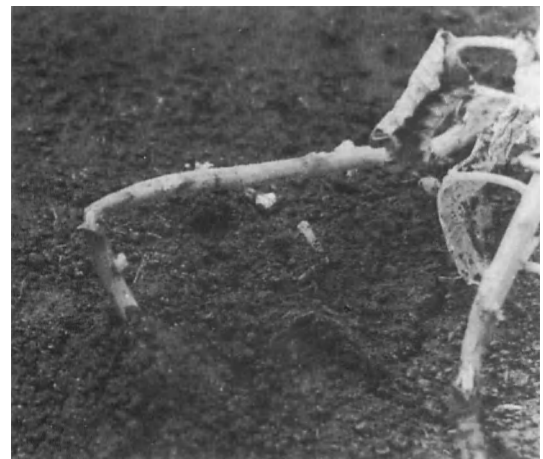
Fig. 2. Young plants attacked by *Rhizoctonia solani*.
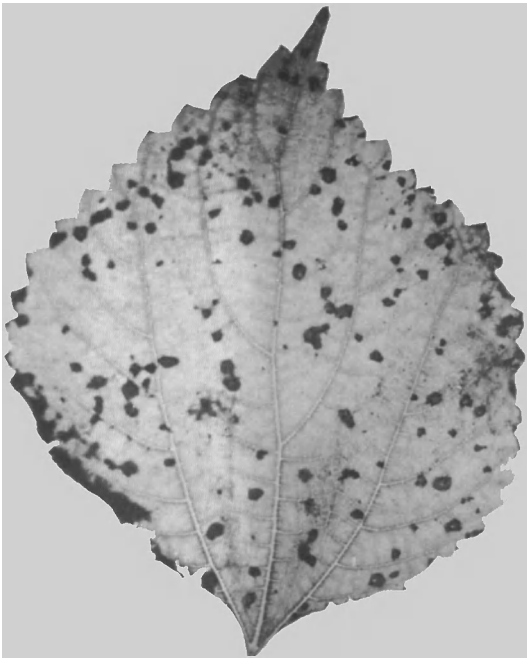
Fig. 3. Leaf spot caused by *Cercospora boehmeria*.

may be controlled by treating planting material with captan. The leaf-spot diseases, caused by *Cercospora boehmeria* and *C. krugliana*, retard the growth of young plants and may defoliate plants of poor vigor (**Fig. 3**). These diseases can be controlled by spraying or dusting infected areas with Dithane.

Many insects feed on ramie, but few are a serious threat. Without control, which is not difficult, several may become harmful. The leaf rollers, *Pilcrosis ramentalis* and *Sylepta silicalis*, are considered the most serious threat and are the most widespread. The larvae of *Pilcrosis* feed on the leaves, then roll up the leaf and pupate within it. Heavy infestations result in defoliation, and the plant ceases to grow. This pest may be controlled by BHC. The black caterpillar, *Cocytodes coerulea*, can also completely defoliate ramie. Dieldrin is an effective control. *See* FLAX; LINEN; NATURAL FIBER; PHLOEM; TEXTILE.
Elton G. Nelson

Bibliography. H. F. Linskens and J. F. Jackson (eds.), *Plant Fibers*, 1989; H. R. Mauersberger, *Mathew's Textile Fibers*, 6th ed., 1954.

Random matrices

Collections of large matrices, chosen at random from some ensemble. Random-matrix theory is a branch of mathematics which emerged from the study of complex physical problems, for which a statistical analysis is often more enlightening than a hopeless attempt to control every degree of freedom, or every detail of the dynamics. Although the connections to various parts of mathematics are very rich, the relevance of this approach to physics is also significant.

Principles. Random matrices were introduced by Eugene Wigner in nuclear physics in 1950. In quantum mechanics the discrete energy levels of a system of particles, bound together, are given by the eigenvalues of a hamiltonian operator, which embodies the interactions between the constituents. This leads to the Schrödinger equation which, in most cases of interest in the physics of nuclei, cannot be solved exactly, even with the most advanced computers. For a complex nucleus, instead of finding the location of the nuclear energy levels through untrustworthy approximate solutions, Wigner proposed to study the statistics of eigenvalues of large matrices, drawn at random from some ensemble. The only constraint is to choose an ensemble which respects the symmetries that are present in the forces between the nucleons in the original problem, and to select a sequence of levels corresponding to the quantum numbers that are conserved as a consequence of these symmetries, such as angular momentum and parity. The statistical theory does not attempt to predict the detailed sequence of energy levels of a given nucleus, but only the general properties of those sequences and, for instance, the presence of hidden symmetries. In many cases this is more important than knowing the exact location of a particular energy level. This program became the starting point of a new field, which is now widely used in mathematics and physics for the understanding of quantum chaos, disordered systems, fluctuations in mesoscopic systems, random surfaces, zeros of analytic functions, and so forth. *See* CONSERVATION LAWS (PHYSICS); EIGENVALUE (QUANTUM MECHANICS); MATRIX THEORY; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Initially Wigner introduced simple gaussian probability distributions for the matrix elements, with the requirement that the ensemble be invariant; that is, it should not favor any preferred direction in the space on which the matrices act. He could then prove that, for large matrices, the density of eigenvalues, that is, their number per unit interval of energy, followed a "semicircle" law. This means that there are essentially no eigenvalues outside an interval, for instance $[-1,1]$ in the appropriate normalization, and within this interval the density ρ varies with eigenvalue λ according to a law of the type given by Eq. (1). [The

$$\rho(\lambda) = \frac{2}{\pi}\sqrt{1 - \lambda^2} \qquad (1)$$

graph of this function is a semicircle.] The statistics of the eigenvalues, of the correlations between their positions, and of the distribution of their spacings led to many intricate new mathematical problems, even for the simple gaussian ensembles, but most of them are now well understood. The distribution of the spacings of successive eigenvalues is one of the characteristics of random matrices that is commonly used for practical applications. It exhibits level repulsion, which reflects the symmetries of the ensemble; indeed, eigenvalues of random matrices

behave quite differently from independent random numbers (whose spacings would follow the Poisson distribution, that is, exhibit no level repulsion). In a comparison with experiment, a spacing distribution must be selected that is appropriate to the basic properties of the system. For instance, when the dynamics is invariant under time reversal, which is true when there is no magnetic field acting on the system, real random matrices can be chosen. In the presence of a magnetic field, the matrices are necessarily complex and the spacing distribution is quite different. *See* DISTRIBUTION (PROBABILITY).

However, some doubts were soon raised concerning the applicability of the theory to real physical situations, since the density of nuclear levels, for instance, increases rapidly with growing excitation energies, and is certainly not given by a semicircular law. It was then conjectured, and proved much later, that the local statistics of levels, that is, the correlations and the spacings at the scale of the mean level separation, were universal, that is, largely independent of the probability distribution of the matrices. This universality is undoubtedly the main reason for the vast range of applicability of the theory, provided the research is limited to the study of the local statistics of levels.

Applications. A few examples outside the initial range of nuclear physics will illustrate some of the modern uses of random matrix theory.

Quantum chaos. Classically the time evolution of a system with several degrees of freedom is generically chaotic; that is, it shows a sensitive dependence upon its initial data (the butterfly effect of meteorologists). It has been conjectured that the same dynamical system, treated now by quantum mechanics instead of newtonian dynamics, exhibits random-matrix-level statistics in its energy levels. Numerical studies of nonintegrable systems, such as the hydrogen atom in a magnetic field, strongly support this conjecture, but it is still a matter of active research to understand the basic reason for the replacement of a deterministic nonrandom problem by a probabilistic treatment. *See* CHAOS; NONLINEAR PHYSICS.

Mesoscopic systems. In semiconductor physics, numerous interesting phenomena and applications have been found for quantum dots, which are small chaotic cavities, coupled via narrow leads to the rest of the system. The theory of quantum transport relies in such cases upon random matrices. Similarly the diffusive motion of electrons in small metallic grains, when there is enough time for the electrons to diffuse over the whole size of the grain, leads to a statistical distribution of energy levels in agreement with random matrix theory. *See* QUANTIZED ELECTRONIC STRUCTURE (QUEST).

The previous examples belong to a long list of phenomena concerning mesoscopic systems, that is, small semiconductor or metallic devices at very low temperatures. In these systems the electrons, scattered by the crystalline ions and by the impurities of the structure, retain the full quantum-mechanical phase coherence which allows for interference between different electronic paths. As a consequence of the small size of these systems, fluctuations of their conductance are observed from sample to sample, or in a single sample as a function of an applied magnetic field. The universality of these fluctuations was soon discovered theoretically and checked by experiment. This striking phenomenon is related to the universality present in the statistics of levels of random matrices. *See* AHARONOV-BOHM EFFECT; MESOSCOPIC PHYSICS.

String theory. Some other domains of application are more surprising. For instance, large random matrices have been found to encode, in some limit, the combinatorics of randomly triangulated surfaces of arbitrary genus. Those surfaces appear in string theory, an approach which attempts to embody Albert Einstein's gravity at the quantum level. The elementary objects of the physical world are no longer points in space, but bits of strings, which sweep out a two-dimensional surface during their motion. Random-matrix theory has brought some partial resolution to such problems, still very incomplete, but in a domain in which there are still very few explicit results. *See* SUPERSTRING THEORY.

Riemann zeta function. In mathematics, this function, defined by Eq. (2), is one of the most studied an-

$$\varsigma(s) = \sum_{1}^{\infty} \frac{1}{n^s} \tag{2}$$

alytic functions of a single complex variable. A famous conjecture of Bernhard Riemann has resisted the efforts of many waves of mathematicians to prove it. It states that, in the band $0 < \text{Re } s < 1$ of the complex s-plane, this function vanishes exclusively on the line $\text{Re } s = 1/2$. A huge number of zeros (hundreds of millions) have been computed numerically, and they have all been found to agree with Riemann's hypothesis. Furthermore, the statistics of those zeros, for instance the distribution of their spacings, measured in units of their local density, is in striking agreement with the spacing distribution of the eigenvalues of random matrices, drawn from an ensemble invariant under unitary transformations. The explanation of these remarkable observations remains a challenge to mathematicians. *See* NUMBER THEORY.

Relation to other fields. The mathematical theory underlying the properties of random matrices overlaps with several active fields of contemporary mathematics, such as the asymptotic behavior of orthogonal polynomials at large-order, integrable hierachies, tau functions, semiclassical expansions, combinatorics, and group theory; and it is the subject of active research and collaboration between physics and mathematics.   Edouard Brézin

Bibliography. C. E. Beenakker, Random-matrix theory of quantum transport, *Rev. Mod. Phys.*, 69:731–808, 1997; D. J. Gross, T. Piran, and S. Weinberg (eds.), *Two Dimensional Quantum Gravity and Random Surfaces*, Jerusalem Winter School, vol. 8, World Scientific, 1992; M. L. Mehta, *Random Matrices*, 2d ed., Academic Press, 1991.

Rangefinder (optics)

An optical instrument for measuring distance, usually from its position to a target point. Light from the target enters the optical system through two windows spaced apart, the distance between the windows being termed the base length of the rangefinder. The rangefinder operates as an angle-measuring device for solving the triangle comprising the rangefinder base length and the line from each window to the target point. Rangefinders can be classified in general as being of the coincidence or the stereoscopic type.

Coincidence rangefinders. In these types, one-eyed viewing through a single eyepiece provides the basis for manipulation of the rangefinder adjustment to cause two images or parts of each to match or coincide. This type of device is used, in its simpler forms, in photographic cameras. The range triangle for such a rangefinder is shown in **Fig. 1**, where B is the base, angle A a right angle, D the range, and L the convergence angle at the target T. The relationship is given by Eq. (1). The basic optical arrangement is

$$D = B \cot L \qquad (1)$$

shown in **Fig. 2**, where M_1 and M_2 are a semitransparent mirror and a reflecting mirror, respectively. When coincidence is obtained, that is, when the target T is seen in the same apparent position along either path, the rangefinder equation is satisfied. In many small rangefinders coincidence is achieved by rotating mirror M_2 through a small angle while viewing the images in mirror M_1; other deviating means which do not require rotating of mirror M_2 are also employed. **Figure 3**a and b shows the appearance of the images in a superposed field rangefinder. If mirror M_1 is made fully reflecting and arranged to fill only the lower half of the field of view, allowing the eye to view target T directly through the upper half, a split-field type of rangefinder is produced; the images appear as shown in **Fig. 4**a and b.
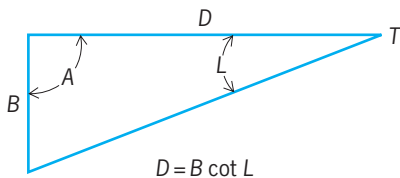
Fig. 1. Range triangle; symbols explained in the text.
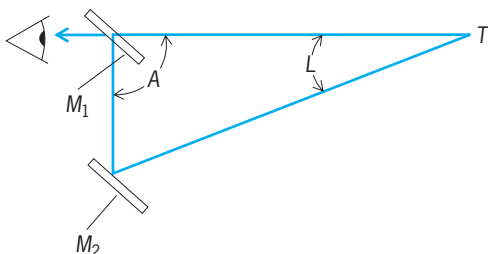
Fig. 2. Simple coincidence rangefinder; symbols explained in the text.
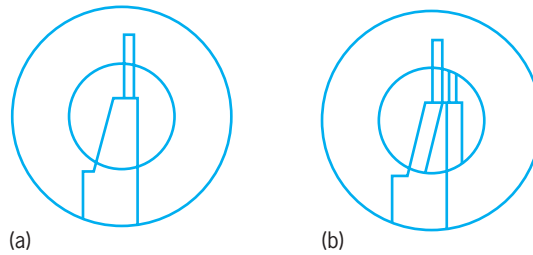
Fig. 3. View in superposed field rangefinder. (*a*) Images in coincidence. (*b*) Images unmatched.
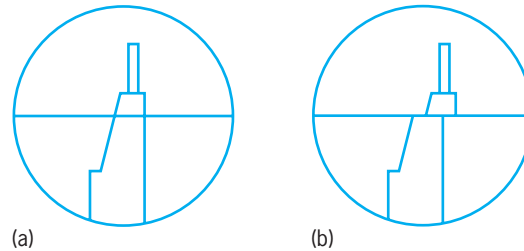
Fig. 4. View in split-field rangefinder. (*a*) Images matched. (*b*) Images unmatched.

Camera rangefinders. These usually have base lengths varying from less than 1 in. (25 mm) to more than 3 in. (75 mm), with magnifications between 0.5 and 2.5. The lower powers are found in arrangements where the rangefinder and viewfinder are combined for viewing through a single eyepiece; here a magnification of unity or less is necessary to cover the usual 42–52° angle of view.

Military types. Coincidence rangefinders of the military type ordinarily employ a pentaprism, or its two-mirror equivalent, at each window. These prisms, permanently fixed in place, reflect the light rays inward at an angle exactly 90° to the angle of incidence. The light from each prism passes through a telescope objective, and an image of the target formed by each objective is combined and directed by a prism assembly in a single eyepiece for simultaneous viewing. **Figure 5** shows the geometry to such a rangefinder. Here b is the base length of the rangefinder, and f' is the focal length of its objective lenses. Light from an infinitely distant target would arrive along lines TG and $T'H$, while light from a target distance R would arrive along lines TG and TK. The objectives, located at G and H and with their axes on lines TG and TH, respectively, form images of the target T at J and at K, respectively. With an infinitely distant target, objective H would form its image at L rather than at K. This displacement d, equal to LK, is called the parallactic displacement and, since triangles TGH and HLK are similar, is the measure of range. The relationship is given by Eq. (2).

$$R = \frac{bf'}{d} \qquad (2)$$

The optical arrangement of a typical military coincidence rangefinder is shown in **Fig. 6**. The rangefinder is adjusted so that if the object being

viewed is at infinity (dotted line path through objective *H*), both images of the object are seen exactly superimposed. If the object is closer than infinity, the axis of the image which is formed by objective *H* is displaced by a deviation means, such as prism *D*, in the manner and amount described by Eq. (2), so that superimposition is achieved. The amount of such displacement is shown on the scale as the distance reading. Other means for obtaining deviation include oppositely rotating wedges (diasporometer), two equal coaxial oppositely arranged prisms with variable spacing, and slidable positive and negative lenses (swing wedge).

Stereoscopic rangefinders. These are entirely different, although externally they resemble coincidence rangefinders except for the fact that they possess two eyepieces. Both eyes are used with this instrument, rangefinding being accomplished by stereoscopic vision. It is essentially a large stereobinocular fitted with special reticles which allow a skilled user to superimpose the stereo image formed by the pair of reticles over the images of the target seen in the eyepieces, so that the reticle marks appear to be suspended over the target and at the
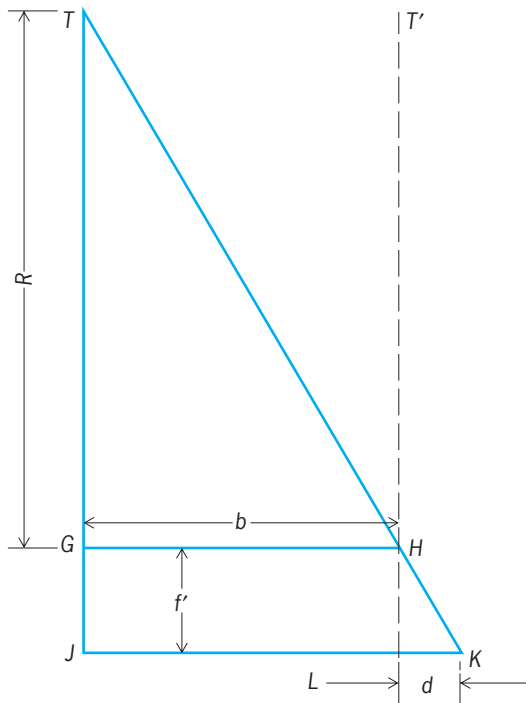
Fig. 5. Range triangle for military coincidence rangefinder; symbols explained in text.
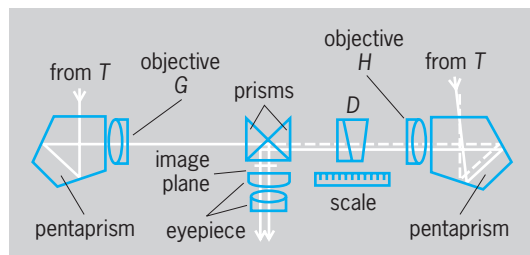
Fig. 6. Military coincidence rangefinder.
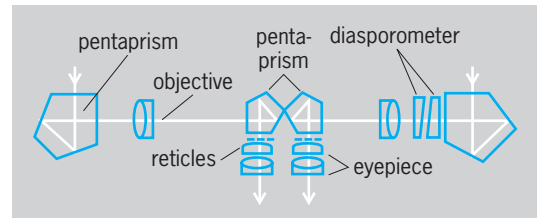
Fig. 7. Stereoscopic rangefinder.

same apparent distance. In the typical stereoscopic rangefinder (**Fig. 7**) the objectives form their images in the plane of their respective reticles. A diasporometer in one half of the instrument is used to vary the angle between the beams reaching the eyes from the target until it equals that between the beams reaching the eyes from the two reticles, at which point the reticle image appears to be at the same distance as the target. *See* STEREOSCOPY.

Rangefinder errors. The accuracy of a rangefinder depends upon the accuracy of the eye in judging the coincidence of two lines, as well as upon the characteristics of the instrument and the range distance. Rangefinder errors are proportional to the square of the range and inversely proportional to the base length and to the magnification. A 1.1-yd (1-m) base and a 3.3-yd (3-m) base rangefinder of 15 power would have errors of 7.3 and 2.7 yd (6.7 and 2.5 m) at 1000-yd (914-m) range and 470 yd (430 m) and 170 yd (155 m) at 8000-yd (7315-m) range, respectively.

Heightfinders. These are modified forms of rangefinders originally used in the directing of antiaircraft fire. A heightfinder depends for its operation upon the knowledge of the range *R* to the aircraft and its angle of elevation θ. The height is given by Eq. (3).

$$H = R \sin \theta \qquad (3)$$

In practice, one way that the height may be measured is by multiplying range and elevation angle in a gear train computer. Another method employs an additional pair of deviation prisms within the rangefinder, geared so as to produce a deviation which varies as $\sin \theta$.

Depression rangefinders. These are used in coast defense positions, where such instruments can be installed on a cliff top, and this elevation is used as the rangefinder base. The device measures the depression angle of the target, which it converts directly to range after including necessary corrections for tide, curvature of the Earth, and refraction of the atmosphere.

Stadia methods of rangefinding. These are suitable for some applications. If the size of the object is known or can be guessed accurately, its distance can be determined by measuring its size in the eyepiece reticle of a telescopic system. *See* LENS (OPTICS); OPTICAL PRISM.   Edward K. Kaprelian

Bibliography. Optical Society of America, *Handbook of Optics*, 2 vols., 2d ed., 1995; W. J. Smith, *Modern Optical Engineering*, 3d ed., 2000.

Rankine cycle

A thermodynamic cycle used as an ideal standard for the comparative performance of heat-engine and heat-pump installations operating with a condensable vapor as the working fluid. Applied typically to a steam power plant, as shown in the **illustration**, the cycle has four phases: (1) heat addition *bcde* in a boiler at constant pressure p_1 changing water at *b* to superheated steam at *e*, (2) isentropic expansion *ef* in a prime mover from initial pressure p_1 to back pressure p_2, (3) heat rejection *fa* in a condenser at constant pressure p_2 with wet steam at *f* converted to saturated liquid at *a*, and (4) isentropic compression *ab* of water in a feed pump from pressure p_2 to pressure p_1.

This cycle more closely approximates the operations in a real steam power plant than does the Carnot cycle. Between given temperature limits it offers a lower ideal thermal efficiency for the conversion of heat into work than does the Carnot standard. Losses from irreversibility, in turn, make the conversion efficiency in an actual plant less than the Rankine cycle standard. *See* CARNOT CYCLE; REFRIGERATION CYCLE; THERMODYNAMIC CYCLE; VAPOR CYCLE.

Theodore Baumeister

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; G. N. Hatsopoulos and J. H. Keenan, *Principles of General Thermodynamics*, 1965, reprint 1981; J. B. Jones and G. A. Hawkins, *Engineering Thermodynamics*: *An Introductory Textbook*, 1960.

Ranunculales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Magnoliidae of the class Magnoliopsida (dicotyledons). The order consists of 8 families and about 3200 species. The vast majority of the species belong to only 3 families, the Ranunculaceae (2000), Berberidaceae (650), and Menispermaceae (425). Within its subclass, the order is characterized by its mostly separate carpels, triaperturate pollen, herbaceous or only secondarily woody habit, frequently numerous stamens, generally more than two sepals, and lack of ethereal oil cells. Many members of the order produce benzyl-isoquinoline alkaloids or aporphine alkaloids, or both. The Ranunculales are generally considered to be derived from the Magnoliales, a woody order in the same subclass. The barberry (*Berberis*), in the family Berberidaceae, and the buttercup (*Ranunculus*), columbine (*Aquilegia*; see **illus.**), larkspur (*Delphinium*), and windflower (*Anemone*), in the family Ranunculaceae, are familiar genera.

Rankine-cycle diagrams (pressure-volume and temperature-entropy) for a steam power plant using superheated steam. Typically, cycle has four phases.

Colorado columbine (*Aquilegia coerulea*) of family Ranunculaceae. (*U.S. Forest Service photograph by C. A. Kutzleb*)

See MAGNOLIALES; MAGNOLIOPHYTA; MAGNOLIOP-SIDA; PLANT KINGDOM.   Arthur Cronquist; T. M. Barkley

Rape

Rape (*Brassica napus*) and turnip rape (*B. campestris*) plants are members of the Cruciferae family. The name is derived from the Latin *rapum*, meaning "turnip," to which these plants are closely related. The aerial portions of rape plants have been bred to produce oilseeds, fodder, and vegetable crops. Rape seed is small (0.07 to 0.18 oz or 2 to 5 g per 1000 seeds), round, and usually black, although varieties with yellow seed coats are also grown. The seeds, borne in long slender pods or siliques, contain over 40% oil. The plant germinates rapidly and forms a rosette of bluish-green leaves from which bolts an indeterminate racemose inflorescence (see **illus.**). Both annual and biennial forms of the crop are grown; the biennial form will not flower without extended exposure to freezing temperatures. *See* CAPPARALES.

Brassica oilseed crops (including *B. juncea*) are the world's third most important source of edible vegetable oil, with production centered in Canada, northern Europe, China, and the Indian subcontinent. Seed, oil, and meal from rapeseed plants that contain superior nutritional qualities are known worldwide as canola. Canola plants were developed by genetically blocking the biosynthesis of the nutritionally undesirable fatty acid, erucic, and sulfur compounds called glucosinolates. Canola oil is a highly nutritious salad and cooking oil and is used in the manufacture of margarine and shortening. It is also used in Europe as a diesel fuel, while rape oil is used industrially in the manufacture of plastics and as a lubricant. Canola meal, the by-product of oilseed extraction, is a high-quality protein feed supplement for livestock and poultry. In some Asian and Indian countries, where conversion to canola quality varieties has not yet occurred, rape meal is used as a fertilizer.

Rape is widely used for forage. In some countries the whole plant is cut and fed to cattle, but more frequently biennial rape, such as the Dwarf Essex variety, is used as summer or late fall and winter pasture for sheep and swine. It is in the latter form that rape is most widely used throughout the United States and New Zealand. Some forms of turnip rape are used as leaf vegetables in Asia. *See* TURNIP.   R. K. Downey

Bibliography. R. K. Downey and R. Rimmer, Agronomic improvement in oilseed brassicas, *Adv. Agron.*, 50:1–66, 1993; G. Röbbelen, R. K. Downey, and A. Ashri (eds.), *Oil Crops of the World*, 1989; F. Shahidi (ed.), *Canola and Rapeseed*, 1990.

Well-podded plants of annual rape (*Brassica napus*) with inset showing the characteristic inflorescence and basal leaves.

Raphidophyceae

A small class of poorly known biflagellate unicellular algae (raphidomonads) in the chlorophyll *a*–*c* phyletic line (Chromophycota). It is sometimes called Chloromonadophyceae, and its members called chloromonads, but this nomenclature is confusing because it calls to mind *Chloromonas*, a totally unrelated genus of green algae (Chlorophyceae). The alternative nomenclature used here is derived from *Raphidomonas*, a generic name within the class.

The cell of a typical raphidomonad is flattened, dorsiventral or bilaterally symmetrical, and naked. Two flagella are borne at the base of an anterior groove. One flagellum has hairs and is directed anteriorly, while the other is smooth and trails close to the cell body. The cytoplasm is segregated into an endoplasm, which contains the nucleus and mitochondria, and an ectoplasm, which contains numerous discoid chloroplasts (in photosynthetic forms), an anterior assemblage of small contractile vacuoles (in fresh-water forms), and commonly ejectile structures (trichocysts and mucocysts). Chloroplasts have four membranes and three thylakoids per lamella. Photosynthetic pigments include chlorophyll *a* and *c*, β-carotene, and various xanthophylls, of which fucoxanthin has been reported only in marine forms. Depending upon pigment composition, the chloroplasts are bright green, yellow-green, or yellow-brown. Pyrenoids are present only in marine forms. Reserve food particles are unknown, and the storage product apparently is oil. Eyespots are absent. The nucleus is relatively large, located in the center or anterior portion of the cell, and capped by an array of Golgi bodies. *See* CELL PLASTIDS; CILIA AND FLAGELLA.

Gelatinous palmelloid colonies composed of nonflagellate spheroidal cells have been observed.

Reproduction apparently is restricted to longitudinal binary fission, which can take place while the cell is motile or palmelloid.

Two families are recognized. All photosynthetic raphidomonads are placed in the Vacuolariaceae. The Thaumatomastigaceae comprises a few colorless forms that bear pseudopodia. They are osmotrophic or phagotrophic or both.

Most genera in both families occur in fresh water (acidic to neutral), but there are brackish-water and marine forms that produce conspicuous blooms. *Heterosigma*, for example, is a frequent cause of red tide in the inland seas of Japan. *See* ALGAE.   Paul C. Silva; Richard L. Moe

Bibliography. E. R. Cox (ed.), *Phytoflagellates*, 1980.

Rare-earth elements

The group of 17 chemical elements with atomic numbers 21, 39, and 57–71; the name lanthanides is reserved for the elements 58–71. The name rare earths is a misnomer, because they are neither rare nor earths. The early Greeks believed that everything in the world was made of four elements: air, earth, fire, and water. The earths were substances which could not be changed with heat by the temperatures then available to the scientist; and in the early part of the nineteenth century, when the first rare earths were discovered, they resembled the common earths, which were really oxides of magnesium, calcium, and aluminum. Since the rare earths were found in very rare minerals, they were thus called rare earths. They are not rare, however, since cerium is reported to be more abundant in the Earth's crust than tin; yttrium more abundant than lead; and even the scarce rare earths, except promethium, more abundant than the platinum-group elements. All these elements form trivalent bonds, and when their salts are dissolved in water, they ionize to form trivalent ions and the solutions exhibit very similar chemical properties. The elements scandium, yttrium, lanthanum, and actinium in the II column of the extended periodic table show similar properties in aqueous solution. Yttrium and lanthanum are always found associated with the rare earths in nature.

The similarity of properties among the lanthanide rare earths (atomic numbers 58 through 71) originates from the fact that as the atomic number increases in this part of the periodic table, the increased charge on the nucleus is compensated for by electrons which start filling an inner incomplete subshell ($4f$). This subshell can hold up to 14 electrons; there are accordingly 14 elements in the lanthanide series. These extra electrons, however, play almost no role in the valency forces between atoms. The elements with atomic numbers 90 through 103 also occur in the periodic table at a place where a similar inner subshell ($5f$) is being completed. In many ways, these elements resemble the lanthanides. They are frequently referred to as the actinide rare earths. Both groups are usually displayed at the bottom of the periodic table as an appendage of two rows. *See* ACTINIDE ELEMENTS; LANTHANIDE CONTRACTION; PERIODIC TABLE.

Properties. The rare-earth elements are metals possessing distinct individual properties which make them potentially valuable as alloying agents. They are usually reduced thermally by treating the anhydrous halide with calcium, lithium, or other alkali metals and then remelting under vacuum to volatilize the last traces of the reductant. They can also be reduced eletrolytically from fused-salt baths, as is done commercially for cerium and misch metal (a mixed rare-earth metal, mainly cerium, with small amounts of iron present).

Samarium, europium, and ytterbium cannot be reduced by the above methods. These elements form divalent halides when so treated. Fortunately, these metals have very low boiling points so they can be distilled from a mixture of lanthanum or cerium and the oxides of these elements. The pure samarium, europium, and ytterbium are collected on the condenser, and lanthanum or cerium oxides are left behind.

Table 1 gives the properties of the metals. The anhydrous salts show greater differences in properties among the elements than do the hydrated salts.

Many of the properties of the metals and alloys are quite sensitive to temperature and pressure. They are also different when measured along different crystal axes of the metal; for example, electrical conductivity, elastic constants, and so on. If the properties are plotted against temperature and pressure, they show anomalies whenever a crystal or magnetic transition takes place.

The rare earths form organic salts with certain organic chelate compounds. These chelates, which have replaced some of the water around the ions, enhance the differences in properties among the individual rare earths. Advantage is taken of this technique in the modern ion-exchange methods of separation. *See* CHELATION; CHROMATOGRAPHY; ION EXCHANGE; MAGNETOCHEMISTRY; TRANSITION ELEMENTS.

Occurrence. Although the rare earths are widely distributed in nature, they generally occur in low concentrations. They also are found in high concentrations as mixtures in a number of minerals (**Table 2**). The relative abundance of the different rare earths in various rocks, geological formations, and the stars is of great interest to the geophysicist, astrophysicist, and cosmologist. The development of precision methods for determining the abundance of the various elements present in trace amounts by means of mass spectroscopy, optical and infrared spectroscopy, and radio and nuclear chemistry permits the relative abundance of the rare earths in any sample to be determined with considerable precision. It is found that the deep basic rocks, such as basalt, contain a few parts per million of mixed rare earths, and that the more acid silicate rocks contain a slightly higher concentration. When these two types of rocks come in contact in the molten state, the more acid rocks extract some of the mixed rare

TABLE 1. Properties of the rare-earth metals

Symbol	Transitions	Transition temperature, °C (°F)	ΔH transition, cal	Lattice parameters, Å			Ionic radius RE^{3+}, nm
				a	c	c/a	
Sc	hcp (ABAB)	—	—	3.3088	5.2680	1.592	0.0732
	bcc	1335 (2435)	958	—	—	—	—
Y	hcp (ABAB)	—	—	3.6482	5.7318	1.571	0.0893
	bcc	1478 (2692)	1189	4.08	—	—	—
La	dhcp (ABAC)	—	—	3.7740	12.171	1.6125*	0.1061
	fcc (ABC)	310 (590)	67	—	—	—	—
	bcc	865 (1589)	753	4.26	—	—	—
Ce	fcc (ABC)	—	—	4.85	—	—	—
	dhcp (ABAC)	−178 (−288)	—	3.68	11.92	1.619*	0.1034
	fcc (ABC)	~100 (heating) (212)	—	5.160	—	—	Ce^{4+}
		−10 (cooling) (14)	—	—	—	—	0.0092
	bcc	726 (1339)	700	4.12	—	—	—
Pr	dhcp (ABAC)	—	—	3.6721	11.832	1.611*	0.1013
	bcc	795 (1463)	760	4.13	—	—	Pr^{4+}
							0.090
Nd	dhcp (ABAC)	—	—	3.6583	11.7966	1.612*	0.0995
	bcc	863 (1586)	713	4.13	—	—	—
Pm							0.0979
Sm	hcp (nonprimitive)	—	—	3.6290	26.207	1.605	0.0964
	bcc	926 (1699)	744	—	—	—	Sm^{2+}
							0.111
Eu	bcc	—	—	4.5827	—	—	0.0950
							Eu^{2+}
							0.109
Gd	hcp (ABAB)	—	—	3.6336	5.7810	1.59	0.0938
	bcc	1238 (2255)	935	4.05	—	—	—
Tb	hcp (ABAB)	—	—	3.6055	5.6966	1.580	—
	bcc	1289 (2352)	1203	—	—	—	0.0923
							TB^{4+}
		—	—	—	—	—	0.084
Dy	hcp (ABAB)	—	—	3.5915	5.6501	1.573	0.0908
	bcc	1381 (2518)	955	—	—	—	—
Ho	hcp (ABAB)	—	—	3.5778	5.6178	1.570	0.0894
Er	hcp (ABAB)	—	—	3.5592	5.5850	1.569	0.0881
Tm	hcp (ABAB)	—	—	3.5375	5.5540	1.570	0.0869
Yb	fcc (ABC)	—	—	5.4848	—	—	0.0858
	bcc	795 (1463)	418	—	—	—	Yb^{2+}
							0.093
Lu	hcp (ABAB)			3.5052	5.5495	1.583	0.0848

*The values are the author's opinion of the best average values. The actual values vary slightly in the original literature, depending on the investigator and the purity of the metal at hand. Most of the values were obtained on 99.8–99.9% pure metal, the principal impurities being hydrogen, oxygen, nitrogen, carbon, and tantalum. Some of these values will change slightly as purer metal are produced.

earths from the basalt, and the light rare earths are extracted to a larger extent than the heavy ones. Thus, by determining a profile of the relative abundance of the rare earths in given rocks, it can be determined whether they have been molten. Such tests were performed on the various rocks brought back from the Moon. Also, if deposits of rich rare-earth minerals are found in a formation, one can determine the physical condition with regard to temperature and pressure to which the formation must have been subjected in order to form such minerals. The relative abundance of the various rare earths present in stars as determined from their spectra gives clues as to what nuclear processes are taking place in those stars. The numerous scientific papers published in this field make interesting reading, since theories as to the formation of the solar system and the universe are in part based on such data. Monazite, xenotime, and bastnasite are among the more important sources.

These minerals are usually concentrated from other rock and minerals by mechanical means, such as flotation or magnetic cross-belt separation methods. The rare earths are leached from the minerals with acid in the case of the phosphate or silicate minerals. Some minerals, such as the columbotantalates, have to be heated with carbon or treated with strong caustic before undergoing the leaching process. *See* ORE DRESSING.

Separation. The mixed rare earths can be separated from the acid solutions by means of oxalate precipitation or by other insoluble precipitates; ignition of the oxalate gives a mixed rare-earth oxide. They are frequently concentrated directly by ion-exchange methods, using the acid leach from the minerals.

The rare earths occur in solution as hydrated trivalent ions whose properties are very much alike; therefore they tend to form mixed crystal precipitates or solid solutions. A single chemical operation

Melting point, °C (°F)	Boiling point, °C (°F)	Heat of vaporization ($\Delta H_{v,0}$), kcal/g-atm	Density 25°C, g/cm³	Molal volume 25°C, cm³/g-atm	Radius metal atom, nm	Electrical resistivity (4.2 K), ohm-cm 10^{-6}	Residual resistivity (4.2 K), ohm-cm 10^{-6}	Compressi- bility, cm²/kg 10^{-6}
1541	2831 (5128)	89.9	2.9890	15.041	0.1640	52	3	2.26
1522	3338 (6040)	101.3	4.4689	19.894	0.1801	59	2	2.68
921 (1690)	3457 (6255)	103.1	6.1453	22.603	0.1879	61–80	Super-conductor	4.04
799 (1470)	3426 (6199)	101.1	6.672	21.001	0.1820	70–80	10	4.10
931 (1708)	3512 (6353)	85.3	6.773	20.805	0.1828	68	1	3.21
1021 (1870)	3068 (5554)	78.5	7.007	20.585	0.1821	65	7	3.0
1168 (2134)	2700 (est.) (4892)	—	—	—	—	—	—	(2.8)
1077 (1971)	1791 (3256)	49.2	7.520	20.001	0.1804	91	7	3.34
822 (1531)	1597 (2907)	(41.9) $\Delta H° = 29$	5.2434	28.981	0.1984	91	1	8.29
1313 (2395)	3266 (5911)	95.3	7.9004	19.9041	0.1801	127	1	2.56
1356 (2473)	3123 (5653)	93.4	8.2294	19.3119	0.1783	114	4	2.45
1412 (2574)	2562 (4646)	70.0	8.5500	19.0058	0.1774	100	5	2.55
1474 (2685)	2695 (4883)	72.3	8.7947	18.7533	0.1766	88	3	2.47
1529 (2874)	2863 (5145)	76.1	9.066	18.4499	0.1757	71	3	2.39
1545 (2813)	1947 (3537)	55.8	9.3208	18.1244	0.1746	74	3	2.47
819 (1506)	1194 (2181)	36.5	6.9654	24.8428	0.1939	28	2	7.39
1663 (3025)	3395 (6143)	102.2	9.8404	17.7808	0.1735	60	2	2.38

only slightly enriches one rare earth over another; thus, to isolate pure compounds of the individual elements by these methods, the processes have to be repeated many times.

Historically, the elements were purified by fractional processes such as fractional crystallization or fractional decomposition. The enormous amount of work that was involved permitted the separation of only very small amounts; consequently, the pure rare earths were very costly and gained the reputation of being rare. Fractionation methods are still used commercially to separate crude rare earths, particularly for lanthanum and cerium, since cerium can then be separated from lanthanum by taking advantage of the quadrivalent state of cerium. The other members of the rare-earth series, if desired in high purity, are separated by means of ion-exchange processes; if great purity is not desired, liquid-liquid extraction processes can be employed. Some of the rare earths which show anomalous valence are separated from the others by taking advantage of this property.

Yttrium and europium, used in the television industry, are usually separated by means of liquid-liquid extraction. Europium shows the divalent characteristic, and advantage can be taken of this. Yttrium, which occurs in another row in the periodic table, can be shifted either in liquid-liquid extraction or ion exchange along the rare-earth series by means of different complexing agents. Advantage can also be taken of the fact that some minerals contain very little of the heavy rare earths, but this fraction contains the yttrium, and liquid-liquid processes can be quite effective. Furthermore, when ton quantities are desired, the liquid-liquid extraction processes are usually more economical.

Uses. Most of the early uses of the rare earths took advantage of their common properties and were centered principally in the glass, ceramic, lighting,

TABLE 2. Some common rare-earth minerals

Mineral	Crystal form	Formula composition*
Monazite[†]	Monoclinic	$CePO_4$ with $Th_3(PO_4)_4$
Xenotime	Tetragonal	YPO_4
Gadolinite	Monoclinic	$2BeO \cdot FeO \cdot Y_2O_3 \cdot 2SiO_2$
Bastnasite[†]	Hexagonal	$CeFCO_3$
Samarskite	Orthorhombic	$3(Fe,Ca,UO_2)_3O \cdot Y_2O_3 \cdot 3(Nb,Ta)_2O_5$
Fergusonite	Tetragonal	$Y_2O_3 \cdot (Nb,Ta)_2O_5$
Euxenite	Orthorhombic	$Y_2(NbO_3)_3 \cdot Y_2(TiO_3)_3 \cdot 1^{1}/_2 H_2O$
Yttrofluorite	Cubic	$2YF_3 \cdot 3CaF_2$

*Only the most abundant rare earth is listed. Ce minerals are rich in light members of the rare-earth group. Y minerals are rich in heavy members of the rare-earth group.
[†]One of the two most important minerals used by the rare-earth industry in the United States.

and metallurgical industries. Today these applications use a very substantial amount of the mixed rare earths just as they are obtained from the minerals, although sometimes these mixtures are supplemented by the addition of extra cerium or have some of their lanthanum and cerium fractions removed.

Cerium oxide or the mixed rare-earth oxides make excellent abrasives for glass polishing, and several million pounds of these materials are sold each year for polishing mirrors, television face plates, lenses, and plate glass. The rare earths have long been used in the manufacture of specialty glasses and in the ceramic industry, particularly as glazes.

The elements exhibit very complex spectra, and the mixed oxides, when heated, give off an intense white light which resembles sunlight. Consequently mixtures of the oxides are used in cored carbon arcs, such as those employed in the movie industry.

The rare-earth metals have a great affinity for the nonmetallic elements, as, for example, hydrogen, carbon, nitrogen, oxygen, sulfur, phosphorus, and the halides. Considerable amounts of the mixed rare earths are reduced to metals, such as misch metal, and these alloys are used as "getters" in the metallurgical industry. The rare-earth metals combine with the nonmetals in other metals and alloys, and these impurities are partially removed along with the slag or, as in nodular cast iron, the rare earths convert the carbon inclusions into nodules, and this procedure imparts considerable strength to the iron. Alloys made of cerium and the mixed rare earths are used in the manufacture of lighter flints.

Rare earths are also used in the petroleum industry as catalysts. The rare-earth ions are introduced into zeolite-type oxide catalyst, say in the form of the chloride. The resulting catalyst, which contains about 5% mixed rare-earth oxides, produces petroleum with a higher yield of the desired petroleum fraction.

Since the individual rare earths have been available commercially at reasonable prices, a sizable market has developed for a number of them. This is not surprising, since they represent, as a group, about one-sixth of the naturally occurring elements and about one-fourth of the elemental metals. The fact that these elements form compounds and alloys, many of whose properties change in a predictable manner as one progresses across parts of the series, makes them of particular interest to scientists working in various fields of research. These scientists develop theories about numerous experimental phenomena, and they can use the rare-earth metals, alloys, compounds, and solutions as test materials to see whether the theories still apply when tested on a series of neighboring elements whose properties shift in a known manner.

The television industry has been using highly purified yttrium and europium oxides in large quantities. A europium-activated yttrium phosphor emits an intense red light when used in a television screen. Use of this type of phosphor greatly improves the color reproduction, and pictures are much more brilliant than when a non-rare-earth phosphor was used. The mercury-activated electric street lights, which are economical to operate, give a purplish-blue color; for about the same energy consumption they can be made to give a brilliant white light by the use of rare-earth phosphors deposited on the glass.

Another very important use of individual rare earths is in the manufacture of solid-state microwave devices widely used in radar and communications systems. Yttrium-iron garnets are especially good, since they transmit short-wave energy with low energy losses. The devices, however, are very small, so the total use of rare earths is not large.

Still another important use of individual rare earths is in the construction of lasers. A high percentage of the patent applications for new lasers involves the use of a rare earth as the active constituent of the laser.

The rare earths show interesting magnetic properties. Alloys of cobalt with the rare earths, such as cobalt-samarium, produce permanent magnets that are far superior to most of the varieties on the market, and many uses for these magnets are developing.

Many phosphors contain rare earths, and barium phosphate-europium phosphor finds applications in x-ray films that form satisfactory images with only half the exposure time of conventional x-rays.

Yttrium aluminum garnets (YAG) are used in the jewelry trade as artificial diamonds. Single crystals of YAG have a very high refractive index, similar to diamond, and when these crystals are cut in the form of diamonds, they sparkle in the same manner as a diamond. Also, they are very hard, so that they scratch

glass, and only an expert can tell the difference between a YAG diamond and a real one.

Other small but important uses are mentioned in the articles on the individual rare earths. (For the list of rare-earth elements see Table 1.)

Frank H. Spedding

Bibliography. R. J. Elliott (ed.), *Magnetic Properties of Rare Earth Metals*, 1972; K. Gschneidner and L. Eyring, *Handbook on the Physics and Chemistry of Rare Earths*, vol. 1: *Metals*, vol. 2: *Alloys and Intermetallics*, 1978; C. M. Lederer et al., *Table of Isotopes*, 6th ed., 1967; J. E. Powell et al., The separation of rare earths, *J. Chem. Educ.*, 37:629–633, 1960; F. H. Spedding and A. H. Daane (eds.), *Rare Earths*, 1961.

Rare-earth minerals

Naturally occurring solids, formed by geological processes, that contain the rare-earth elements—the lanthanides (atomic numbers 57–71) and yttrium (atomic number 39)—as essential constituents. In a rare-earth mineral, at least one crystallographic site contains a total atomic ratio of lanthanides and yttrium that is greater than that of any other element. The mineral name generally has a suffix, called a Levinson modifier, indicating the dominant rare-earth element; for example, monazite-(La) [$LaPO_4$] contains predominantly lanthanum, and monazite-(Ce) [$CePO_4$] contains predominantly cerium. *See* MINERAL; MONAZITE; PERIODIC TABLE; RARE-EARTH ELEMENTS.

So far, about 170 distinct species of rare-earth minerals have been described. A large number of carbonates, phosphates, silicates, niobates, and fluorides are known as rare-earth minerals. However, no sulfide of the rare earths has been reported as a mineral species. It is necessary to obtain structural as well as chemical information about a mineral to judge the essentiality of its rare-earth elements (that is, whether the rare-earth element is part of the mineral's ideal formula or is basically an impurity). Sometimes, minerals with significant rare-earth content are treated as rare-earth minerals, even if the rare-earth element content appears unessential to the mineral. More than 60 mineral species, including the apatite group minerals, garnet group minerals, and fluorite, are in this category. *See* APATITE; FLUORITE; GARNET.

Occurrence. Despite their name, rare-earth elements are not extremely rare in the Earth's crust. Their abundance is not as low as the rarest elements such as gold and mercury. Rare-earth elements are contained in many common rock-forming minerals as trace impurities, and are concentrated by geological activities to form rare-earth minerals. The concentration depends on many factors such as the relative content of the rare-earth elements in source melts or fluids, oxidation states, temperature, pressure, and fit of ionic radius to developing crystal structure. Rare-earth minerals can be observed \pagebreak as accessory minerals in igneous rocks, such as monazite-(Ce) in granite. Carbonatite is the typical host rock of rare-earth minerals such as bastnäsite-(Ce) [$CeCO_3F$] and monazite-(Ce). Rare-earth minerals also often occur in pegmatite. In both carbonatite and pegmatite, rare-earth elements are concentrated by primary crystallization from melt and by hydrothermal reactions. Carbonatite deposits containing rare-earth elements are found throughout the world in such places as Mountain Pass, California, and Bayan Obo, China. Chemically stable, rare-earth minerals are not weathered easily. As a result, they have been deposited as heavy minerals in beach sand. Such deposits are found in Southeast Asia and Western Australia. *See* CARBONATITE; PEGMATITE.

Among the rare-earth minerals, bastnäsite-(Ce) is the most important source of rare-earth elements. Monazite-(Ce), synchysite-(Ce) [$CaCe(CO_3)_2F$], xenotime-(Y), britholite-(Ce) [$(Ce_3Ca_2)(SiO_4)_3(OH)$], and allanite-(Ce) [$CaCeAl_2Fe(Si_2O_7)(SiO_4)O(OH)$] are also sources. Rare-earth elements have been leached with acid from the surface of clay minerals. Rare-earth minerals containing radioactive nuclear species, such as thorium and uranium, are not used as source materials. *See* CLAY MINERALS.

Chemistry. The rare-earth minerals always contain a variety of rare-earth elements. The study of the rare-earth elemental distribution in minerals and rocks has been an active subject in geochemistry and mineral chemistry for many years. Elemental distribution follows Oddo-Harkins's rule: in natural minerals, a lanthanide with an even atomic number is more abundant than the neighboring lanthanides with odd atomic numbers. *See* GEOCHEMISTRY; MINERALOGY.

Rare-earth minerals can be classified into the following three groups based on their distribution of rae-earth elements: those with a predominance of the Y-group rare-earth elements, such as yttrium, having relatively small ionic radii; those rich in the Ce-group rare-earth elements, such as cerium, having larger ionic radii; and those in which either Y-group or Ce-group rare-earths elements may predominate (for example, gadolinite-(Y) [$Y_2FeBe_2Si_2O_{10}$] and gadolinite-(Ce) [$Ce_2FeBe_2Si_2O_{10}$]).

Rare-earth elements are usually trivalent in minerals. Notable exceptions are the divalent europium (Eu^{2+}) and tetravalent cerium (Ce^{4+}) ions. Crystal structures have been reported for about half of the known rare-earth minerals. The structure analyses have revealed that many of them are solid solutions of rare-earth ions and other heterovalent ions, such as sodium (Na^+), calcium (Ca^{2+}), and thorium (Th^{4+}), with ionic radii comparable to those of rare-earth ions. It is known that the similarity of ionic radii and electronegativities are important factors for isomorphous substitutions (substitution of similar-sized ions within crystal structures), whereas the similarity of chemical properties is not as significant. Studies of isomorphous substitution in rare-earth minerals are important for understanding the behavior of rare-earth elements in solids. *See* ELECTRONEGATIVITY.

Crystal structure. Crystallographic studies are impeded by the fact that some rare-earth minerals always occur in the metamict state, an amorphous (glasslike) state mainly caused by radiation damage to crystal structures due to radioactive decay of elements such as uranium and thorium. Crystallographic data and crystal structures of metamict minerals are therefore determined using samples recrystallized by annealing. In such cases, there is some doubt about the correlation between the crystal structures of recrystallized phases and the original structures of pre-metamict minerals. *See* METAMICT STATE; RADIOACTIVE MINERALS.

Anionic groups. Rare-earth minerals fall into six classes based on the anionic groups in their crystal structure:

1. Minerals with crystal structures containing isolated triangular anionic groups—carbonates such as lanthanite-(Nd) $[Nd_2(CO_3)_3 \cdot 8H_2O]$ and bastnäsite-(Ce) $[Ce(CO_3)F]$.

2. Minerals with crystal structures containing only isolated tetrahedral anionic groups—phosphates such as xenotime-(Y) $[YPO_4]$ and monazite-(Ce) $[CePO_4]$.

3. Minerals with crystal structures containing linked anionic groups of tetrahedral ions—silicates such as thalenite-(Y) $[Y_3Si_3O_{10}(OH)]$ and yttrialite-(Y) $[Y_2Si_2O_7]$.

4. Minerals with crystal structures containing tetrahedral and octahedral ions—aluminum silicate and titanium silicate such as allanite-(Ce) $[CaCeAl_2Fe(Si_2O_7)(SiO_4)O(OH)]$ and mosandrite $[Na(Na,Ca)_2(Ca,Ce)_4Ti(Si_2O_7)_2(O,F)_2F_2]$.

5. Minerals with crystal structures containing anionic groups of octahedral ions—titanates and niobates such as aeschynite-(Ce) $[CeTi_2(O_5OH)]$ and fergusonite-(Y) $[YNbO_4]$.

6. Minerals with crystal structures without anionic groups—fluorides and simple oxides such as fluocerite-(Ce) $[CeF_3]$ and cerianite-(Ce) $[CeO_2]$.

Isomorphous substitution. The coordination numbers of the rare-earth atoms are generally high, and their coordination polyhedra are as large as those of sodium ions, calcium ions, thorium ions, and so on. Y-group rare-earth elements and Ce-group rare-earth elements show a difference in their coordination numbers. The coordination numbers of the Y-group range from 6 to 11. Those of the Ce-group range from 7 to 12. The coordination numbers of calcium resemble those of the Ce-group; hence isomorphous substitution between calcium and Ce-group rare-earth elements has been observed often.

The charge compensation mechanisms for mutual substitution of heterovalent ions found in the structures of rare-earth minerals have been classified into the following four types (where RE represents a generic rare-earth element):

1. Substitution accompanying vacancies (denoted by □).

$$3Ca^{2+} \rightleftharpoons 2RE^{3+} + \square$$
$$RE^{3+} + \square \rightleftharpoons Ca^{2+} + Na^+$$
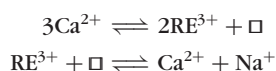
2. Coupled substitution within a rare-earth site.

$$2RE^{3+} \rightleftharpoons Ca^{2+} + Th^{4+}$$
$$2Ca^{2+} \rightleftharpoons RE^{3+} + Na^+$$
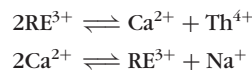
3. Coupled substitution at two independent sites.

$$RE^{3+} + Na^+ \rightleftharpoons Ca^{2+} + Ca^{2+}$$
$$RE^{3+} + Si^{4+} \rightleftharpoons Ca^{2+} + (P,As)^{5+}$$
$$RE^{3+} + Ti^{4+} \rightleftharpoons Ca^{2+} + (Nb, Ta)^{5+}$$
$$RE^{3+} + Mg^{2+} \rightleftharpoons Ca^{2+} + Al^{3+}$$
$$RE^{3+} + (OH)^- \rightleftharpoons Ca^{2+} + H_2O$$
$$RE^{3+} + O^{2-} \rightleftharpoons Ca^{2+} + (OH, F)^-$$
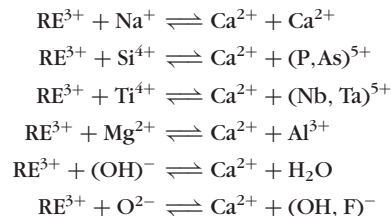
4. Substitution accompanying valence variation.

$$RE^{3+} + Fe^{2+} \rightleftharpoons Ca^{2+} + Fe^{3+}$$

In almost all rare-earth minerals, complex solid solutions result from these isomorphous substitutions. Type 1, accompanying a vacancy site, may occur only in crystal structures having tight frameworks formed by covalent bonding, because the number of atoms changes before and after the substitution. Type 2 is a simple one occurring at an independent site. These are the most common substitutions observed in the crystal structures of rare-earth minerals because they do not require any other concomitant substitution. The complex solid solutions formed by types 3 and 4 make it especially difficult to study their chemical compositions and crystal structures.

Polyhedra shape. The shapes of the coordination polyhedra of rare-earth atoms are, in almost all cases, not regular polyhedra, but distorted polyhedra because the bonds between the rare-earth atoms and anions are ionic rather than covalent and their coordination numbers are high. Considering the distortion of polyhedra, the coordination polyhedra composed of rare earths and surrounding anions can be expressed as an octahedron, a monocapped trigonal prism, a cube, a dodecahedron, a square antiprism, an icosahedron, and so on.

The differences in coordination numbers and sizes of coordination polyhedra between the Y- and Ce-groups are reflected in the crystal structures of carbonates and phosphates. These minerals do not have infinite framework structures, and the isolated anionic groups in the structures are connected by rare-earth atoms to form three-dimensional structures. Therefore, the size of the rare earth–polyhedron affects the whole structure. On the other hand, structures having a tight framework, as seen in the structures of gadolinite-(Y), allanite-(Ce), and aeschynite-(Y), are not affected by the sizes of the rare earth–polyhedra, because they can accept rare earths of various sizes without rearrangement of the structure. *See* CRYSTAL STRUCTURE; CRYSTALLOGRAPHY.   Ritsuro Miyawaki

Bibliography. K. A. Gschneidner, Jr., and L. Eyring (eds.), *Handbook on the Physics and Chemistry of Rare Earths*, vol. 16, Elsevier Science, Amsterdam, 1993; A. P. Jones, F. Wall, and C. T.

Williams (eds.), Rare Earth Minerals: Chemistry, Origin and Ore Deposits, *Mineralogical Society Series*, no. 7, Chapman & Hall, London, 1996; B. R. Lipin and G. A. McKay (eds.), Geochemistry and Mineralogy of Rare Earth Elements, *Reviews in Mineralogy*, vol. 21., Mineralogical Society of America, Washington DC, 1989.

Rarefied gas flow

Flow of gases below standard atmospheric pressure, sometimes called low-pressure gas flow. The flow may be confined to pipes between a chamber or vessel to be evacuated and a pump, or it may be the beam of molecules issuing from an orifice into a large evacuated chamber or the plume of exhaust gases from a rocket launched into the upper atmosphere, for example. The flow velocity is measured with respect to a fixed boundary such as the wall of a pipe, the surface of a rocket or jet plane, or a model in a wind tunnel. *See* FLUID FLOW; GAS; MOLECULAR BEAMS; PIPE FLOW; VACUUM PUMP; WIND TUNNEL.

For flow through ducts, the gases concerned are initially those of the original atmosphere inside a chamber that must be evacuated. However, even after the bulk of the original gas has been removed, the pumps must continue to remove gas evolved from surfaces and leaking in through imperfections in the walls. In some cases, gas is introduced through valves at a controlled rate as part of the process being carried out at a low pressure.

Since the flow through pipes involves an interaction or drag at the walls, a pressure drop is generated across the entrance and exit of the pipe. Also, gaseous impurities from the pump may flow toward the chamber when the pressure is very low. Proper design of the duct system therefore involves selecting pipes and valves of adequate internal diameter to ensure a minimal pressure drop and the insertion of baffles or traps to prevent impurities from the pumps from entering the process chamber.

Flow regimes. The resistance due to the walls depends on the mass flow velocity v, and may depend on the gas viscosity η and the pressure P or density ρ of the gas. The mean free path λ of molecules between collisions with other molecules in the gas can be calculated from Eq. (1), where v_m is the

$$\lambda = \frac{2\eta}{\rho v_m} \qquad (1)$$

mean molecular speed in a gas obeying the Maxwell-Boltzmann distribution law in a reference frame moving with the stream velocity v. The speed v_m is given by Eq. (2), and the density ρ is given by the perfect gas law, Eq. (3). Here, M is the molecular weight

$$v_m = \left(\frac{8R_0T}{\pi M} \right)^{1/2} \qquad (2)$$

$$\rho = \frac{MP}{R_0T} \qquad (3)$$

(in kg/mol in SI units, and in g/mol in cgs units);

P is the pressure; R_0, given by Eq. (4), is the molar

$$R_0 = 8.314 \frac{J}{mol \cdot K} \qquad \text{(SI units)}$$

$$= (8.314 \times 10^7) \frac{erg}{mol \cdot K} \qquad \text{(cgs units)} \quad (4)$$

gas constant; and T is the absolute temperature. *See* KINETIC THEORY OF MATTER; MEAN FREE PATH; VISCOSITY.

The analysis of low-pressure flow is divided into three or four flow regimes depending on the value of the Knudsen number Kn defined as the ratio of the mean free path λ to a characteristic length d [Eq. (5)],

$$Kn = \frac{\lambda}{d} \qquad (5)$$

and the dimensionless Reynolds number, defined by Eq. (6). The characteristic length d may be chosen

$$Re = \frac{v\rho d}{\eta} \qquad (6)$$

as the mean pipe diameter in the case of confined flow or as some length associated with a test model suspended in a wind tunnel, for example. *See* GAS DYNAMICS; KNUDSEN NUMBER; REYNOLDS NUMBER.

Another dimensionless number used in gas flow dynamics is the Mach number, defined as the ratio of the mass flow velocity v to the local velocity of sound v_a in the gas. It is given by Eq. (7), in which γ

$$Ma = \frac{v}{v_a} = \frac{2v}{v_m} \left(\frac{2}{\pi\gamma} \right)^{1/2} \qquad (7)$$

is the ratio of the specific heat at constant pressure to the specific heat at constant volume. When the gas molecules are diffusely scattered from the walls, these numbers are related by Eq. (8).

$$Kn = \left(\frac{\pi\gamma}{2} \right)^{1/2} \frac{Ma}{Re} = \frac{\lambda}{d} \qquad (8)$$

See MACH NUMBER.

When the mean free path is much smaller than the pipe diameter (Kn < 0.01), the gas flows as a continuous viscous fluid with velocity near the axis of the pipe at locations well beyond the pipe entrance much higher than the velocity in gas layers near the wall. The velocity profile as a function of radial distance from the axis depends on the distance from the entrance and the viscosity of the gas. When the Reynolds number is less than 2000, the profile is a simple curved surface so that the flow is laminar (laminar flow regime). When the mean free path becomes greater than about 0.01 times the diameter, the profile is distorted by boundary-layer effects, and the velocity near the wall does not approach zero (sometimes referred to as slip flow). *See* LAMINAR FLOW.

For Reynolds numbers above the critical value (approximately 2100), the flow is subject to instabilities depending on the geometry of the boundary and at high Reynolds numbers becomes turbulent (turbulent flow regime). *See* TURBULENT FLOW.

When the mean free path is about equal to or greater than the pipe diameter (Kn ≥ 1), the gas molecules seldom collide with each other, but can either pass through the pipe without striking the wall or scatter randomly back and forth between various points on the wall and eventually escape through the exit or pass back through the entrance. This type of gas flow is known as free-molecule flow (molecular flow regime). The transition region (0.01 < Kn < 1) between the laminar flow regime and the molecular flow regime is referred to as the Knudsen or transition flow regime.

The flow may also be classified by the boundary conditions or by the Mach number. For example, Couette flow involves the flow of rarefied gas between two surfaces that are moving with respect to each other with different parallel tangential velocities. For hypersonic flow, Ma ≥ 5.

Vacuum system pressure. Pressure may be expressed by the SI unit, the pascal (Pa; one newton of force per square meter), or the commonly used unit in vacuum technology, the torr (approximately 133 Pa). Standard atmospheric pressure is 101,325 Pa. Pressure measurement usually involves inserting the tubulation of a vacuum gage through the wall of a vacuum system with the plane of the opening parallel to the axis of flow so that so-called static pressure is measured, but ionization-type vacuum gages are sometimes installed directly inside the walls with no envelope or tubulation so that gas molecules can arrive from all directions. Such so-called nude gages then measure gas density or molecular concentration rather than pressure. When the temperature varies widely over different parts of a vacuum system, it is usually preferable to analyze the gas flow in terms of mass flow or molecular flux and the average number of molecules in unit volume. *See* PRESSURE MEASUREMENT; VACUUM MEASUREMENT.

Many applications require a high vacuum, corresponding to pressures less than 0.1 Pa. These usually involve a process chamber connected by a short duct of large cross section equal to or greater than the intake area of a high-vacuum pump, such as a diffusion pump, turbomolecular pump, or cryopump. Diffusion pumps and turbomolecular pumps require a forepressure below a limiting value and therefore must be connected by a pipe to a backing pump or forepump, such as a mechanical rotary-piston pump or rotary blower pump (**Fig. 1**). The connecting pipe can be relatively long and of small diameter since the mean pressure in the foreline is generally high enough for the flow to be viscous and the pressure drop is small up to the entrance of the backing pump.

Pumping speed. Since most high-vacuum pumps and mechanical forepumps have a broad range of intake pressures over which the volumetric flow rate of gas is approximately constant, it is customary to express the quantity of gas flowing into the pump in pressure–volume units (at room temperature) rather than mass units. In designing vacuum systems, it is essential to know the maximum volu-
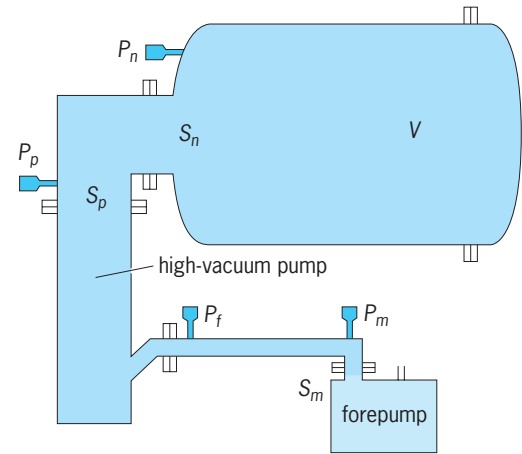
Fig. 1. Vacuum system consisting of a process chamber of volume *V* at pressure P_n, high-vacuum pump of speed S_p, and forepump of speed S_m with connecting pipes. $P_p =$ pressure near entrance to high-vacuum pump; $P_f =$ discharge pressure (forepressure) at exit of high-vacuum pump; $P_m =$ intake pressure of forepump; $S_n =$ flow rate of gas from the process chamber through the opening to the pipeline.

metric flow rate at room temperature, or pumping speed S_0, for each pump. The speed is measured by the manufacturer by bleeding air (or other gases) through a calibrated control valve into a chamber (or test dome) placed directly over the intake entrance and measuring the pressure near the entrance according to an industry standard. The actual speed S_p varies with the absolute temperature T and molecular weight M of the gas. For high-vacuum pumps it can usually be assumed that Eq. (9) holds, where S_0 is the speed for air at 296 K (73°F).

$$S_0 = \left(\frac{28.98\,T}{296\,M} \right)^{1/2} \tag{9}$$

Common units for pumping speed are liters per second (L/s or l/s); cubic feet per minute (usually abbreviated cfm; sometimes ft³/min); or the SI unit, cubic meters per second (m³/s).

Throughput. The quantity of gas in pressure–volume units, at a specified temperature, flowing through a cross section of the pipe in unit time is called the throughput. Unless otherwise specified, it is understood that the prevailing gas temperature is that of the walls near the point at which the pressure is measured; in most cases this will be room temperature. For steady flow through a pipe with walls at a uniform temperature, the throughput is a conserved quantity when no adsorption or desorption occurs on the walls.

When the pressure P_p in the gas near the entrance to the pump is measured under conditions similar to those in the standard test dome used to measure pump speed S_p for the same gas, the throughput is given by Eq. (10).

$$Q = S_p P_p \tag{10}$$

Most gases at low pressures obey the ideal gas law of Eq. (3), so that the mass rate of flow in g/s

is given by Eq. (11), where the numerical value of

$$Q_m = Q\frac{M}{R_v T} \qquad (11)$$

the molar gas constant R_v is 62.36 when Q is in torr · L/s, 8.314×10^3 when Q is in Pa · L/s, or 8.314 when Q is in Pa · m³/s.

Conductance and resistance. The pressure P_n in the process chamber depends on the resistance to flow in the pipeline between the chamber and the pump. This resistance W_{np} is defined by Eq. (12), where P_p

$$Q = \frac{P_n - P_p}{W_{np}} = C_{np}(P_n - P_p) \qquad (12)$$

is the pressure at the junction with the pump and C_{np} is the conductance, or reciprocal of the resistance. Similarly, for the pipe between a high-vacuum pump and a mechanical forepump of speed S_m at the intake pressure P_m, the steady-state throughput is given by Eq. (13), where P_f is the discharge pres-

$$Q = C_{fm}(P_f - P_m) = S_m P_m \qquad (13)$$

sure (or forepressure) at the exit of the high-vacuum pump and C_{fm} is the conductance of the pipe between the points at which P_f and P_m are measured (Fig. 1).

The conductance C_{np} is independent of internal pressure since the flow occurs in the molecular flow regime, but the flow in the line between high-vacuum pump and forepump may be in the transition regime for certain values of Q and will be in the viscous flow regime during the initial pump-down from atmospheric pressure. The conductance of a pipe in the viscous or transition flow regime increases with the mean pressure.

Net pumping speed. The volumetric flow rate of gas from the chamber of volume V through the opening to the pipeline at the system temperature is known as the net speed, S_n as given by Eq. (14), where P_n is the

$$S_n = \frac{Q_0}{P_n} - \frac{V}{P_n}\frac{dP_n}{dt} \qquad (14)$$

pressure in the chamber, and Q_0 is the quasi-steady-state gas load due to outgassing, leaks, and so forth, in throughput units, which can be regarded as constant during the time required for initial evacuation. When dP_n/dt is small, equations of continuity can be written as Eq. (15), and Eq. (16) follows. Equa-

$$Q = S_n P_n = C_{np}(P_n - P_p) = S_p P_p$$
$$= C_{fm}(P_f - P_m) = S_m P_m \qquad (15)$$

$$\frac{1}{S_n} = \frac{1}{C_{np}} + \frac{1}{S_p} \qquad (16)$$

tions (12)–(16) are valid only for steady-state flow and uniform temperature throughout the system, or when the variables have been referred to standard room temperature by correction factors.

Molecular flow regime. A conductance C_{12} can be defined in terms of the pressure drop, $P_1 - P_2$, across any section of a complex pipe system, but depends on the geometry of the system at the points where P_1 and P_2 are measured. If a section of the pipeline of length L and uniform cross section is located between two relatively large chambers in which the pressures, as measured some distance from the pipe entrance or exit planes, are respectively P_1 and P_2, then the molecular flow conductance of the pipe section will include end corrections.

The molecular flow conductance can be calculated from Eq. (17). Here, f is the effusion law factor

$$C_{12} = \frac{Q}{P_1 - P_2} = fA_1 w_{12} \qquad (17)$$

for molecular flow through an orifice in a thin plate, given by Eq. (18) [where Eq. (2) has been used to ex-

$$f = \frac{v_m}{4} = \left(\frac{R_0 T}{2\pi M}\right)^{1/2} \qquad (18)$$

press v_m]; A_1 is the entrance area; and w_{12} is defined as the probability that a molecule entering the pipe will cross the exit and not return to the entrance chamber.

The transmission probability w_{12} for a circular pipe of radius R between two large chambers is given approximately by Eq. (19), where L' is an end-

$$w_{12} = \frac{L'}{L + L'} \qquad (19)$$

correction term, which varies from $L' = \frac{8}{3}R$ for pipes with $L/R > 100$ to $L' = 2R$ for L/R approaching zero. More exact equations for w_{12} for straight pipes of circular (and some other) cross sections have been derived. The transmission probability factor w for straight circular tubes between large chambers is plotted as a function of L/R in **Fig. 2**.

The overall transmission probability w_{13} for gas from a large chamber entering a pipeline consisting of a first pipe of uniform cross section A_1 connected in series by a short transition piece, or by an annular junction, to a second pipe of uniform cross section A_2
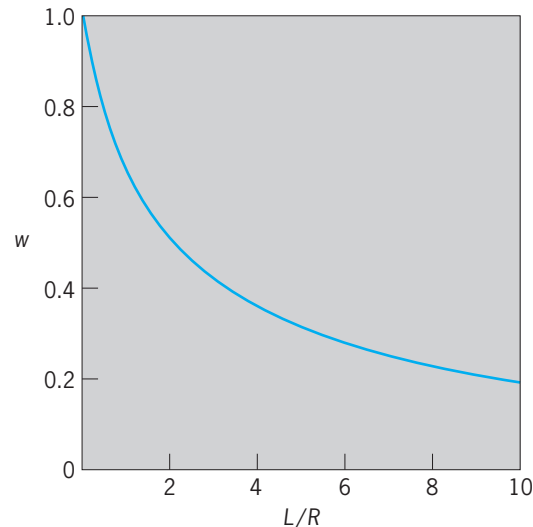
Fig. 2. Molecular flow transmission probability *w* for straight cylindrical tube of length *L* and radius *R*.

exiting into a large chamber is given approximately by Eq. (20). Here, w_{12} is the transmission probabil-

$$\frac{(1/w_{13}) - 1}{A_1} = \frac{(1/w_{12}) - 1}{A_1} + \frac{(1/w_{23}) - 1}{A_2} \quad (20)$$

ity through the first pipe, calculated as though the pipe were located by itself between two large chambers, and w_{23} is similarly the transmission probability for the second pipe alone. This equation can be extended to any number of different pipe sections in series.

The molecular flow transmission probability for pipes with elbows and for ducts with baffles or complex shapes can be obtained from Monte Carlo computer programs for the trajectory of molecules crossing the entrance. *See* MONTE CARLO METHOD.

For an orifice of area A_1 in a thin plate between two chambers of cross section much larger than A_1, the value of w_{12} is approximately equal to 1, corresponding to $L = 0$ in Eq. (19).

In common units, when L is much greater than L', the conductance (in L/s) of a straight circular tube of radius R (in cm), length L (in cm), and effusion law factor f (in cm/s) is given by Knudsen's formula, Eq. (21). For air at $23°C$ ($73°F$), the effusion law factor $10^{-3}f = 11.63$ L/(s · cm²).

$$C_{12} = \frac{{}^8/_3 (10^{-3}f)\pi R^3}{L} \quad (21)$$

Laminar flow regime. For long straight tubes ($L > 20R$) of length L (in cm) and uniform radius R (in cm) in the laminar flow regime, the conductance (in L/s) is given by the Poiseuille equation (22),

$$C_{fm} = \frac{(\pi R^4/8\eta L)P_{fm}}{100} \quad (22)$$

where P_{fm} is the mean pressure in pascals, given by Eq. (23), and η is the viscosity in poise. For short tubes a correction factor must be introduced.

$$P_{fm} = \frac{P_f + P_m}{2} \quad (23)$$

Transition flow regime. The Knudsen semiempirical formula for conductance (in L/s) of a long circular tube of length L (in cm) and radius R (in cm) in the transition regime is Eq. (24), where C_{fm} is the

$$C_t = C_{fm} + F_k C_{12} \quad (24)$$

Poiseuille flow term given by Eqs. (22) and (23), C_{12} is the molecular flow conductance given by Eq. (21), and F_k is given by Eq. (25). Here, f is given by

$$F_k = \frac{1 + (2/\pi)^{1/2}R(10P_{fm}/f\eta)}{1 + 1.24(2/\pi)^{1/2}R(10P_{fm}/f\eta)} \quad (25)$$

Eq. (18) in centimeters per second, and P_f by Eq. (23) in pascals. 　　　　　　　Benjamin B. Dayton

Bibliography. G. A. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, 1994; W. Fiszdon, *Rarefied Gas Flows*, 1982; E. P. Muntz, D. H. Campbell, and D. P. Weaver (eds.), *Rarefied Gas Dynamics*, 3 vols., 1994; H. Oguchi, *Rarefied Gas Dynamics*, 2 vols., 1985; A. Roth, *Vacuum Technology*, 3d ed., 1990, reprint 1998; G. L. Saksaganskii, *Molecular Flow in Complex Vacuum Systems*, 1988.

Raspberry

The horticultural name for certain species of the genus *Rubus*, plant order Rosales. In these species the fruit, when ripe (unlike the blackberry), separates thimblelike from the receptacle. Raspberry plants are upright shrubs with perennial roots and prickly, biennial canes (stems). There are several species, both American and European, from which the cultivated raspberries have been developed. Varieties are grouped as to color of fruit—black, red, and purple, the last being hybrids between the red and black types. Red raspberries, with upright canes, are propagated by suckers or by root cuttings (**Fig. 1**);

Fig. 1. Red raspberry branch, Loudon variety. (*USDA*)

the black varieties, with long canes which arch over and touch the ground, are propagated by tip layers. The hybrid purple varieties are usually progagated by tip layers. Raspberry breeding has been carried on so extensively that all the important red and purple and most of the black varieties are the result of breeding experiments. Raspberries are grown extensively in home gardens over most of the United States. Leading states in commercial production are Michigan, Oregon, New York, Washington, Ohio, Pennsylvania, New Jersey, and Minnesota. The fruit is sold fresh for dessert purposes, is canned, and is made into jelly or jam, but quick freezing is the most important processing method. *See* BREEDING (PLANT); ROSALES.
　　　　　　　　　　　　　　　　　J. Harold Clarke

Raspberry plants are seriously damaged by diseases caused by viruses and viruslike organisms, bacteria, fungi, and nematodes.

Raspberry mosaic and raspberry leaf curl are spread by aphids and are common in most areas

Fig. 2. Anthracnose disease on canes (stems) of black raspberry.

in the United States where raspberries are grown. Tomato ringspot virus, spread by dagger nematodes (*Xiphinema* sp.), and pollen-spread raspberry bushy dwarf virus weaken plants and cause crumbly fruit. The use of resistant cultivars, certified stocks, isolation from infected plantings, and nematode control (for tomato ringspot virus) are recommended controls.

Crown gall caused by *Agrobacterium tumefaciens* is characterized by tumorlike outgrowths at ground level on infected raspberry plants. Certified stocks should be planted where fruit crops have been grown. *See* CROWN GALL.

Fungi cause diseases of raspberry leaves, canes, roots, fruits, and of the whole plant. Raspberry leaf spot, caused by *Sphaerulina rubi*, causes whitish spots on leaves. Powdery mildew, caused by *Sphaerotheca humuli*, involves entire young leaves in dusty white growth. Anthracnose, caused by *Elsinoë veneta* (**Fig. 2**), spur blight, caused by *Didymella applanata*, and yellow rust, caused by *Phragmidium rubi-idaei*, can produce spots on leaves and cankers on canes. *Leptosphaeria coniothyrium* causes cane blight when it gains entrance into tissues through wounds. Resistant cultivars, proper pruning, and approved fungicides help control these diseases. Verticillium wilt caused by *Verticillium dahliae* is a root-infecting fungus that produces wilting and death of canes. Avoidance of land previously planted with solanaceous crops (like potatoes) helps control this disease. Red raspberry fruits are susceptible to gray mold fruit rot caused by *Botrytis cinerea*, and less commonly to rots by other fungi including *Alternaria, Cladosporium,* and *Rhizopus* sp. Prompt picking and marketing, cool handling, and application of approved fungicides aid in fruit rot control. Orange rust, caused by *Gymnoconia peckiana*, systemically infects black raspberry plants. Growing points are invaded, and the fungus produces abundant orange pustules on leaves at harvest time. Use of certified planting stock controls the disease. Red raspberries are immune. *Pratylenchus* and *Xiphinema* are the most impor-

tant plant-parasitic nematodes that feed on raspberry roots and weaken the plants. *Xiphinema* also spreads certain viruses. Control is achieved by using certified planting stock and by not planting in soil with high population of these nematodes. *See* PLANT PATHOLOGY.   Richard H. Converse

Bibliography. M. A. Ellis et al., *Compendium of Raspberry and Blackberry Diseases and Insects*, 1991; N. W. Frazier, *Virus Diseases of Small Fruits and Grapevines*, USDA Farmers Bull. 2208, 1978.

Ratchet

A wheel, usually toothed, operating with a catch or a pawl so as to rotate in a single direction (see **illus.**).
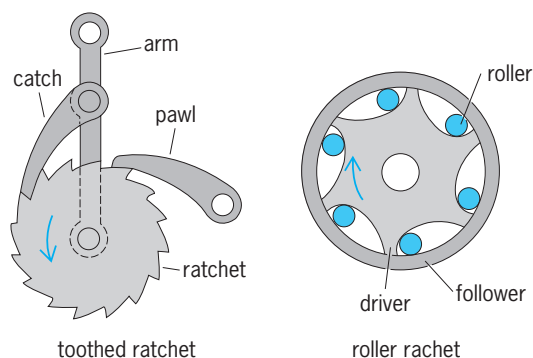
Toothed ratchet is driven by catch when arm moves to left; pawl holds ratchet during return stroke of catch. In roller ratchet, rollers become wedged between driver and follower when driver turns faster than follower in direction of arrow.

A ratchet and pawl mechanism locks a machine such as a hoisting winch so that it does not slip. The locking action may serve to produce rotation in a desired direction and to disengage in the undesired direction as in a drill brace. A further adaptation is to drive the catch in a to-and-fro motion against the ratchet to produce intermittent circular motion. The catch or pawl may be of various shapes such as an eccentrically mounted disk or ball bearing. Gravity, a spring, or centrifugal force (with the catch mounted internal to the ratchet) are commonly used to hold the pawl against the ratchet. A ratchet and pawl provides an arresting action, whereas an escapement provides an arresting action followed by a self-initiated momentary release. In high-speed machines, the abrupt action of a toothed ratchet produces severe shock. In such situations, a continuously variable yet directionally sensitive action is achieved by wedging rollers or specially shaped sprags between the input and output members. *See* BRAKE; ESCAPEMENT; PAWL.
                Frank H. Rockett

Bibliography. R. H. Creamer, *Machine Design: Mechanical Engineering Technology*, 3d ed., 1984; A. Esposito and J. R. Thrower, *Machine Design*, 2d ed., 1991; K. Lingaiah, *Machine Design Data Handbook*, 2d ed., 2001.

Rate-of-climb indicator

An aircraft instrument that provides an indication of the vertical change of the aircraft position within the air mass. It is more commonly known as the vertical-velocity or vertical-speed indicator. Contained within a sealed case, it is connected to the aircraft static pressure source, the so-called at-rest air pressure outside the aircraft, through a calibrated leak (see **illus.**). Although the instrument operates

(a)

(b)
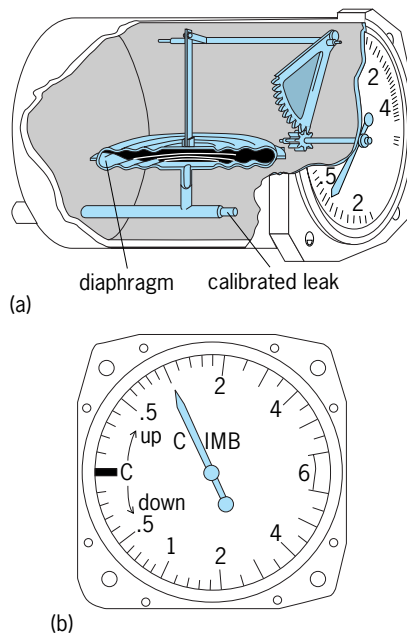
Rate-of-climb indicator or vertical-speed indicator. (*a*) Components. (*b*) Face. (*After Federal Aviation Administration, Airframe and Powerplant Mechanics Airframe Handbook, AC 65-15A, 1976*)

from a static pressure source, it is a differential pressure indicator. The differential pressure is established between the static pressure in the diaphragm or pressure capsule and the trapped static pressure within the case. When the aircraft changes vertical position, the static pressure in the diaphragm changes immediately but, because of the metering action of the calibrated leak, the case pressure will remain at its prior value and cause the needle to show a change in vertical speed. The needle is usually calibrated in feet per minute but may be calibrated in any appropriate unit of length over time. Because of the calibrated leak in the system, a finite time (usually 6–9 s) is required for the pressure inside the instrument to equalize and a reliable rate indication to become available.

This lag induced by the calibrated leak has led to the introduction of a more reliable instrument known as the instantaneous vertical-speed indicator. It incorporates acceleration pumps to eliminate the limitations associated with the calibrated leak. During climb entry, the vertical acceleration causes the pumps to supply extra air into the diaphragm to stabilize the pressure differential without the usual lag time. The opposite phenomenon occurs during

descent. During stabilized climbs and descents or level flight, the instrument operates on the same principles as the vertical-speed indicator. Both the vertical-speed indicator and instantaneous vertical-speed indicator usually incorporate a small screw on the lower left face of the instrument to allow ground adjustment to zero rate of climb.

In many modern aircraft with flight computers, the rate of climb or descent is electronically calculated by differentiating the altitude from the pitot-static source. *See* AIRCRAFT INSTRUMENTATION; PITOT TUBE.   Grady W. Wilson

Ratites

A group of flightless, mostly large, running birds characterized by their flat, keelless sternum, which were formerly segregated as a superorder of birds, the Palaeognathae, but whose interrelationships have been a long-standing controversy. Species include the ostriches (**Fig. 1**), emus, cassowaries, and rheas (**Fig. 2**). Traditionally the flying tinamous, in spite of their also possessing the typical palaeognathous palate, were not included in this group despite their close similarities to the rheas. Ornithologists, stressing the disjunctive distribution of these flightless birds and the many possible convergent features due to their large size, have concluded that the several groups of ratite birds represent end points of unrelated phyletic lineages from several different ancestral avian stocks and hence should be placed in separate orders. Comparative morphologists, on the other hand, stressing the basic similarity of the bony

Fig. 1. Ostrich (*Struthio camelus* spp. *massaicus*). (*Gerald and Buff Corsi © 2002 California Academy of Sciences*)

Fig. 2. Greater rhea (*Rhea americana*). (*Dr. Lloyd Glenn Ingles* © *2001 California Academy of Sciences*)

palate and other anatomical features, have concluded that the ratites are a strictly monophyletic group of birds. The latter view is supported by evidence that the structure of the bony palate, type of cranial kinesis, and other cranial features are homologous in all ratites in spite of their other differences. Additional evidence such as general behavior, structure of the rhamphotheca, and some biochemical features also support this conclusion.

The ratites apparently represent two or three phyletic lineages [the African ostriches and elephant birds; the Australasian emus, cassowaries, moas and kiwis; and the Neotropical rheas and tinamous, the Australasian and Neotropical groups being perhaps more closely related to one another than to the African group] which evolved from a common ancestral stock—the Holarctic fossil Lithornithidae of the Paleocene-Eocene periods, which were possibly much like the still volant (capable of flight) tinamous of Central and South America. Each phyletic line probably reached its present center of distribution (African, Australasia, and the Neotropics) before becoming gigantic in size and flightless. The ratites are advanced birds, not a primitive avian group as argued by a number of earlier workers. *See* AVES; NEOGNATHAE; PHYLOGENY; STRUTHIONIFORMES.   Walter J. Bock

Bibliography. S. Davis, *Ratites and Tinamous*, Oxford University Press, 1999; A. Feduccia, *The Origin and Evolution of Birds*, 2d ed., Yale University Press, 1999; B. Monroe and C. G. Sibley, *A World Checklist of Birds*, Yale University Press, 1997.

Rauwolfia

A genus of mostly poisonous, tropical trees and shrubs of the dogbane family (Apocynaceae). Certain species are the source of valuable emetics and cathartics. The species *Rauwolfia serpentina* (see **illus.**) has received special attention as the source of

Branch of *Rauwolfia serpentina.*

tranquilizing drugs. For centuries in India the drug has been used in the treatment of hypertension. It came into use in Western countries because of its effect in reducing blood pressure. Although not a sedative, as that term is usually construed, it often has a quieting influence on the patient. Among the purified alkaloids obtained from *R. serpentina*, reserpine (Serpasil) is perhaps the one most used as a tranquilizing agent. *See* GENTIANALES; TRANQUILIZER.   Perry D. Strausbaugh; Earl L. Core

Raw water

Water obtained from natural sources such as streams, reservoirs, and wells. Natural water always contains impurities in the form of suspended or dissolved mineral or organic matter and as dissolved gases acquired from contact with earth and atmosphere. Industrial or municipal wastes may also contaminate raw water. *See* WATER POLLUTION.

If admitted to a steam-generating unit, such contaminants may corrode metals or form insulating deposits of sediments or scale on heat-transfer surfaces, with resultant overheating and possible failure of pressure parts.

Principal scale-forming impurities are compounds of calcium and magnesium, or silica. Principal corrosive agents are dissolved oxygen and carbon dioxide. In some localities, raw water has a mineral acidity. Oil and grease impair wetting and heat removal from the steam-generating surfaces and may also form corrosive scale or sludge. Certain organic materials or a high concentration of dissolved solids in the boiler water may cause foaming which contaminates the steam.

Raw water can be treated to remove objectionable impurities or to convert them to forms that can

be tolerated. For steam generation, suspended solids are removed by settling or filtration. Scale-forming hardness is diminished by chemical treatment to produce insoluble precipitates that are removable by filtration, or soluble compounds that do not form scale. Essentially complete purification is achieved by demineralizing treatment or evaporation. Demineralization consists of passing the water through beds of synthetic ion-exchange resin particles. Certain of these exchange hydrogen for metallic cations; others exchange hydroxyl for sulfate, chloride, or other anions in solution. The hydrogen and hydroxyl ions combine to form water. The resins may be used in separate or mixed-bed arrangements. They require periodic regeneration by acid and alkaline solutions, respectively; the mixed resins can be separated for such treatment by virtue of the fact that they differ in specific gravity. *See* WATER SOFTENING; WATER TREATMENT.

Evaporation requires expenditure of heat for complete vaporization of the water. The vapor is subsequently condensed and collected as purified distillate. Low-pressure steam is used as a heat source; multiple-effect heat exchange provides thermal economy.   F. G. Ely

Reactance

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

The impedance Z of an alternating current circuit is a complex number given by Eq. (1). The imaginary

$$Z = R + jX \qquad (1)$$

part X is the reactance. The units of reactance, like those of impedance, are ohms. Reactance may be positive or negative. For example, the impedance of an inductor L at frequency ω is given by Eq. (2), so

$$Z = jL\omega = jX \qquad (2)$$

X is positive, and X is referred to as the inductive reactance. The impedance of a capacitor C is given by Eq. (3), so X is negative, and X is referred to as the capacitive reactance.

$$Z = -\frac{j}{C\omega} = jX \qquad (3)$$

The reactance of a circuit may depend on both the resistors and the inductors or capacitors in the circuit. For example, the circuit in the **illustration** has admittance [Eq. (4)] and impedance [Eq. (5)], so that the value of the reactance [Eq. (6)],
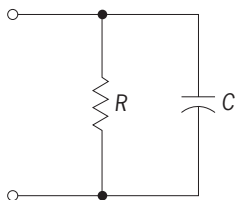
Circuit with a resistor and capacitor in parallel.

depends on both the capacitor C and the resistor R.

$$Y = \frac{1}{R} + jC\omega \qquad (4)$$

$$Z = \frac{R}{1 + jRC\omega} \qquad (5)$$

$$X = -\frac{R^2 C\omega}{1 + R^2 C^2 \omega^2} \qquad (6)$$

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

Reaction turbine

A power-generation prime mover utilizing the steady-flow principle of fluid acceleration, where nozzles are mounted on the moving element. The rotor is turned by the reaction of the issuing fluid jet and is utilized in varying degrees in steam, gas, and hydraulic turbines. All turbines contain nozzles; the distinction between the impulse and reaction principles rests in the fact that impulse turbines use only stationary nozzles, while reaction turbines must incorporate moving nozzles. A nozzle is defined as a fluid dynamic device containing a throat where the fluid pressure drops and potential energy is converted to the kinetic form with consequent acceleration of the fluid. For details of the two basic principles of impulse and reaction as applied to turbine design *see* IMPULSE TURBINE.   Theodore Baumeister

Reactive intermediates

Unstable compounds that are formed as necessary intermediate stages during a chemical reaction. Thus, if a reaction in which A is converted to B requires that A first be converted to C, then C is an intermediate in the reaction (A → C → B). The term reactive further implies a certain degree of instability of the intermediate; reactive intermediates are typically isolable only under special conditions, and most of the information regarding the structure and properties of reactive intermediates comes from indirect experimental evidence.

Organic Reactions

In organic reactions, the most common types of reactive intermediates arise from dissociative reactions, in which a carbon atom dissociates from some appropriate leaving group, resulting in a carbon reactive intermediate with a decreased valence (**Table 1**). Associative reactions, in which some reactive species forms a new bond to carbon, can also give rise to some of the same classes of carbon reactive intermediates, as well as other classes in which carbon has increased valence. The most common types of reactive species to initiate associative reactions are nucleophiles which bring in a pair of electrons for the new bond; free radicals which bring in a single electron for the new bond; and electrophiles which contribute no electrons toward the new bond (**Table 2**). *See* VALENCE.

TABLE 1. Most common types of organic reactive intermediates*

Dissociative reaction	Intermediate	Simplest example
$X-C- \longrightarrow X{:}^- + {}^+C-$	Carbocation	CH_3^+ (methyl cation)
$X-C- \longrightarrow X{\cdot} + {\cdot}C-$	Free radical	$CH_3{\cdot}$ (methyl radical)
$X-C- \longrightarrow X^+ + {}^-{:}C-$	Carbanion	$CH_3{:}^-$ (methyl anion)
$X-C- \longrightarrow XY + :C$	Carbene	$CH_3{:}$ (methylene)

*X^-, $X^{\cdot}$, X^+, XY represent appropriate leaving groups.

Carbocations. These are compounds in which carbon bears a positive charge. The simplest type of carbocation is a carbon atom that makes only three bonds and has no additional electrons. With a total of only six valence electrons (from the three bonds), such a carbon atom bears a positive charge. The simple trivalent carbocations have a planar structure. The presence of only six valence electrons represents an electron deficiency, and any structural features that offer electron donation stabilize the carbocation. Particularly effective are resonance effects, in which electron density is donated from a pair of electrons from an adjacent atom or from an adjacent pi bond or aromatic ring, as in notation (1).

$$(1)$$
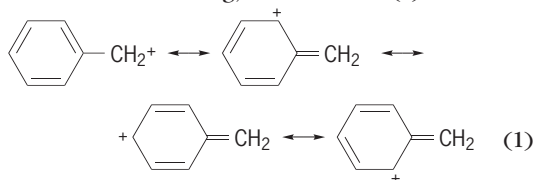
See RESONANCE (MOLECULAR STRUCTURE).

Reactivity of carbocations is dominated by their electron deficiency. Virtually all reactions of carbocations involve bonding to an electron pair, by covalent bonding to an anion or neutral molecule with an available pair of electrons (a nucleophile), or by bonding to a pair of electrons from a pi bond, or by shifting a pair of electrons from an adjacent bond. Carbocations are intermediates in a large variety of organic reactions (**Table 3**). *See* CHEMICAL BONDING.

Carbocations also exist in which carbon has the normal complement of eight valence electrons but still bears a positive charge. In these cases, sometimes called nonclassical carbocations, the key structural feature is a three-center, two-electron bond. The simplest example is the methonium ion, CH_5^+; and a typical reaction, electrophilic substitution of methane by deuterium ion, is shown in reaction (2).

TABLE 2. Organic reactive intermediates produced by associative reactions

Associative reaction	Intermediate
$C=C + X^+ \longrightarrow X-C-C^+$	Carbocation (X^+ = electrophile)
$C=C + X{\cdot} \longrightarrow X-C-C{\cdot}$	Free radical (X^- = free radical)
$C=C + X{:}^- \longrightarrow X-C-C{:}^-$	Carbanion (X^- = nucleophile)
$C=\ddot{O}{:} + X{:}^- \longrightarrow X-C-\ddot{O}{:}^-$	Tetrahedral intermediate (X^- = nucleophile)
$-C-X + Y^+ \longrightarrow -C{\cdots}{}^+$ with Y, X	Carbonium ion (Y^+ = electrophile)

TABLE 3. Some reactions involving carbocation intermediates

Reaction	Reaction type
	Nucleophilic substitution (X^- = leaving group, Y^- = nucleophile)
	Elimination (X^- = leaving group)
	Electrophilic aromatic substitution (X^+ = electrophile)
	Electrophilic addition (X^+ = electrophile, Y^- = nucleophile)
	Cationic polymerization (X^+ = electrophile)

$$(2)$$
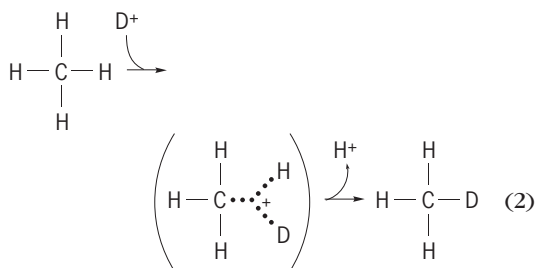
Carbanions. These are compounds in which carbon bears a negative charge. A carbanion will always have a positive counterion in associated with it; depending upon the particular cation and the stability of the carbanion, the association may be ionic, coavalent, or some intermediate combination of ionic and covalent bonding, as in notation (3), where M = metal.

$$(3)$$

Covalent   Ionic
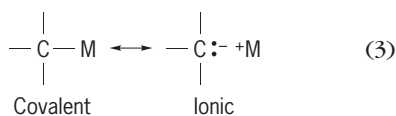
Compounds involving primarily covalent bonding of carbon to a metal are considered organometallic compounds, which frequently have properties similar to carbanions but substantially modified by the presence of the metal. *See* ORGANOMETALLIC COMPOUND.

Carbanions are trivalent, with eight valence electrons. Simple alkyl carbanions assume a pyramidal structure, analogous to ammonia, and are extremely strong bases. Carbanions are stabilized by structural features that allow electron withdrawal, in particular, inductive effects of halogen substituents or res-

onance effects of nitro (—NO_2), cyano (—CN), or carbonyl

substituents. Such resonance-stabilized carbanions have a planar structure and are relatively easily formed in acid-base reactions. An example is shown in reaction (4).

$$(4)$$

Resonance-stabilized enolate anion
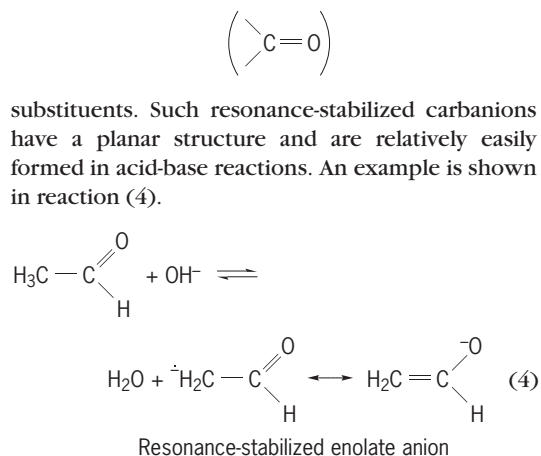
Carbanions are electron-rich; their reactions are mostly with electron-deficient species (**Table 4**). The use of carbanions (or organometallic analogs) in alkylation and condensation reactions generates new carbon-carbon bonds and provides some of the most useful reactions for organic synthesis. *See* ORGANIC SYNTHESIS.

Free radicals. These are neutral compounds having an odd number of electrons and therefore one unpaired electron. Carbon free radicals are trivalent, with seven valence electrons, and typically assume a planar structure. Free radicals are primarily electron-deficient species and are stabilized by structural features that donate electron density or delocalize the odd electron by resonance. Structural features that prevent the normal reactions of radicals, such as large substituents that surround the free-radical site, are very effective in increasing radical stability. *See* DELOCALIZATION.

TABLE 4. Some reactions involving carbanion intermediates

Reaction	Reaction type
$-\overset{\mid}{\underset{\mid}{C}}-H \xrightarrow{\text{base}} -\overset{\mid}{\underset{\mid}{C}}\!:^- \xrightarrow{X^+} -\overset{\mid}{\underset{\mid}{C}}-X$ $\left(\xrightarrow{RY} -\overset{\mid}{\underset{\mid}{C}}-R + Y:^- \right)$	Electrophilic substitution (alkylation) (X^+ = electrophile, Y^- = leaving group)
$O_2N-\!\!\!\bigcirc\!\!\!-X \xrightarrow{Y:^-} O_2N-\!\!\!\bigcirc\!\!\!\overset{X}{\underset{Y}{<}} \xrightarrow{X:^-} O_2N-\!\!\!\bigcirc\!\!\!-Y$ (with NO_2 substituents)	Nucleophilic aromatic substitution (Y^- = nucleophile, X^- = leaving group)
$-\overset{H}{\underset{\mid}{C}}-\overset{X}{\underset{\mid}{C}}- \xrightarrow{\text{base}} -\overset{..}{\underset{\mid}{C}}-\overset{X}{\underset{\mid}{C}}- \xrightarrow{X:^-} \,>\!C\!=\!C\!<$	Elimination (X^- = leaving group)
$>\!C\!=\!C\!< \xrightarrow{X:^-} -\overset{X}{\underset{\mid}{C}}-\overset{..}{\underset{\mid}{C}}- \xrightarrow{H^+} -\overset{X}{\underset{\mid}{C}}-\overset{H}{\underset{\mid}{C}}-$	Nucleophilic addition (X^- = nucleophile)
$>\!C\!=\!C\!< \xrightarrow{X:^-} X-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}\!:^- \xrightarrow{>C=C<} X-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}\!:^- \longrightarrow$ (etc.)	Anionic polymerization (X^- = nucleophile)
$-\overset{\mid}{\underset{\mid}{C}}-H \xrightarrow{\text{base}} -\overset{\mid}{\underset{\mid}{C}}\!:^- \xrightarrow{C=O} -\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}-O^- \xrightarrow{H^+} -\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}-OH$	Condensation

An important characteristic of many reactions involving free-radical intermediates is the tendency to follow a chain reaction mechanism. An example is shown.

Initiation:

$$Cl_2 \rightarrow 2Cl\cdot$$

Propagation:

$$Cl\cdot + CH_4 \rightarrow CH_3\cdot + HCl$$
$$CH_3\cdot + Cl_2 \rightarrow CH_3Cl + Cl\cdot$$

Termination:

$$2Cl\cdot \rightarrow Cl_2$$
$$2CH_3 \rightarrow CH_3CH_3$$
$$CH_3\cdot + Cl\cdot \rightarrow CH_3Cl$$
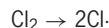
Overall substitution reaction:

$$CH_4 + Cl_2 \rightarrow CH_3Cl + HCl$$
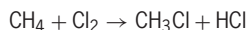
Addition by a free-radical chain reaction:

$$X\cdot + \;>\!C\!=\!C\!< \longrightarrow X-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}\!\cdot \xrightarrow{XY}$$

$$X-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}-Y + X\cdot$$
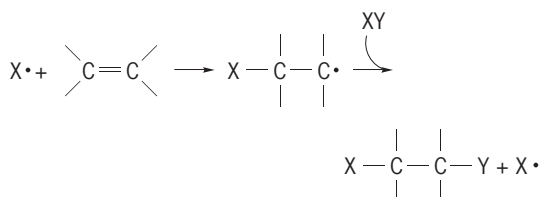
Radical chain polymerization:

$$X\cdot + \;>\!C\!=\!C\!< \longrightarrow X-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}\!\cdot \xrightarrow{>C=C<}$$

$$X-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}-\overset{\mid}{\underset{\mid}{C}}\!\cdot \longrightarrow \text{(etc.)}$$
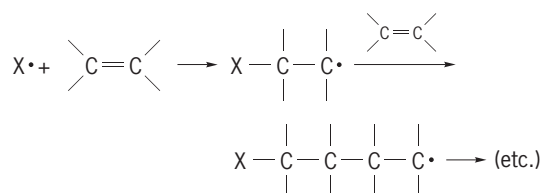
Chain reactions have three fundamental components: initiation, propagation, and termination steps. Initiation steps generate the reactive free radical from an appropriate precursor (or radical initiator). Propagation steps utilize the reactive free radical to generate product, and also regenerate another reactive free radical so that those steps may continue recycling. The overall reaction is the sum of the propagation steps, which may recycle thousands of times. Termination steps involve the combination of free radicals; each combination effectively terminates two chains.

Other common free-radical reactions include addition reactions, including polymerization. *See* CHAIN REACTION (CHEMISTRY); FREE RADICAL; POLYMERIZATION.

Radical ions. These compounds are charged and have an unpaired electron; they are either radical cations (positively charged) or radical anions (negatively charged). In many cases a radical ion is derived from a stable neutral molecule by addition of

one electron (radical anion) or removal of one electron (radical cation); thus, radical ions are frequently encountered as intermediates in electrochemical or other one-electron reduction or oxidation processes, as shown in the pinacol reduction reaction (5).

$$2(CH_3)_2CO + 2Na \longrightarrow$$

$$2(CH_3)_2C - O^- \longrightarrow (CH_3)_2C - C(CH_3)_2 \quad (5)$$

Acetone
radical anion

$$\downarrow 2H^+$$

$$(CH_3)_2C - C(CH_3)_2$$
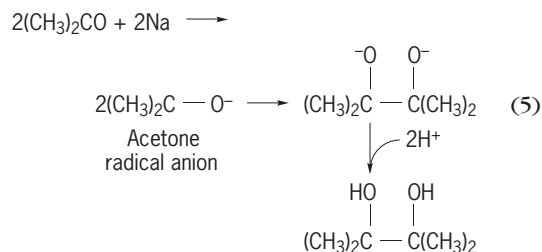
See ELECTROCHEMISTRY; OXIDATION-REDUCTION.

Carbenes. These are compounds which have a divalent carbon. The divalent carbon also has two nonbonded electrons, for a total of six valence electrons. The two nonbonded electrons may have either the same spin quantum number, which is a triplet state, or opposite spin quantum number, which is a singlet state. The triplet state of carbenes is typically lower in energy than the singlet state. Singlet and triplet carbenes normally undergo the same reactions, but because of the spin differences they must follow different mechanisms, so there are differences in reactivity, selectivity, and stereochemical details of the reactions. Typical singlet carbene reactions are concerted, involving a single step, and usually maintain stereochemistry. Triplet carbene reactions typically proceed through a triplet radical pair or a triplet biradical intermediate, which frequently randomizes stereochemistry. *See* ELECTRON SPIN; TRIPLET STATE.

Generation of carbenes is most commonly by photolysis or thermolysis diazo compounds or ketenes, or by alpha-elimination reactions. Carbenes are extremely reactive because of their electron deficiency. Typical reactions may be classified as reactions with single bonds (insertions) and those with double bonds (cycloadditions). Carbene cycloaddition reactions are among the best synthetic procedures for making cyclopropane derivatives. Internal rearrangements of carbenes are also common, depending on the particular structure of the carbene (**Table 5**). *See* PERICYCLIC REACTION.

Carbenoids are carbene complexes which have reactivity similar to carbenes. Often their reactivity is moderated in a way that is very useful synthetically, such as the Simmons-Smith reagent (CH_2I_2-Zn), which undergoes cycloaddition reactions well without any competing insertion reactions.

Nitrenes. These are compounds which contain a monovalent nitrogen atom. The monovalent nitrogen will have a nonbonded pair of electrons and two additional electrons which may also be spin-paired (a singlet nitrene) or unpaired (a triplet nitrene). Nitrenes, like carbenes, have only six valence electrons and are highly electron-deficient. There are strong similarities between nitrenes and carbenes in terms of both their structure and their chemical reactivity. Nitrene insertion reactions form new C—N bonds, and nitrene cycloadditions provide a good synthetic route to aziridines. Nitrenes are conveniently generated by thermolysis or photolysis of azides.

Tetrahedral intermediates. These are compounds derived from addition of a nucleophile to a carbonyl group. Depending upon the substituents on the carbonyl group, the initially formed tetrahedral intermediate may give either an addition or a substitution product. Aldehydes and ketones give addition products, whereas carboxyl derivatives (including carboxylic acids, esters, amides, and acid anhydrides) give substitution products (**Table 6**).

Excited states. These are unstable electronic states of a molecule in which the electron distribution is different from the most stable distribution for the molecule (the ground state). Occasionally, vibrational excited states are encountered, in which a molecule contains an abnormally large amount of vibrational energy. Vibrational excited states lose their excess vibrational energy very rapidly, particularly in condensed phases, and generally the term excited states refers to electronically excited molecules. Typically, electronically excited states are generated by absorption of a photon of light; hence the chemistry of excited states is called photochemistry. Absorption promotes an electron from a normally occupied molecular orbital into a normally unoccupied molecular orbital (see **illus.**). Because of this electron redistribution, excited states are always better oxidizing agents as well as better reducing agents that the molecule in its ground state, as measured by ionization potential (IP; the energy to remove one electron) and electron affinity (EA; the energy gained on adding an electron). Thus many photochemical reactions involve one-electron transfer as their initial step.

TABLE 5. Reactions of carbenes

Reaction	Classification
:CH_2 + —C—H ⟶ —C—CH_2—H	Insertion
:CH_2 + $\diagdown$C=C$\diagup$ ⟶ —C—⟨CH_2⟩—C—	Cyclo-addition
R—C(=O)—CH: ⟶ R—CH=C=O	Wolff rearrangement
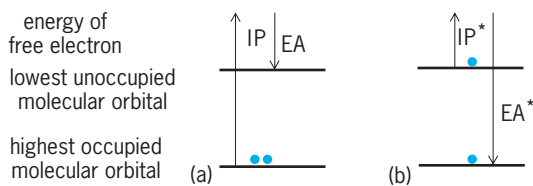
Ionization potentials (IP) and electron affinities (EA) for a molecule in (*a*) its ground state (A) and (*b*) its lowest electronically excited state (A*).

TABLE 6. Addition and substitution reactions via tetrahedral intermediates*

Reaction	Product

Nucleophilic addition
(X⁻ = nucleophile) — shown in table for the Aldehyde or ketone reaction

Nucleophilic substitution
(X⁻ = nucleophile, Y⁻ = leaving group) — shown in table for the Carboxyl derivative reaction

*R, R′ = C or H; Y = O, N, S, halogen.

Like carbenes and nitrenes, excited states have two unpaired electrons whose spin quantum numbers may be opposite (a single excited state) or parallel (a triplet excited state). A triplet excited state will always be lower in energy than the corresponding singlet excited state.

Because an excited state differs from the more stable ground state only in its electron distribution, the predominant unimolecular reactions of excited states involve return to the ground state, by emission of light (fluorescence from a singlet state, phosphorescence from a triplet state) or by conversion of the excess energy to heat. These unimolecular processes are very rapid, and excited state lifetimes are very short (for singlets, typically 10^{-6} to 10^{-12} s; for triplets, 10^{-6} to 10 s). Bimolecular reactions between excited states and other molecules also contribute to the deactivation of excited states. Such reactions include energy transfer, electron transfer, and a variety of specific reaction types depending on the compounds involved. However, to compete effectively with the rapid unimolecular deactivation of excited stated, bimolecular reactions must be extremely rapid, typically with rate constants close to diffusion control. *See* ELECTRON-TRANSFER REACTION.

Photochemistry results from excited states when a new compound is formed, either by unimolecular reaction of the excited state (rearrangement or fragmentation, for example) or by bimolecular reaction with another molecule (photoadditions, photooxidations, and photoreductions, for example). Some of the special features of photochemical reactions which differentiate them from normal (thermal) chemical rections include the following: (1) Compounds in a mixture may be selectively excited, depending upon the absorption spectrum of the compounds and the wavelength of light used. (2) Absorption of a photon provides a molecules with large amounts of energy (ultraviolet photons of 286 nanometers provide a total of 100 kcal or 420 kilojoules per mole) relative to thermal energy (around 1 kcal or 4.2 kJ per mole).

Considering these two points together, low-temperature photochemistry offers the opportunity for both the selective formation and the stabilization of compounds which would be highly unstable at ambient temperatures. Many highly strained reactive intermediates have been isolated and characterized under these conditions, including benzyne [reaction (6)]. Other intermediates which have been

$$\xrightarrow[\text{8 K}]{\text{254 nm}} \quad + 2CO_2 \qquad (6)$$
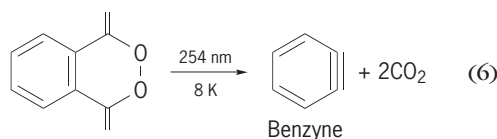
Benzyne

immobilized at low temperatures include free radicals, carbenes, nitrenes, and ylides. *See* PHOTOCHEMISTRY; YLIDE.

Other types. There are many kinds of organic reactive intermediates which do not fit into the previous classifications. Benzyne is a reactive intermediate which can be derived from an aryl halide by elimination with strong base: benzyne rapidly undergoes nucleophilic addition to give a net result of a substitution, reaction (7a), or may undergo cycloaddition, reaction (7b).

$$\qquad (7a)$$

Benzyne

$$\qquad (7b)$$
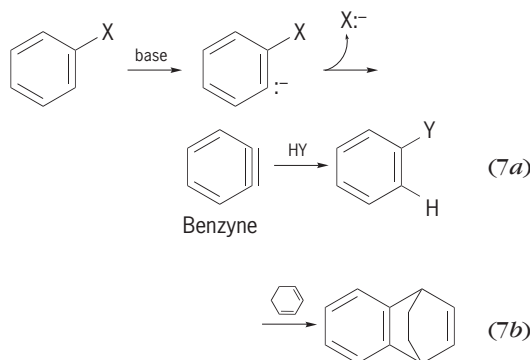
See SUBSTITUTION REACTION.

Other reactive intermediates are simply compounds which are unstable for a variety of possible reasons, such as structural strain or an unusual oxidation state.   Carl C. Wamser

Inorganic Reactions

Inorganic reactions may involve reactive intermediates as atoms, ionic species, covalently bound radicals or radical ions, and coordination compounds. In the gas phase, intermediates occur mainly as free radicals. The complexity of inorganic intermediates arises mainly from the possible participation of more than 100 elements and their varied oxidation states. However, with the tremendous upsurge of inorganic and organometallic research since the mid-1960s, some of the complex mechanisms leading to these intermediates as well as their reactivity are more clearly understood. It is almost impossible to provide a uniform classification of the common inorganic intermediates in the same manner that is possible for organic reactive intermediates.

On the other hand, some organometallic intermediates resemble organic intermediates. Divalent species of group 14 metals such as germylenes with the structure

$$Ge \begin{smallmatrix} X \\ \\ X \end{smallmatrix} \quad (X = F^-, Cl^-, \text{ or phenyl})$$
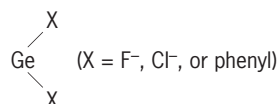
exhibit properties analogous to those of carbenes. Other examples are compexes containing transition metal (M) to carbon double bonds, such as cationic alkylidene complexes, $M^+{=}CH_2$. In some cases, these reactive intermediates could be regarded as metallocarbocations.

Inorganic reactive intermediates, such as metal atoms, radicals, radical ions, or metal ions in unusual valency states, are very reactive species. They can be generated either by ordinary chemical processes or by special methods, such as continuous or pulse radiolysis, photolysis, or electron beam vaporization. In particular in this section, intermediates proposed or "detected" in oxidation-reduction and substitution reaction of the metal complexes are discussed in detail. *See* OPTICAL PULSES; PHOTOLYSIS; ULTRAFAST MOLECULAR PROCESSES. For other types of processes or methods which could give rise to reactive inorganic intermediates; *see* ELECTROCHEMICAL PROCESS; HOMOGENEOUS CATALYSIS; INORGANIC PHOTOCHEMISTRY; LASER PHOTOCHEMISTRY; OSCILLATORY REACTION.

Inorganic radicals and radical ions. Three major short-lived intermediates, hydrated electron (e_{aq}^-), hydrogen atom (H·), and hydroxyl radical (OH·), are produced by irradiation of aqueous solutions, as shown by reaction (8). These primary radicals react

$$H_2O \xrightarrow{\text{irradiation}} e_{aq}^- + H\cdot + OH\cdot \qquad (8)$$
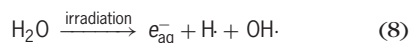
with solutes in water to form other short-lived and also very reactive intermediates, such as oxyanion, halide, pseudohalide, and cation-radical ions.

Hydrated electron. This radical is a very strong reducing agent ($E^\circ = -2.77$ V, where E° represents standard electrode potential), and has a broad absorption band with maximum at 715 nm, corresponding to a blue solution. In the presence of oxygen, e_{aq}^-

is rapidly converted into other reactive species, superoxide ion O_2^-, and hydroperoxyl radical HO_2 as shown by reaction (9).

$$e_{aq}^- + O_2 \rightarrow O_2^- \longrightarrow H + HO_2 \qquad (9)$$
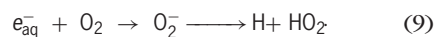
Hydrated electrons also can be completely converted into hydroxyl radicals when a totally oxidizing system is obtained. This can be achieved when N_2O is present as a solute, as shown by reaction (10).

$$e_{aq}^- + N_2O \rightarrow N_2 + O^- (+H_2O) \rightarrow OH\cdot + OH^- \qquad (10)$$
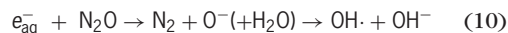
If other solutes are present, such as peroxodisulfate ions ($S_2O_8^{2-}$), reaction (11) occurs. In this exam-

$$e_{aq}^- + S_2O_8^{2-} \longrightarrow SO_4^- \cdot + SO_4^{2-} \qquad (11)$$
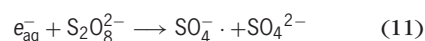
ple, a new reactive sulfate radical anion is produced.

A radical-cation intermediate can be generated by the very fast reaction of hydrated electrons with cobalt(III)-ammine complexes containing a mono- or dinitrobenzoato ligand, as shown by reaction (12).

$$\longrightarrow \text{products} \qquad (12)$$
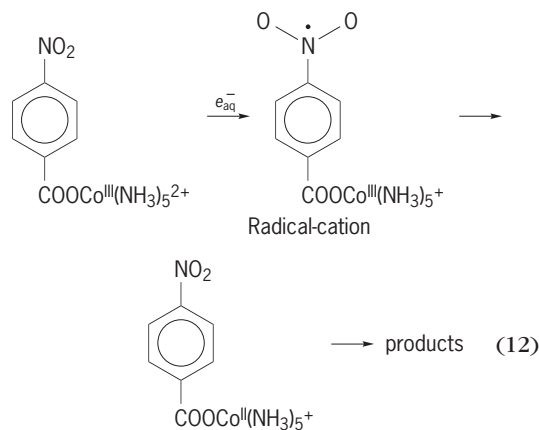
Hydrogen atom. This is the simplest chemical species, carrying a single electron, which decays very rapidly in aqueous solution. In alkaline solution the hydrogen atom is converted into the hydrated electron [reaction (13)], and in acidic solution the hydrated electron is reconverted to the hydrogen atom [reaction (14)].

$$H\cdot + OH^- \longrightarrow e_{aq}^- + H_2O \qquad (13)$$

$$e_{aq}^- + H^+ \longrightarrow H\cdot \qquad (14)$$
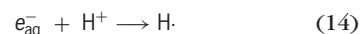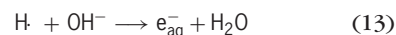
Although the hydrogen atom is a strong reducing agent, there are reactions where it acts as an oxidant: as an example, the formation of the reactive hydrido complex, $Fe^{3+}H^-$ (**Table 7**).

Hydroxyl radical. This radical is a strong oxidant ($E^\circ = 1.9$ V) which reacts rapidly with organic and inorganic compounds, often producing new radicals. A hydroxyl radical is frequently produced as a reactive intermediate in reactions in solutions and gas phase, for example, reactions (15), where hydroxyl

$$Co(CN)_5^{3-} + H_2O_2 \rightleftharpoons Co(CN)_5OH^{3-} + OH\cdot \qquad (15a)$$

$$Co(CN)_5^{3-} + OH\cdot \longrightarrow Co(CN)_5OH^{3-} \qquad (15b)$$

TABLE 7. Some reactions of radicals and radical ions with inorganic and organic solutes

Radical or radical ion		Solute		Products
$OH\cdot$	$+$	Fe^{2+}	$\rightarrow$	$Fe^{3+} + OH^-$
O^-	$+$	$Fe^{2+}(+H_2O)$	$\rightarrow$	$Fe^{3+} + 2OH^-$
$HO_2\cdot$	$+$	Fe^{2+}	$\rightarrow$	$Fe^{3+}\ OH_2^-\ (+H^+) \rightarrow$
				$Fe^{3+} + H_2O_2$
O_2^-	$+$	Fe^{3+}	$\rightarrow$	$Fe^{2+} + O_2$
$H\cdot$	$+$	Fe^{2+}	$\rightleftharpoons$	$Fe^{3+}\ H^-\ (+H^+) \rightarrow$
				$Fe^{3+} + H_2$
$H\cdot$	$+$	Fe^{3+}	$\rightarrow$	$Fe^{2+} + H^+$
$Cl_2\cdot^-$	$+$	Fe^{2+}	$\rightarrow$	$FeCl^{2+} + Cl^-$
$Cl_2\cdot^-$	$+$	Fe^{2+}	$\rightarrow$	$Fe^{3+} + 2Cl^-$
e_{aq}^-	$+$	C_2H_5OH	$\rightarrow$	$C_2H_5O^- + H\cdot$
$H\cdot$	$+$	C_2H_5OH	$\rightarrow$	$C_2H_4OH + H_2$
$SO_4\cdot^-$	$+$	C_2H_5OH	$\rightarrow$	$C_2H_4OH + HSO_4^-$

radicals, OH, are short-lived reactive intermediates generated in the oxidation-reduction step (15*a*) and consumed in the addition step (15*b*).

In the pulse radiolysis of aqueous solution, hydroxyl radical is completely converted into a reactive carbon dioxide radical anion, CO_2^-, in the presence of a solute, such as methanol or formate. In this case, a totally new reducing system is achieved, as is shown in reaction (16). An additional example of

$$OH\cdot + HCOO^- \longrightarrow CO_2^- + H_2O \qquad (16)$$

hydroxyl radical reactivity is shown in reaction (17),

$$OH\cdot + Cl^- \rightleftharpoons ClOH^- + (H^+) \rightleftharpoons$$
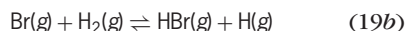$$H_2O + Cl\cdot + (Cl^-) \rightleftharpoons Cl_2^- \qquad (17)$$
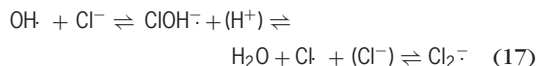
where the chloride radical, Cl_2^-, is produced. *See* HYDROXYL.

Gas-phase systems. There are many systems, such as in the gas phase, where the radicals are also common intermediates. An example is in the decomposition of ozone, O_3, to oxygen, which involves oxygen atoms as intermediates [reactions (18)].

$$O_3(g) \rightleftharpoons O_2(g) + O(g) \qquad (18a)$$
$$O_3(g) + O(g) \rightarrow 2O_2(g) \qquad (18b)$$
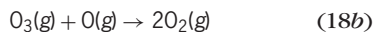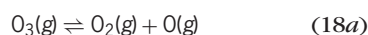
See OZONE.

Radicals are common reactive intermediates in gas-phase synthesis. An example is in the formation of hydrogen bromide, HBr, by the reaction of hydrogen and bromine gases at high temperature [reactions (19)]. This reaction proceeds by a chain mechanism with H and Br radicals as intermediates.
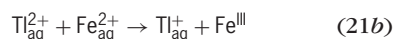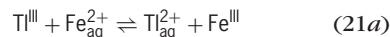
$$Br_2(g) \rightarrow 2Br(g) \qquad (19a)$$
$$Br(g) + H_2(g) \rightleftharpoons HBr(g) + H(g) \qquad (19b)$$
$$H(g) + Br_2(g) \rightarrow HBr(g) + Br(g) \qquad (19c)$$
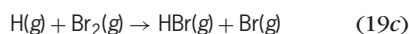
anism with H and Br radicals as intermediates.

Metal intermediates: oxidation-reduction. Reactive intermediates such as metal atoms and metal ions in unusual valency states and with metal complexes as intermediates are mainly classified according to the types of reactions in which they may arise. A major class involves oxidation-reduction.

Metal ions in unusual valency states. Some metals can exhibit higher reduced or higher oxidized or intermediate states than normally shown. In such states, metal ions are usually very reactive, and they are generally produced by radiation methods. A simple example is provided by reaction (20), where transition metal
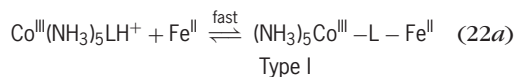
$$e_{aq}^- + M^{n+} \rightarrow M^{(n-1)+} \qquad (20)$$
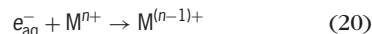
(M) in reduced state can be generated by the hydrated electron, e_{aq}^-. A few examples illustrate the wide variety of reactions of metals in an unusual oxidation state, such as Cu(III), Co(I), Eu(II), Ni(I), and Tl(II). (**Table 8**). It is also believed that the metal ions are reactive intermediates in some oxidation-reduction reactions in solutions. For example, it is assumed that the thallium ion Tl^{2+} is an intermediate in the oxidation of the ferrous ion Fe^{2+} by Tl^{3+} [reactions (21)].

$$Tl^{III} + Fe_{aq}^{2+} \rightleftharpoons Tl_{aq}^{2+} + Fe^{III} \qquad (21a)$$
$$Tl_{aq}^{2+} + Fe_{aq}^{2+} \rightarrow Tl_{aq}^+ + Fe^{III} \qquad (21b)$$

Metal atoms. Metals in the atomic state, also sometimes known as the high-surface-area state, are the most reactive form and hence are technologically very important in the fields of synthetic and catalytic chemistry. The reactive form of the metal can be obtained, for example, by electron beam or resistive heating vaporization. Atomic metals participate in a variety of chemical reactions (**Table 9**).

Binuclear complexes. The most common intermediates arising in the inner-sphere oxidation-reduction reactions between two metal complexes of transition elements are binuclear complexes. Depending on the state of valencies of the metals in the binuclear centers, these reactive intermediates may be subdivided into three types. The first type is generated when the oxidant and the reductant become mutually attached, usually through a ligand (L) of the oxidant. In this type, the metals in the binuclear site exhibit the initial oxidation states. This is apparent in reactions (22), where type I represents a binuclear

$$Co^{III}(NH_3)_5LH^+ + Fe^{II} \underset{}{\overset{fast}{\rightleftharpoons}} (NH_3)_5Co^{III}-L-Fe^{II} \quad (22a)$$
$$\text{Type I}$$

$$\text{Type I} \rightarrow Co^{II} + Fe^{III}L \qquad (22b)$$

intermediate, termed the precursor complex [reaction (22*a*)].

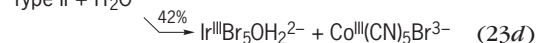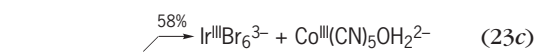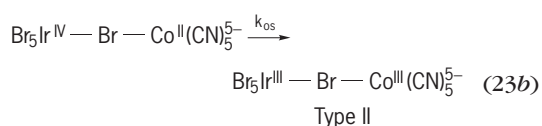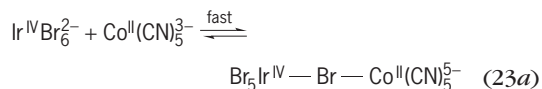**TABLE 8. Some reactions involving metals in unusual valency states**

Reaction	Reaction type
$Cu(OH)_2^+ + Cu(OH)_2^+$	Adduct formation
$Tl^{2+} + Tl^{2+}$	Dismutation
$Co^+ + Cu^{2+}$	Electron-transfer
$Eu^{2+} + Co(NH_3)_5Cl^{2+}$	Inner-sphere electron transfer
$Ni(L)^+ + H_3O^+$	Proton transfer

TABLE 9. Some chemical reactions of metal atoms

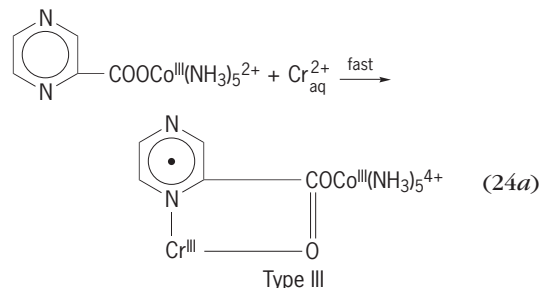| Atoms	Reaction substrate	Products	Reaction type
$2Cu$ +	$2BCl_3$	$2CuCl + B_2Cl_4$	Abstraction
Pd +	$RC(O)Cl$	$RC(O) - Pd - Cl$	Oxidative addition
K +	O_2	$K^+ O_2^-$	Electron transfer
Fe +	(cyclopentadiene ring)	(ferrocene) $Fe + H_2$	Simple orbital mixing

Reactions (23) give an example of the second type of binuclear complex. Usually termed the successor complex, type II could originate in reaction step (23b), where k_{is} represents rate constant for the inner-sphere electron transfer process. Both metals in this type exhibit their final oxidation states.

$$Ir^{IV}Br_6^{2-} + Co^{II}(CN)_5^{3-} \xrightleftharpoons{fast}$$
$$Br_5Ir^{IV} - Br - Co^{II}(CN)_5^{5-} \quad (23a)$$

$$Br_5Ir^{IV} - Br - Co^{II}(CN)_5^{5-} \xrightarrow{k_{is}}$$
$$Br_5Ir^{III} - Br - Co^{III}(CN)_5^{5-} \quad (23b)$$
$$\text{Type II}$$

$$\text{Type II} + H_2O \begin{cases} \xrightarrow{58\%} Ir^{III}Br_6^{3-} + Co^{III}(CN)_5OH_2^{2-} \quad (23c) \\ \xrightarrow{42\%} Ir^{III}Br_5OH_2^{2-} + Co^{III}(CN)_5Br^{3-} \quad (23d) \end{cases}$$

The third type of binuclear intermediate contains metals in the mixed valence: the oxidant metal ion is in the initial oxidation state, and the reductant metal ion is already in the final oxidation state. The most widely known example is the radical-cation intermediate, type III, shown in reactions (24). (The dot inside the ring represents electron-rich pyrazine ligand.) *See* LIGAND.

Other types. Inevitably, many intermediates generated in the oxidation-reduction reactions of metal compounds cannot be neatly placed in the above categories. It has been reported that in some systems the "dead-end" intermediate can be formed. The best-known example is illustrated in reactions (25), where EDTA represents ethylenediaminetetraacetic acid and k_{os} represents the outer-sphere rate constant. In the first step [reaction (25a)], the dead-end intermediate appears, but the only productive oxidation-reduction process is the outer-sphere step [reaction (25b)].


$$(24a)$$
$$\text{Type III}$$

$$\text{Type III} \xrightarrow{decay} \quad C = O + Co^{2+} \quad (24b)$$

$$Co^{II}EDTA^{2-} + Fe^{III}(CN)_6^{3-} \xrightleftharpoons{fast}$$
$$EDTACo^{III} - NC - Fe^{II}(CN_5)^{5-} \quad (25a)$$
$$\text{"Dead-end"}$$

$$Co^{II}EDTA^{2-} + Fe^{III}(CN)_6^{3-} \xrightarrow{k_{os}}$$
$$Co^{II}EDTA^- + Fe^{II}(CN)_6^{3-} \quad (25b)$$
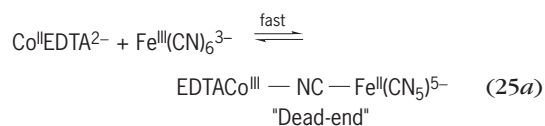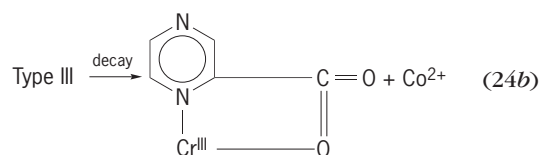
Metal intermediates: substitution and homogeneous catalysis. Reactive intermediates may exist for a very short time in dissociative and associative substitution reactions of metal complexes, or during homogeneous catalysis by transition metal complexes. The most common type is five-coordinate intermediates.

Five-coordinate: dissociative reactions and catalysis. This type of coordinatively unsaturated reactive intermediate is generated by dissociative substitution reactions in octahedral Werner-type complexes (**Table 10**) or in six-coordinate organometallic complexes.

TABLE 10. Some dissociative substitution reactions of Werner-type complexes involving five-coordinate intermediates

Reaction type	Metal complex	Entering ligand	Intermediate
Anation	$trans\text{-}Coen_2SO_3OH_2^+$	N_3^-	$Coen_2SO_3^+$
Base hydrolysis	$Co(NH_3)_5Cl^+$	OH^-	$Co(NH_3)_4NH_2^{2+}$
Induced aquation	$Co(NH_3)_5Br^+$	$(Hg^{2+})H_2O$	$Co(NH_3)_5^{3+}$
Replacement	$Fe(CN)_5SO_3^{5-}$	CN^-	$Fe(CN)_5^{3-}$
Solvolysis	$Cr(H_2O)_5I^{2+}$	$H_2O - CH_3OH$	$Cr(H_2O)_5^{3+}$

The intermediate of reduced coordination number of Werner-type complexes lives long enough to show some discrimination between potential ligands in its vicinity. For example, in the ligand replacement [reactions (26)], the generation of a five-coordinate

$$MA_5L \underset{k_2}{\overset{k_1}{\rightleftharpoons}} MA_5 + L \qquad (26a)$$

$$MA_5 + Y \xrightarrow{k_3} MA_5Y \qquad (26b)$$
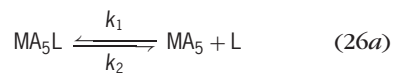
complex, MA_5, is shown in step (26a). The rate constants k_1, k_2, and k_3 are given for the individual reactions; k_1 represents the unimolecular dissociative rate constant, k_2 represents the rate constant for the addition of ligand (L) to the five-coordinate intermediate (MA_5), and k_3 represents the formation of a product with a different ligand (Y).

The competition ratio, k_2/k_3, is most often used to test the ability of the intermediate to discriminate between two ligands. It is helpful to recognize that the shape of a five-coordinate intermediate could be trigonal-bipyramidal or square-pyramidal, and it may be either rigid or fluxional. It is most probable that in the base hydrolysis of Co(III)-ammine and -amine complexes, the intermediate has trigonal bipyramidal geometry. On the other hand, in the base hydrolysis of Rh(III), its shape is probably square pyramid. There is apparently very little energy difference between the two geometries. A retention of stereoconfiguration is expected if reaction proceeds through square pryamid. A reaction proceeding through a trigonal bipyramid, on the other hand, would lead to a mixture of stereoisomers.

The nature of intermediates in organometallic reactions and catalysis is more complex. It is believed that dissociative reactions of 18-electron organotransition metal complexes proceed via 16-electron intermediates, and that 16-electron complexes proceed via 14-electron intermediates. Most of the common intermediates are three- or five-coordinated. The ligand-deficient intermediates are highly reactive, and the competition ratio shows that they are nondiscriminatory toward ligands of widely varying nucleophilicity. An example of a reaction which proceeds via five-coordinate 16-electron intermediate involves an acyl intermediate [reactions (27)].

$$RMn(CO)_5 \rightleftharpoons [RC(O)Mn(CO)_4] \qquad (27a)$$

$$[RC(O)Mn(CO)_4] + L \rightarrow RC(O)Mn(CO)_4L \qquad (27b)$$

This intermediate is generated by CO insertion [reaction (27a)] in alkyl transition metal carbonyl complexes.

Generally, the cyclic nature of catalysis requires that homogeneous catalytic reactions proceed by a series of ligand dissociation and ligand association steps in which intermediates with 16- and 18-electron configuration occur alternately.

Five-coordinate: associative reactions. It is generally accepted that substitution reactions of square-planar Pt(II), Pd(II), Au(II), Rh(I), and Ni(II) complexes occur by an associative mechanism with the possibility of the formation of a five-coordinate in-

termediate with increased coordination number. The trigonal-bipyramidal geometry of the intermediate has been assumed to be more favorable than square-pyramidal. The reaction of square-planar $Pt(NH_3)_3Cl^+$ complex with Cl^- ion to give trans-$Pt(NH_3)_2Cl_2$ is a typical example, which probably involves a five-coordinate trigonal bipyramidal intermediate, $Pt(NH_3)_3Cl_2$, as shown in reaction (28). The

Trigonal-bipyramidal
intermediate

(28)
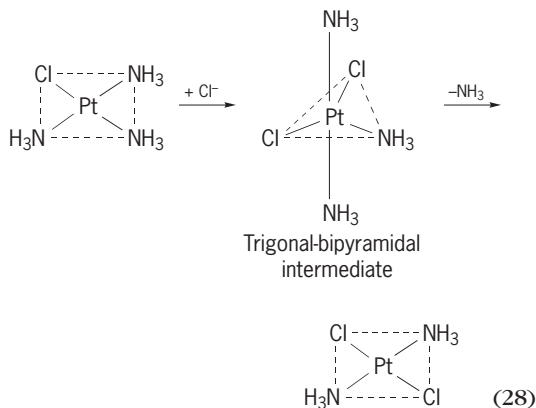
stability, structure, and reactivity of this type of intermediate is greatly influenced by solvents, ligands, and the nature of the metals involved in the reactions. In some systems, an intermediate exists for a sufficient time for it to undergo pseudorotation. *See* CHEMICAL DYNAMICS; COORDINATION CHEMISTRY; COORDINATION COMPLEXES; ELECTROPHILIC AND NUCLEOPHILIC REAGENTS; MOLECULAR ORBITAL THEORY; TRANSITION ELEMENTS.   Zdravko Bradic

Bibliography. J. D. Atwood, *Inorganic and Organometallic Reaction Mechanisms*, 1985; F. A. Cotton and G. Wilkinson, *Advanced Inorganic Chemistry*, 5th ed., 1988; A. B. Lever (ed.), *Excited States and Reactive Intermediates: Photochemistry, Photophysics, and Electrochemistry*, 1986; T. H. Lowry and K. S. Richardson, *Mechanism and Theory in Organic Chemistry*, 3d ed., 1987; J. March, *Advanced Organic Chemistry*, 5th ed., 2000; C. J. Moody and G. H. Whitham, *Reactive Intermediates*, 1992.

Reactive power

The concept of reactive power arises in electrical circuits operated from alternating-current sources. Electric fields (associated with capacitive effects) and magnetic fields (associated with inductive effects) are alternately charged and discharged according to the transfer of energy that takes place in each cycle. Reactive power is related to the peak of power absorbed by either capacitance or inductance in the circuit. The term "real or active power" is used to describe the average value of the power waveform. All power system equipment generates or absorbs reactive power, including synchronous generators, synchronous condensers, static compensators, capacitive and inductive compensators, overhead lines and underground cables, transformers, and consumer loads. Real power controls the
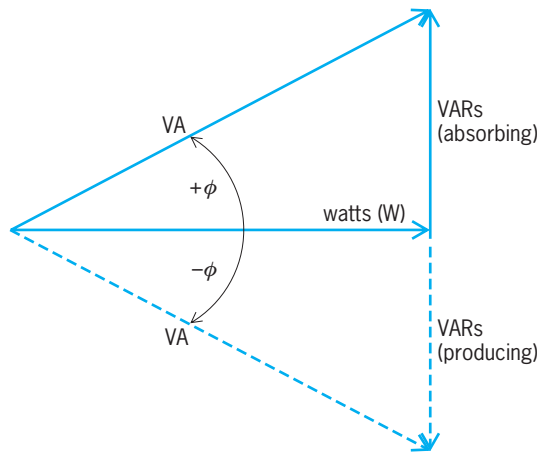
Fig. 1. The power triangle.

frequency of the system, while reactive power controls the voltage. Both are required for successful operation of the power system, and a balance must be respected between the real and reactive power at every point in the network. To violate the real power balance is to invite a condition known as transient instability, while violating the reactive power balance may lead to voltage instability or eventually collapse.

Power triangle. The real power consumed by a single-phase device will be denoted as P, the reactive power as Q, and the complex power as S, where $S = P + jQ$, and where j is the imaginary number $\sqrt{-1}$. The reactive power is measured by a unit called a

VAR, for volt-ampere reactive, while the real power is measured in watts (W). As shown in **Fig. 1**, the vector (complex) sum of these two components defines the complex power, S, whose magnitude $|S|$, termed the apparent power, is measured in units of volt-amperes (VA). The graphical relation between S, P, and Q is known as the power triangle. *See* POWER; VOLT-AMPERE.

Denoting the voltage magnitude across the device as V, the current magnitude as I, and the phase angle between the voltage and the current as ϕ, we can express the real power as $P = VI\cos(\phi)$, the reactive power as $Q = VI\sin(\phi)$, and the apparent power as VI. As shown in Fig. 1, the term $\cos(\phi) = P/S$, defined as the power factor, is a dimensionless quantity measured in watts/VA denoting the fraction of the apparent power which can be usefully converted into other forms of energy. As seen in Fig. 1, a device can absorb or produce VARs. By convention, a capacitor produces reactive power, while an inductor absorbs reactive power. *See* CAPACITOR; INDUCTOR.

Power-factor correction. The idea of power-factor correction is to compensate for reactive power consumption by inductive loads by generating reactive power as close as possible to the load which requires it, rather than supplying it from a remote power station. Alternatively, if a load generates reactive power, power-factor correction would be provided by consuming VARs close to the load.

Figure 2a shows a single-line diagram of a load with admittance $Y = G + jB$ supplied by a constant voltage source, V. The supply current is, therefore, $I = YV = VG + jVB = I_R + j_X$. The voltage and current, V and I, are represented by the phasor diagram of Fig. 2b, in which V is the reference phasor. The load current has a resistive component, I_R, in phase with V, and a reactive component, I_X, which is in quadrature with V. The current supplied by the power system is larger than necessary to supply the real power alone by the factor $I/I_R = 1/\cos(\phi)$. *See* ADMITTANCE.

Figure 3 shows the load after 100% compensation by placing a device at the load bus with a purely reactive admittance $-jB$. Then, the compensator current is $I_C = -V(jB)$, while the current supplied by the power system is $I_+ = V(G + jB) - V(jB) = VG = I_R$. This then means that the compensated supply current I_+ is in phase with V, making the overall power-factor unity. The supply current I_+ now has the smallest value capable of supplying full power at the voltage V, and all the reactive power required by the load is supplied locally by the compensator. Relieved of the reactive requirements of the load, the supply now has excess capacity which is available for supplying other loads.

The terms "leading" and "lagging" are often written in conjugate with the power factor. They are defined by the current through the load. If the current through a load leads the voltage across it, the load is said to have a leading power factor. If the current through a load lags the voltage across it, the load is said to have a lagging power factor. In other words, capacitive networks have a leading
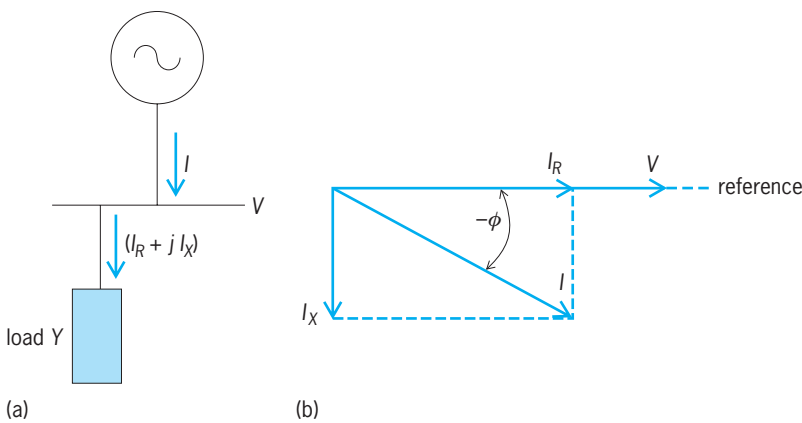

Fig. 2. Diagrams of uncompensated load. (*a*) Single-line diagram. (*b*) Phasor diagram.
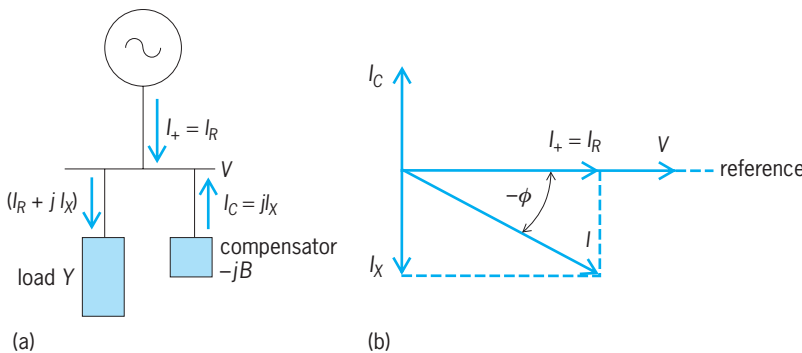

Fig. 3. Diagrams of compensated load. (*a*) Single-line diagram. (*b*) Phasor diagram.

power factor and inductive networks have a lagging power factor. *See* ALTERNATING-CURRENT CIRCUIT THEORY; ELECTRIC POWER TRANSMISSION.

D. McGillis; K. El-Arroudi; Francisco D. Galiana

Bibliography. A. E. Emanuel, On the definition of power factor and apparent power in unbalanced polyphase circuits with sinusoidal voltage and currents, *IEEE Trans. Power Delivery*, 8(3): 841–852, July 1993; A. E. Knowlton, Reactive powers in need of clarification, *AIEE Trans.*, 52:744–805, September 1933; R. M. Mathur, Static *Compensator for Reactive Power Control*, Cantext Publication, Canadian Electricity Association, Winnipeg, 1986; T. J. E. Miller, *Reactive Power Control in Electric Systems*, Wiley, 1982; Reactive Power Reserve Work Group Technical Studies Subcommittee, *Voltage Stability Criteria, Undervoltage Load Shedding Strategy and Reactive Power Reserve Monitoring Methodology*, Western Electricity Coordinating Council (WECC), Final Report, May 1998; C. W. Taylor, *Power System Voltage Stability*, McGraw-Hill, New York, 1994; U.S.-Canada Power System Outage Task Force, *Final Report on the August 14, 2003 Blackout in the United States and Canada: Cases and Recommendations*, April 2004; Western Electricity Coordinating Council (WECC), *WECC Policy Regarding Extreme Contingencies and Unplanned Events*, no. 12, May 2002; J. Willems, J. Ghijselen, and A. Emanuel, The apparent power concept and the IEEE standard 1459–2000, *IEEE Trans. Power Delivery*, 20(2):870–884, April 2005.

Reactor (electricity)

A device for introducing an inductive reactance into a circuit. Inductive reactance x is a function of the product of frequency f and inductance L; thus, $x = 2\pi f L$. For this reason, a reactor is also called an inductor. Since a voltage drop across a reactor increases with frequency of applied currents, a reactor is sometimes called a choke. All three terms describe a coil of insulated wire. *See* CHOKE (ELECTRICITY); INDUCTOR.

According to their construction, reactors can be divided into those that employ iron cores and those where no magnetic material is used within the windings. The first type consists of a coil encircling a circuit of iron which usually contains an air gap or a series of air gaps. The air gaps are used to attenuate the effects of saturation of the iron core. The second type, called an air-core reactor, is a simple circular coil, wound around a cylinder constructed of nonmagnetic material for greater mechanical strength. This strength is necessary for the coil to withstand the electromagnetic forces acting on each conductor. These forces become very large with heavy current flow, and their direction tends to compress the coil into less space: radial forces tend to elongate internal conductors in the coil and to compress the external ones while the axial forces press the end sections toward the center of the coil.

Both iron-core and air-core reactors may be of the air-cooled dry type or immersed in oil or a similar cooling fluid. Both types of reactors are normally wound with stranded wire in order to reduce losses due to eddy currents and skin effect. In addition, it is important to avoid formation of short-circuited metal loops when building supporting structures for air-core reactors since these reactors usually produce large magnetic fields external to the coil. If these fields penetrate through closed-loop metal structures, induced currents will flow, causing both losses and heating of the structures. Which of these two reactor types should be used depends on the particular application, which also provides a reactor designation. *See* EDDY CURRENT; SKIN EFFECT (ELECTRICITY).

Ballast reactor. A ballast reactor consists of a coil wound on an iron core and connected in series with a fluorescent or vapor lamp. The coil compensates for a negative-resistance characteristic of the lamp by providing an increased voltage drop as the current through the lamp is increased. Normally, two coils are magnetically coupled (wound on the same core) to ensure an equal current distribution between two fluorescent lamps operating in parallel. *See* FLUORESCENT LAMP; VAPOR LAMP.

Commutating reactor. The commutating reactor is found primarily in silicon controlled rectifier (SCR) converters, where it is connected in series with a commutation capacitor to form a highly efficient resonant circuit. This circuit is used to cause a current oscillation that turns off (commutates) the conducting SCR. The inductor is normally made of litz wire to reduce losses caused by skin effect and is usually of air-core type. *See* CONTROLLED RECTIFIER; RECTIFIER; SEMICONDUCTOR RECTIFIER.

Current-equalizing reactor. This type is used to achieve desired division of current between several circuits operating in parallel.

Figure 1 shows two interphase reactors connected between two parallel SCR converters. The reactor provides balanced system operation when both converters are conducting by acting as an inductive voltage divider. Since the net direct current through the reactor is zero, the reactor supports only the harmonic voltage generated by SCR switchings.

Figure 2 shows two magnetically coupled reactors used to force equal currents through two parallel conducting paths. The voltage induced across each
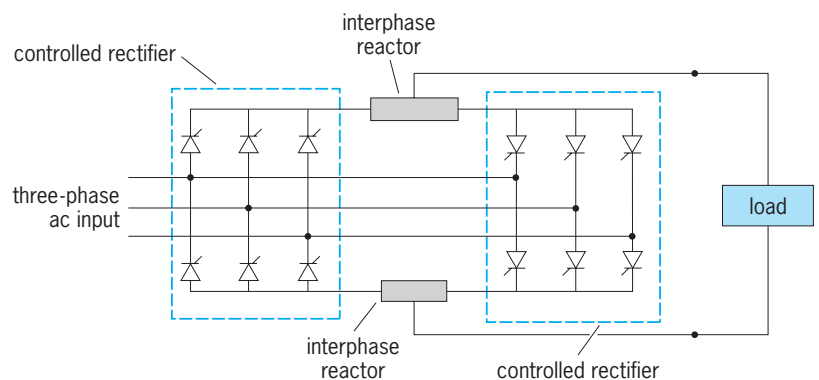

Fig. 1. **Two interphase reactors that enable simultaneous (circulating current) operation of two controlled rectifiers.**
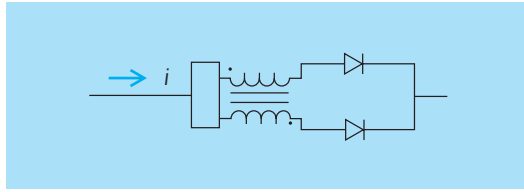
Fig. 2. Two magnetically coupled reactors used to force equal currents through two parallel diodes.

inductor is proportional to the difference between currents through the two diodes and with a polarity which tends to equalize the two currents. Since the reactors should be as linear as possible and have minimum hysteresis, magnetic materials are not used in the cores.

Current-limiting reactor. This type is used for protection against excessively large currents under short-circuit or transient conditions. By increasing the circuit's time constant (defined as L/R, where R is the total resistance), an inductor decreases the rate at which the current can change. In this way, a reactor provides more time for various protection devices (such as circuit breakers) to act before the current reaches a dangerous level. *See* ELECTRIC PROTECTIVE DEVICES.

Current-limiting reactors are always connected in series with a circuit which they are protecting, are of the air-core type, and can be divided into the following groups.

Bus reactor. This is an air-core inductor connected between two buses or two sections of the same bus in order to limit the effects of voltage transients on either bus. During normal operation both buses are at approximately the same voltage, and a free exchange of current takes place. As a fault occurs, the voltage on that bus abruptly changes and a heavy current starts to flow from one to the other bus. At that instant, a significant voltage drop is developed across the reactor, the magnitude of which depends on the rate of change of the current and on the reactor size (inductance). This voltage drop limits the current flow, minimizing the effect of the fault on the rest of the system and allowing a circuit breaker to disconnect the faulted section.

The di/dt reactor. This is a small inductor connected in series with a silicon controlled rectifier in order to limit the rate of rise of the SCR current. In small static converters, the *di/dt* reactor consists of an air-core inductor. In larger units, a saturable reactor is used, with a saturation point being just below the SCR rated current.

Feeder reactor. This is essentially the same type (but usually of higher power rating) used as a bus reactor. It is connected in series with a feeder circuit in order to limit and localize disturbances due to faults on the feeder. The principle of operation is the same as for bus reactors.

Generator reactor. This is the same as the feeder reactor but is connected between power plant generators and the rest of the network.

Starting reactor. This type is normally used to limit starting current of electric motors. A starting reactor

usually consists of iron-core inductors connected in series with the machine stator winding. The starting reactor is short-circuited by a switch once the motor reaches a predetermined speed.

Filter reactor. This type is used to attenuate (filter) high-frequency currents. Since the voltage drop across an inductor is equal to a product between inductance, current, and frequency, the inductor opposes the flow of high-frequency currents, while a direct current is limited only by the resistance of the wire used to make the reactor. Filter reactors are connected in series and are usually wound on an iron core with an air gap to prevent saturation. Filter reactors are used extensively in both alternating-current and direct-current circuits, especially in conjunction with static power converters. *See* ELECTRIC FILTER.

Grounding reactor. This type is used in grounded power systems and connected between the system's neutral point and a ground. The main purpose of a grounding reactor is to limit disturbances due to ground faults and atmospheric discharges on high-voltage transmission lines. A special grounding reactor, tuned to the rest of the network, which helps to extinguish a line-to-ground fault current, is called a Petersen coil. All grounding reactors are made of heavy cable that is wound on an iron core. *See* GROUNDING.

Saturable reactor. This is an iron-core reactor, adjusted to saturate at a predetermined current (**Fig. 3**), and used primarily in alternating-current circuits. Since the slope of the voltage-current curve (Fig. 3) is proportional to the reactor inductance, saturation substantially decreases this inductance, meaning that any further increase in the reactor current will result in very little change in the voltage across the reactor. Saturable reactors are used in a wide variety of circuit, whenever a nonlinear inductance is needed. By placing an additional, direct-current-supplied, control winding on the reactor core, one obtains a magnetic amplifier (amplistat or magamp). The direct current is used to control the point at which the reactor saturates and thus the voltage drop across the reactor. Magnetic amplifiers are often used in conjunction with diode rectifiers to regulate dc power flow, but are now being replaced by semiconductor (thyristor or silicon controlled rectifier) converters. *See* SATURABLE REACTOR.
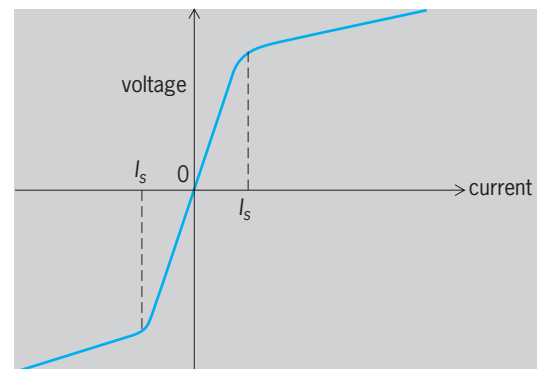

Fig. 3. Voltage-current characteristic of a saturable reactor; I_s is the current level at which the inductor saturates.
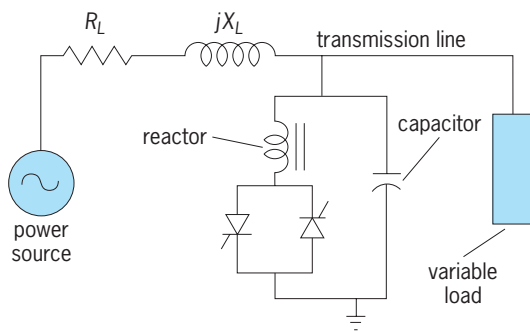
Fig. 4. Simplified single line diagram of a static reactive power system. Silicon controlled rectifiers are used to regulate the current through the reactor. R_L and X_L are the equivalent resistance and reactance of the transmission line.

Shunt reactor. A shunt reactor consists of an iron-core inductor and is used to compensate for capacitance in a transmission line. For this reason, it is also called a compensating reactor. Shunt reactors are normally connected at selected points between a transmission line and a ground. Examples of its application are the following:

1. Compensating reactors, connected at predetermined points along an underground transmission line to compensate the leading (capacitive) characteristic of an underground cable.

2. The Pupin coil, connected at regular intervals on long telephone lines to compensate line capacitance and to provide a constant line characteristic (impedance) at lower frequencies.

3. A parallel reactor in a static reactive power system (**Fig. 4**), used in conjunction with silicon controlled rectifiers to regulate voltages on an extra-high-voltage (EHV) overhead transmission line. The regulation is achieved by controlling the amount of current flow through the reactor and is used to prevent voltage variations caused by switching loads.

Tesla reactor. Also known as a tesla coil, this consists of two magnetically coupled coils wound on an air core. The primary winding is made of a few turns of heavy wire, while the secondary consists of fine wire with a very large number of turns. By circulating a high-frequency current in the primary, very high voltage is induced in the secondary. The tesla coil is used in radio and television receivers as well as for producing high discharge voltages. *See* INDUCTION COIL; TRANSFORMER.

Victor R. Stefanovic

Bibliography. B. J. Baliga, *Power Semiconductor Devices*, PWS, 1995; S. B. Dewan and A. Straughen, *Power Semiconductor Circuits*, Wiley, 1975; A. J. Pansini, *Power Transmission and Distribution*, CRC Press, 2d ed., 2005.

Reactor physics

The science of the interaction of the elementary particles and radiations characteristic of nuclear reactors with matter in bulk. These particles and radiations include neutrons, beta (β) particles, and gamma (λ) rays of energies between zero and about 10^7 eV. *See* BETA PARTICLES; GAMMA RAYS.

The study of the interaction of beta and gamma radiations with matter is, within the field of reactor physics, undertaken primarily to understand the absorption and penetration of energy through reactor structures and shields. For a discussion of problems of reactor shielding *see* RADIATION SHIELDING

With this exception, reactor physics is the study of those processes pertinent to the chain reaction involving neutron-induced nuclear fission with consequent neutron generation. Reactor physics is differentiated from nuclear physics, which is concerned primarily with nuclear structure. Reactor physics makes direct use of the phenomenology of nuclear reactions. Neutron physics is concerned primarily with interactions between neutrons and single nuclei or with the use of neutron beams as analytical devices, whereas reactor physics considers neutrons primarily as fission-producing agents. In the hierarchy of professional classification, neutron physics and reactor physics are ranked as subfields of the more general area of nuclear physics. *See* NEUTRON; NUCLEAR PHYSICS.

Reactor physics borrows most of its basic concepts from other fields. From nuclear physics comes the concept of the nuclear cross section for neutron interaction, defined as the effective target area of a nucleus for interaction with a neutron beam. The total interaction is the sum of interactions by a number of potential processes, and the probability of each of them multiplied by the total cross section is designated as a partial cross section. Thus, a given nucleus is characterized by cross sections for capture, fission, elastic and inelastic scattering, and such reactions as (n,p), (n,α) and $(n,2n)$. An outgrowth of this is the definition of macroscopic cross section, which is the product of cross section (termed microscopic, for specificity) with atomic density of the nuclear species involved. The symbols $N\sigma^j_i$ or Σ^j_i are used for macroscopic cross section, the subscript referring to the nuclear reaction involved, and the superscript to the isotope. The dimensions of Σ are cm^{-1}, and Σ_t (total cross section) is usually of the order of magnitude of unity. *See* NEUTRON SPECTROMETRY.

Cross sections vary with energy according to the laws of nuclear structure. In reactor physics this variation is accepted as input data to be assimilated into a description of neutron behavior. Common aspects of cross section dependence, such as variation of absorption cross section inversely as the square root of neutron energy, or the approximate regularity of resonance structure, form the basis of most simplified descriptions of reactor processes in terms of mathematical or logical models.

The concept of neutron flux is related to that of macroscopic cross section. This may be defined as the product of neutron density and neutron speed, or as the rate at which neutrons will traverse the outer surface of a sphere embedded in the medium, per unit of spherical cross-sectional area. The units of flux are neutrons/(cm^2)(s). The product of flux and

macroscopic cross section yields the reaction rate per unit volume and time. The use of these variables is conventional in reactor physics.

The chain reaction, a concept derived from chemical kinetics, is the basis of a physical description of the reactor process. The source of energy in a nuclear reactor is the fission of certain isotopes of heavy elements (thorium, uranium, and plutonium, in particular) when they absorb neutrons. Fission splits the elements into two highly radioactive fragments, which carry away most of the energy liberated by the fission process (about 160 MeV out of about 200 MeV per fission), the bulk of which is rapidly transformed into heat. Neutrons of high energy are also liberated, so that a chain of events alternating between neutron production in fission and neutron absorption causing fission may be initiated. Because more than one neutron is liberated per fission, this chain reaction may rapidly branch out to produce an increasing reaction rate, or divergent reaction; or if the arrangement of materials is such that only a small fraction of the neutrons will ultimately produce fission, the chain will be broken in a convergent reaction. *See* NUCLEAR FISSION.

Criticality. The critical condition is what occurs when the arrangement of materials in a reactor allows, on the average, exactly one neutron of those liberated in one nuclear fission to cause one additional nuclear fission. If a reactor is critical, it will have fissions occurring in it at a steady rate. This desirable condition is achieved by balancing the probability of occurrence of three competing events: fission, neutron capture which does not cause fission, and leakage of neutrons from the system. If ν is the average number of neutrons liberated per fission, then criticality is the condition under which the probability of a neutron causing fission is $1/\nu$. Generally, the degree of approach to criticality is evaluated by computing k_{eff}, which is the ratio of fissions in successive links of the chain, as a product of probabilities of successive processes.

Reactor constants are the parameters used in determining the probabilities of the various processes which together define k_{eff}. They comprise two sets, those used to characterize nuclear events, and those used to characterize leakage. In the former group are fast effect ϵ, resonance escape probability p, thermal utilization factor f, and neutrons emitted per fuel absorption η. In the latter group are neutron age or slowing-down area τ or L^2_s, migration area M^2, thermal diffusion area L^2, diffusion coefficient D, and buckling B^2.

Fast effect. The fast effect occurs in thermal reactors containing significant quantities of ^{238}U or ^{232}Th. These reactors comprise the bulk of plutonium production and civilian power reactors. The isotopes mentioned can undergo fission only when struck by very energetic neutrons; only a little more than one-half of the neutrons born in fission can cause fission in them. Moreover, these fast neutrons (energy greater than about 1.4 MeV) are subject to energy degradation by the competing reaction of inelastic scattering and by collision with moderator. In con-

sequence, only a small number of fast fissions occur.

The fast effect is characterized by a ratio R of fast to nonfast fissions. From this ratio, a quantity ϵ, the fast effect, is derived, which determines the neutrons available to the chain reaction after the convergent chain of fast fissions has been completed, per neutron born of nonfast fission.

The common magnitude of ϵ varies between 1.02 and 1.05. R varies between 0 and 0.15 commonly. When $R < 0.01$, the fast effect is usually omitted from consideration.

Resonance escape probability. The resonance escape probability p is a significant parameter for the same group of reactors that have a significant fast effect and is defined as the probability that a neutron, in the course of being moderated, will escape capture by ^{238}U (or ^{232}Th) at any of the many energies at which the capture cross section is unusually high (resonance energies). The existence of this resonance absorption prevents most reactors homogeneously fueled with natural uranium from going critical, because resonance absorption, coupled with other losses, causes neutron depletion below the requirement for criticality. Therefore, low-enrichment reactors are heterogeneous, the fuel being deposed in lumps. The lumping of fuel, originally proposed by E. Fermi and E. P. Wigner, increases the probability of resonance escape. Spatial isolation of resonance absorber from moderator increases the number of neutrons which are slowed down without making any collisions with the absorber. Also, because absorption is very probable at resonance energies, the neutrons can travel only very short distances into the fuel lump before being absorbed, so that the interior of the lump is hardly exposed to resonant neutrons. Another effect of lumping is the removal of excess neutron scattering near the absorber so that there is a lesser probability of absorption following multiple collision.

With slight enrichment, homogeneous assemblies can be made critical. The homogeneous problem also provides the formalism by which p can be calculated. Use is made of the resonance integral RI, defined as the absorption probability per absorbing nucleon per neutron slowed down from infinite source energy in a moderator of unit slowing-down power. In a highly dilute system the resonance interval is given by Eq. (1), where σ_a is the neutron absorp-

$$RI = \int_{E_0}^{\infty} \sigma_a \frac{dE}{E} \qquad (1)$$

tion cross section of the absorber as a function of energy, and E_0 is an energy taken as the lower limit of the resonance region. The slowing-down power of a moderator nucleus is $\xi \sigma_s$, where ξ is the mean increase in lethargy of the neutron per scattering, and σ_s is the moderator scattering cross section. Thus the resonance absorption probability of a system would appear to be given by Eq. (2), where N_r and N_m are,

$$1 - p = \frac{N_r(RI)}{N_m \xi \sigma_s} \qquad (2)$$

respectively, atomic concentrations of absorber and moderator in the reactor volume. However, because the probabilities of escaping capture by successive nuclei must be multiplied, Eq. (3) is better.

$$p = \exp\left[\frac{-N_r(RI)}{N_m \xi \sigma_s}\right] \qquad (3)$$

The value of E_0 used in the definition of *RI* follows one of two conventions: it is either taken as some defined thermal cut-off energy between about 0.3 and 1 eV, or as an energy just below the lowest resonance, about 6 eV for U and 20 eV for Th. In the latter case, *RI* is spoken of as $1/v$ corrected.

Thermal utilization factor. Thermal utilization factor *f* is the fraction of neutrons which, once thermalized, are absorbed in fuel. In a homogeneous array, *f* may be calculated from the atomic densities (that is, the number of atoms per cubic centimeter) and thermal-absorption cross sections of the various constituents of the reactor. The problem becomes more complex in a highly absorbing system and in the presence of nuclei (such as Pu or Cd) whose absorption cross sections do not vary with energy in the usual $1/v$ fashion. In this case, it becomes necessary to evaluate the neutron spectrum and average the absorption cross sections over this spectrum in order to obtain reaction ratios.

For low absorptions, the neutron spectrum is given by Eq. (4), the maxwellian expression, where

$$N(E)dE = \frac{2\pi}{(\pi kT)^{3/2}}E^{1/2} \times \exp\left(\frac{-E}{kT}\right)dE \qquad (4)$$

k is the Boltzmann constant and T the absolute temperature of the medium. Deviations from this shape caused by absorption were first formulated by Wigner and J. E. Wilkins. This problem has become very significant with the advent of highly absorbing systems operating with considerable quantities of Pu.

In heterogeneous systems the problem is further complicated by the spatial nonuniformity of the neutron flux. It is therefore necessary to calculate reaction rates in various regions of the lattice by multiplying local absorption cross sections and atomic densities by local neutron fluxes; or, in effect the same thing, by introducing flux weights into the cross sections.

The calculation of neutron flux may be performed by diffusion theory or by more exact and elaborate methods for solving the neutron-transport problem. The problem may be further complicated by spatial effects on spectrum.

The term disadvantage factor, applied in simple systems originally to describe the ratio of mean moderator to mean fuel flux, has fallen into disfavor because of vague and local definitions. *F*, called the fuel disadvantage factor, is still in use to describe the ratio of surface to mean volume flux in a fuel lump with isoperimetric flux. *See* THERMAL NEUTRONS.

Fission neutrons per fuel absorption. This constant, η, is a characteristic of the fuel and of the neutron spectrum, but not of the spatial configuration of the system. Pure fissionable materials do not always undergo fission when they absorb neutrons; sometimes they lose their energy of excitation by emission of a gamma ray. The number of neutrons per fission ν varies only slightly with incident neutron energy (a few percent per megaelectronvolts), but η can fluctuate considerably even within a fraction of an electronvolt. Consequently, the specification of η is dependent upon a good evaluation of neutron spectrum.

For unirradiated, low-enrichment assemblies, the custom of defining fuel as all uranium, ^{235}U and ^{238}U, persists. Thus, *f* considers total uranium captures, and η is lowered from the value for pure ^{235}U by the fractional absorption rate of ^{235}U in uranium. This custom leads to excessive complexity as plutonium builds into the fuel and is therefore declining in use.

Infinite multiplication constant. The infinite multiplication constant, k_∞, is the ratio of neutrons in successive generations of the chain in the absence of leakage. In the formalism just described, the chain is taken from thermal neutron through fission and back to thermal neutron and from the definition of terms, $k_\infty = \eta\epsilon pf$. This formula is known as the four-factor equation.

For reactors other than weakly absorbing thermal systems, the simplified description breaks down. Thus in a fast reactor significant fission and capture occur at all neutron energies, and no moderator is present; in a very strongly absorbing thermal system, an appreciable fraction of neutrons react at energies between the thermal region and the lowest ^{238}U resonance. For such systems, the definition of the neutron chain is usually made in terms of a total time-dependent fission-rate expression, and the parametric representation of k_∞ becomes appreciably more complex. Generally, the spectrum is broken up into energy groups, and k_∞ is defined as the sum of the fission neutron production rate over all groups divided by the sum of absorption rates over all groups.

Neutron age. The neutron age τ is a reactor parameter defined in various ways, all related to the probability of leakage of a fast neutron from a reactor system. The basic definition is that 6τ measures the mean square distance of travel between injection of a neutron at one energy or energy spectrum and its absorption at some other energy in an infinite system. The term age is used because in weakly moderating systems the equation describing neutron slowing down in space and energy has the same form as the time-dependent heat-conduction equation, with τ substituted for time. The British and Canadian usage is L_s^2 for slowing-down length (squared), which more accurately describes the physical parameter.

The common injection spectrum is a fission spectrum, and the common points of measurement or application are absorption at the indium resonance energy, 1.4 eV, or at some arbitrarily defined thermalization energy. The ages of these energies are denoted as τ_{In} or τ_{Th}. When a source other than the fission spectrum is considered, other subscripts are used.

A. M. Weinberg has pointed out that the shape of the spatial distribution in an infinite system for which τ is the second moment can be closely correlated with the fast leakage of a bare finite system. If the finite system has a source and sink distribution which is a solution of Laplace's equation, $\nabla^2\phi \times B^2\phi = 0$, then the Fourier transform of the τ distribution for given B will yield a quantity $P_\infty(B^2)$, which is almost exactly the nonleakage probability during slowing down. Two particularly significant cases are those for which the τ distribution is a gaussian or an exponential curve. In the former case, $P_\infty(B^2) = e^{-\tau B2}$; in the latter, $P_\infty(B^2) = 1/(1 + \tau B^2)$. The gaussian distribution is experimentally and theoretically verified for moderators as heavy as Be or heavier. The exponential distribution is crudely applicable to H_2O-moderated systems, and D_2O has a definitely mixed distribution.

In some cases, τ is used in a synthetic way to describe the leakage under some simple approximation to the slowing-down distribution. Thus, τ_{2G} is a number which is used in a two-energy group neutron model to give correct fast nonleakage probability as $P(B^2) = 1/(1 + \tau_{2G}B^2)$. When this model is not a good approximation, $6\tau_{2G}$ is not the second moment of the τ distribution.

For a heavy moderator, the age between two energies is given by the simple approximation of Eq. (5),

$$\tau_{E_1 \to E_2} = \int_{E_2}^{E_1} \frac{D(E)}{3(N\xi\sigma_s)} \frac{dE}{E} \qquad (5)$$

where D is the diffusion coefficient, N is atomic density, σ_s is scattering cross section, and ξ is mean lethargy (logarithmic energy) gain/collision.

Thermal diffusion area. The thermal diffusion area L^2 is one-sixth the mean square distance of travel between thermalization and absorption. It is thus the analog of τ for thermal neutrons. Because thermal migration is usually well represented by a diffusion equation, which has an exponential absorption distribution, thermal nonleakage probability is given by $1/(1 + L^2B^2)$. L^2 is defined by $L^2 = D_{Th}/N\sigma_{a,Th}$, where D and σ_a are spectrum-averaged diffusion coefficient and absorption cross section, respectively.

Migration area. The migration area M^2 is one-sixth the mean square distance of travel from birth to death of a neutron. For large, small-leakage systems, $1/(1 + M^2B^2)$ is an excellent approximation to the total nonleakage probability. When only thermal absorption exists, $M^2 = \tau + L^2$.

Diffusion coefficient. The diffusion coefficient D is essentially a scaling factor applied to validate Fick's law, a relation between neutron flux and current (the latter being defined as a vector describing net rate of flow of neutron density). Fick's law is expressed by Eq. (6), where $\mathbf{j}$ is current, $\nabla\phi$ flux gradient, and D

$$\mathbf{j} = -D\nabla\phi \qquad (6)$$

diffusion coefficient. For generalized systems, D is a tensor, but in regions where $|\nabla\phi|/\phi$ is small, D is approximated by a scalar of magnitude $1/(3\Sigma)$.

Buckling, B^2, is mathematically defined as $\nabla^2\phi/\phi$ in any region of a reactor where this quantity is constant over an appreciable volume. The name is derived from the relationship between force and deflection in a mechanical system with constrained boundaries. In a bare reactor, the buckling is constant except within one neutron mean free path ($\lambda \equiv 1/\Sigma$) of the boundary. For a slab of width t, $B^2 = (\pi/t)^2$; for an infinitely high cylinder of radius a, $B^2 = (2.404/a)^2$; for a sphere of radius a, $B^2 = (\pi/a)^2$; and for other geometries, it is again a geometrical constant. Because B^2 is a definite eigenvalue of Laplace's equation, it is the appropriate number to be used in the formulations of $P(B^2)$ previously described.

Effective multiplication k_{eff} is the ratio of neutron production in successive generations of the chain reaction. It is given by the product of k_∞ and $P(B^2)$, where $P(B^2)$ is itself the product of fast and thermal (or equivalent) nonleakage probabilities.

Reactivity. Reactivity is a measure of the deviation of a reactor from the critical state at any frozen instant of time. The term reactivity is qualitative, because three sets of units are in current use to describe it.

Percent k and millikay are absolute units describing the imbalance of the system from criticality per fission generation. Because $k_{eff} = 1$ describes a critical system, one says that it is 1% super- or subcritical, respectively, if each generation produces 1.01 or 0.99 times as many neutrons as the preceding one. Millikay are units of $0.1\% k$, and are given plus sign for supercriticality and minus sign for subcriticality. In both cases, k_{eff} is the base, as in Eqs. (7). These units

$$\%k = \frac{100\,|k_{eff} - 1|}{k_{eff}}$$
$$\text{Millikay} = \frac{1000(k_{eff} - 1)}{k_{eff}} \qquad (7)$$

are used primarily in design and analysis of control rods.

Dollars describe reactivity relative to the mean fraction of delayed neutrons per fission. Because the delayed neutrons are the primary agents for permitting control of the reaction, supercriticalities of less than 1 dollar are considered manageable in most cases. Thus, there is 1 dollar to "spend" in maneuvering power level. The dollar is subdivided into 100 cents. Because the delayed neutron fraction is a function of both neutron energy and fissionable material, the conversion rate between dollars and % k varies among reactors.

Inhours are reactivity units based on rate of change of power level in low-power reactors. If a low-power reactor is given enough reactivity so that its level would steadily increase by a factor of e per hour (which is also known as a 1-h period), that much reactivity is 1 inhour (from inverse hour). It is only for very small reactivities, however, that reactivity in inhours may be obtained from reciprocal periods.

Reactivity is measured in inhours primarily by operators of steady-state reactors, in which only small reactivities are normally encountered.

Reflectors. Reflectors are bodies of material placed beyond the chain-reacting zone of a reactor, whose

function is to return to the active zone (or core) neutrons which might otherwise leak. Reflector worth can be crudely measured in terms of the albedo, or probability that a neutron passing from core to reflector will return again to the core.

Good reflectors are materials with high scattering cross sections and low absorption cross sections. The first requirement ensures that neutrons will not easily diffuse through the reflector, and the second, that they will not easily be captured in diffusing back to the core.

Beryllium is the outstanding reflector material in terms of neutronic performance. Water, graphite, D_2O, iron, lead, and ^{238}U are also good reflectors. The use of Be, H_2O, C, and D_2O as reflectors permits conversion of neutrons leaking at high energy into thermal neutrons diffusing back to the core. Because the reverse flow of neutrons is always accompanied by a neutron-flux gradient, these reflectors show characteristic thermal flux peaks outside the core. They are, therefore, desirable materials for research reactors, in which these flux peaks are useful for experimental purposes.

The usual measurement of reflector worth is in terms of reflector savings, defined as the difference in the reflected dimension between the actual core and one which would be critical without reflector. Reflector savings are close to reflector dimensions for thin reflectors and approach an asymptotic value dependent upon core size and reflector constitution as thickness increases.

Reactor dynamics. Reactor dynamics is concerned with the temporal sequence of events when neutron flux, power, or reactivity varies. The inclusive term takes into account sequential events, not necessarily concerned with nuclear processes, which may affect these parameters. There are basically three ways in which a reactor may be affected so as to change reactivity. A control element, absorbing rod, or piece of fuel may be externally actuated to start up, shut down, or change reactivity or power level; depletion of fuel and poison, buildup of neutron-absorbing fission fragments, and production of new fissionable material from the fertile isotopes ^{232}Th, ^{234}U, ^{238}U, and ^{240}Pu make reactivity depend upon the irradiation history of the system; and changes in power level may produce temperature changes in the system, leading to thermal expansion, changes in neutron cross sections, and mechanical changes with consequent change of reactivity.

Reactor control physics. Reactor control physics is the study of the effect of control devices on reactivity and power level. As such, it includes a number of problems in reactor statics, because the primary question is to determine the absorption of the control elements in competition with the other neutronic processes. It is, however, a problem in dynamics, given the above information, to determine what motions of the control devices will lead to stable changes in reactor output.

Particular problems occurring in the statics of reactor control stem from the particular nature of control devices. Many control rods are so heavily

absorbing for thermal neutrons that elegant refinements of neutron-transport theory are needed to estimate their absorption. Other types of control rods include isotopes with heavy resonance absorption, and the interaction of such absorbers with ^{238}U resonances must be examined. The motion of control rods changes the material balance of reactor regions, and with water-moderated reactors, peaks in the fission rate occur near empty rod channels. By virtue of high absorption of their constituent isotopes, some absorbing materials (for example, cadmium and boron) burn out in the reactor, and a rod made of these materials loses absorbing strength with time. As a final example, the motion of a control rod may change the shape of the power pattern in the reactor so as to bring secondary pseudostatic effects into play. *See* NUCLEAR REACTOR.

Reactivity changes. Long-term reactivity changes may represent a limiting factor in the burning of nuclear fuel without costly reprocessing and refabrication. As the chain reaction proceeds, the original fissionable material is depleted, and the system would become subcritical if some form of slow addition of reactivity were not available. This is the function of shim rods in a typical reactor. The reactor is originally loaded with enough fuel to be critical with the rods completely inserted. As the fuel burns out, the rods are withdrawn to compensate.

In order to decrease requirements on the shim system, many devices to overcome reactivity loss may be used. A burnable poison may be incorporated in the system. This is an absorbing isotope which will burn out at a rate comparable to or greater than the fuel. Burnable poisons are therefore limited to isotopes with very high effective neutron-absorption cross sections. Combinations of poisons, and the use of self-shielding of poisons can, in principle, make the close compensation of considerable reactivity possible without major control rod motions, but the technological problems in their use are formidable.

A more popular method for compensating reactivity losses is the incorporation of fertile isotopes into the fuel. This is desirable because the neutrons captured in the fertile material are not wasted, but used to manufacture new fissionable material; and also because (as with ^{238}U in ^{235}U reactors or ^{240}Pu in ^{239}Pu systems) the fertile material is normally found mixed with the fissionable, and isotopic separation may be circumvented or minimized. Depending on the conversion ratio (new fissionable atoms formed per old fissionable atom burned) and the fission parameters of the materials, a reactor so fueled loses reactivity relatively slowly, and in some cases, may show a temporary reactivity increase. The various isotopes produced by successive neutron capture in uranium and plutonium must also be considered, the higher isotopes becoming prominent at very long exposures.

A final consideration of long-term reactivity is the extra parasitic absorption of the fission products as formed. At long exposure, this absorption becomes significant because of the relatively high absorption

of many of the fission products. At shorter exposures, isotopes of very high absorption, mainly ^{149}Sm and ^{135}Xe, are more prominent. These materials have such high cross sections that they reach a steady-state concentration relatively quickly, burning out by neutron capture as rapidly as they are formed in fission.

^{135}Xe is particularly interesting because its cross section is abnormally large, its fission yield is high, it is preceded by an isotope of low cross section and approximately 7-h half-life (^{135}I), and it undergoes beta decay to the low absorption ^{135}Cs with a half-life of about 10 h. This combination of properties gives several interesting effects. Chief of these is that, at high flux, a reservoir of ^{135}I is formed which continues to decay to ^{135}Xe even after the reactor is shut down. Because ^{135}Xe is maintained at steady state during operation by a balance of buildup against burnout, the shutdown also removes the chief mode of ^{135}Xe destruction. Hence, the ^{135}Xe concentration increases immediately after shutdown. The reservoir of ^{135}I is so large that reactivity is rapidly lost as ^{135}Xe builds up; and in some cases, it may be impossible to restart the reactor after a short shutdown. The operator must then wait almost 2 days for the ^{135}Xe to disappear by radioactive decay. At very high fluxes, this effect becomes so severe that even a small temporary reduction in power may lead to ultimate subcriticality of the reactor.

Another effect caused by ^{135}Xe in high-flux reactors is that, if the reactor is large enough, the ^{135}Xe may force the power pattern into oscillations of 1- or 2-day periods. Although this is not a serious dynamic problem, it does emphasize the necessity of monitoring not only the total power, but also the power pattern, so that appropriate countermeasures may be taken. *See* NUCLEAR FUELS; NUCLEAR FUELS REPROCESSING.

Reactor kinetics. This is the study of the short-term aspects of reactor dynamics with respect to stability, safety against power excursion, and design of the control system. Control is possible because increases in reactor power often reduce reactivity to zero (the critical value) and also because there is a time lapse between successive fissions in a chain resulting from the finite velocity of the neutrons and the number of scattering and moderating events intervening, and because a fraction of the neutrons is delayed. *See* DELAYED NEUTRON.

Prompt-neutron lifetime. This is the mean time between successive fissions in a chain, and it is the basic quality which determines the time scale within which controlling effects must be operable if a reactivity excursion is touched off. In thermal reactors the controlling feature is the time between thermalization and capture, because fast neutrons spend less time between collisions, and the total time for moderation of a neutron is at most a few microseconds. Prompt-neutron lifetimes vary from 10 to a few hundred microseconds for light-water reactors, the shorter times being found in poorly reflected, highly absorbing systems, and the longer in well-reflected systems, in which time spent in the reflector is the dominating factor. Other thermal reactors have longer lifetimes,

with some heavy-water reactors having lifetimes as long as a few milliseconds. Fast reactors have lifetimes of the order of 0.01–0.1 μs, the controlling factor being the amount of scattering material used as diluent.

Delayed neutrons. Delayed neutrons are important because a complete fission generation is not achieved until these neutrons have been emitted by their precursors. A slightly supercritical reactor must wait until the delayed neutrons appear, and this delay allows time for the system to be brought under control. The fraction of delayed neutrons β ranges from about $^1/_3$% for thermal fission of ^{239}Pu and ^{233}U, to about $^3/_4$% for thermal fission of ^{235}U, to several percent for some fast fission events. In these latter cases, however, the extra delayed neutrons have such short lifetimes that they are of only slight extra utility.

In any case, the influence of delayed neutrons is felt only to the extent that they are needed to maintain criticality. When the system is supercritical enough that the delayed neutrons are not needed to complete the critical chain, it is known as prompt critical. Prompt criticality represents in a qualitative sense the threshold between externally controllable and uncontrollable excursions, and it is for this reason that the dollar unit is popular in excursion analysis (a prompt critical system has a reactivity of 1 dollar).

Although it has now been established that a larger number of fission products emit delayed neutrons, the distribution in time after fission of the delayed neutron emission rate is accurately represented for all purposes by a sum of six negative exponentials. For many purposes, however, a three, two, or one group approximation is adequate.

Reactors moderated by D_2O and Be have additional delayed neutrons contributed by photoneutron reactions between the moderator and fission product gamma rays. Although not a large fraction, this effect makes such reactors unresponsive to small reactivity fluctuations and gives them unusual operational smoothness.

Reactor period. This is the asymptotic time required for a reactor at constant reactivity to increase its power by a factor e. When a critical reactor is given extra reactivity, its power will rise. At first, the power production rate has a complex shape on a time plot, but ultimately the power will rise exponentially. The period is the measure of this exponential rate.

The relation between the reactor period and the reactivity is known as the inhour equation. If l is the reactor lifetime in seconds, β_i the fraction of delayed neutrons in group i, λ_i the delay constant of group i delayed neutrons in s^{-1}, S reactor period in seconds and ρ reactivity in thousands of millikay, then Eq. (8) is the inhour equation. The equation

$$\rho = \frac{l/S + \sum_i [\beta_i/(1 + \lambda_i S)]}{1 - \sum_i [\beta_i/(1 + \lambda_i S)]} \qquad (8)$$

has $I + 1$ solutions for S for a given ρ, I being the number of groups; and there is always a real value

of S with a higher value than the real part of any other solution. This highest S is the period. For very small values of ρ, that is, for very large periods, this value is approximatelythat given by notation (9). For

$$\frac{l + \sum\limits_{i}(\beta_i/\lambda_i)}{\rho} \qquad (9)$$

very large ρ, and therefore small S, the period is approximately $l/(\rho - \beta)$. This same result would be found if the delayed neutrons were thrown away completely.

Reactivity coefficients. There are several functions relating changes in reactivity to changes in the physical state of the reactor. The power coefficient is the change in reactivity per unit change in reactor power; the temperature coefficient relates reactivity to temperature change, and is often broken down into fuel, moderator, and coolant coefficients; for low-power graphite reactors there exists a barometric coefficient; one may define also coolant circulation rate coefficients and void coefficients.

Because the reactivity is commonly a complicated function of all the pertinent variables, the reactivity coefficient generally is the coefficient of the first term in a series expansion of the reactivity about the operating point. This in turn describes a linear theory of reactor dynamics. The theory may be extended to reactivity effects of arbitrary type by considering reactivity coefficients as functionals.

The basic problem of reactor dynamics is specifying the power coefficient of reactivity. The chain, power affects reactivity which affects power, is thus analyzable, using the power coefficient functional along with the reactor kinetic equations. The power coefficient is, however, predictable only in terms of changes in temperature and flow resulting from power changes in the system. The analysis of the power coefficient implies exhaustive knowledge of system behavior.

Reactivity coefficients may be prompt or delayed, and most delayed effects can be characterized as either of decay or transport type. An example of the decay type of coefficient is the contribution of coolant temperature change to power coefficient. Here, a power pulse gives a thermal effect on the coolant which is instantaneously observable, and which decreases exponentially with a time constant imposed by the heat-transfer equations. An example of a transport type of delay is the delay attributable to coolant circuit times. Here, a finite time lapse exists between the cause and the observable response. Effects due to fuel heating are examples of prompt effects.

Some types of power coefficient yield dangerous or unstable situations. Thus a power coefficient may contain a prompt positive (autocatalytic) term and a larger delayed negative term. Even though such a system may be stable against slow power-level increases, it will undergo a violent excursion whenever power is raised rapidly enough to outstrip delayed effects. Again, a system with prominent delayed effects of the transport type is always unstable beyond some critical power, even if the effect opposes the power shift; here, there is a possibility of phase instability.

The dynamic behavior of a reactor is usually analyzed by techniques common to all feedback systems. *See* CONTROL SYSTEMS; SERVOMECHANISM.

Bernard I. Spinrad

Bibliography. G. I. Bell and S. Glasstone, *Nuclear Reactor Theory*, 1970, reprint 1979; A. R. Foster and R. L. Wright, Jr., *Basic Nuclear Engineering*, 4th ed., 1983; S. Glasstone and A. Sesonske, *Nuclear Reactor Engineering*, 4th ed., 1993; D. L. Hetrick, *Dynamics of Nuclear Reactors*, 1993; J. R. Lamarsh, *Introduction to Nuclear Engineering*, 2d ed., 1983; R. Stammler and M. J. Abate, *Methods of Steady State Reactor Physics in Nuclear Design*, 1983; J. Weisman, *Elements of Nuclear Reactor Design*, 2d ed., 1983.

Reagent chemicals

High-purity chemicals used for analytical reactions, for the testing of new reactions where the effects of impurities are unknown, and in general for chemical work where impurities must either be absent or at known concentrations. If the concentration of impurity in any reagent is critical, an analysis should be made.

Methods of purification. Chemicals are purified by a variety of methods, the most common being recrystallization from solution. For many inorganic chemicals a saturated solution is prepared in water at the boiling point. After filtration to remove insoluble matter, the chemical crystallizes out of solution as it cools. The crystals are removed by filtration on a small scale or by centrifugation on a large scale, washed with water to remove impurity-containing solution on the surface, and dried. If the substance to be purified is not appreciably more soluble in hot solution than in cold solution, then one can use isothermal crystallization, the removal of solvent at constant temperature by reducing the pressure. The recrystallization process is not always successful as a purification method. In some cases precipitation of crystals from a saturated water solution by adding a second solvent, for example, ethyl alcohol, is the simplest method. One difficulty is that the impurity crystals may have the same structure as the desired ones, that is, the two crystals may be isomorphous. If water of hydration is present, storage in an atmosphere of known humidity may be necessary to obtain a definite hydrate. Occasionally it is easier to remove the impurity by dissolving it in a solvent in which the desired material is not very soluble.

If the desired chemical is volatile and the impurities are not volatile, sublimation is an effective method of purification. For example, iodine and arsenious oxide are easily purified by sublimation. For liquid chemicals, distillation is an effective procedure. Finally, the simplest procedure may be to synthesize the desired reagent from pure materials; for example, the addition of pure ammonium carbonate solution

to pure calcium chloride solution precipitates pure calcium carbonate.

Standards of purity. Commercial chemicals are available at several levels of purity. Chemicals labeled "technical" or "commercial" are usually quite impure. The grade "USP" indicates only that the chemical meets the requirements of the *United States Pharmacopeia*. The term "CP" means only that the chemical is purer than "technical." Chemicals designated "reagent grade" or "analyzed reagent" are specially purified materials which usually have been analyzed to establish the levels of impurities. The last two classes are the ones usually used in the laboratory. The American Chemical Society has established specifications and tests for purity for some chemicals. Materials which meet these specifications are labeled "Meets ACS Specifications."

A special group of extremely pure chemicals are called "primary standard" reagents. These reagents are usually readily available, easily purified, and unreactive with components of air such as water and carbon dioxide. The total sum of impurities should be less than 0.02%. These reagents are used to determine the concentrations of solutions used in volumetric analysis or for other purposes in which impurities must be quite low in concentration.

Care is necessary to prevent contamination by dust or by other chemicals. Transfers should be made directly from the container by pouring and not with spatulas or other tools. No material should be returned to the container. The maintenance of purity in an opened container is a problem.

Selective and specific reagents. Chemical reagents are often classified on the basis of utility. Many chemicals are general reagents, that is, they may react with many others. For example, any acid will be neutralized to some extent by a base. However, there are some reagents which react with only a limited number of other chemicals. These reagents are called selective reagents. The silver salts of chloride, bromide, iodide, thiocyanate, and a few other ions are insoluble in water; therefore silver nitrate is a selective precipitating reagent for these ions. A limited number of reagents are known which react appreciably with only one ion under specified conditions. These reagents are called specific reagents. The specific property is determined by the product formed. The ion dimensions, the charge densities, and the electron arrangements of the reagent and the ion which react must fall within certain limits or no reaction will occur. Variations in the acidity of the solution and the presence of other ions can change the condition of the ion so that no reaction can occur. Most inorganic reagents are at best selective. Most specific reagents are organic in nature.

These organic reagents have two types of reactive groups. One group forms electrovalent bonds by charge neutralization, and the second group forms covalent bonds by sharing electrons. The products are cyclic because both bonds are to the same ion, and they are called chelate compounds. These chelates are frequently very insoluble in water and intensely colored, making them very useful in anal-

ysis. *See* CHEMICAL SEPARATION TECHNIQUES; CRYSTALLIZATION; DISTILLATION.   Kenneth G. Stone
Charles Rulfs

Bibliography. American Chemical Society, *Reagent Chemicals: ACS Specifications*, 9th ed., 1999; L. Fieser and M. Fieser (eds.), *Fieser and Fieser's Reagents for Organic Synthesis*, vols. 9 and 10, 1981–1982; M. Fieser, *Reagents for Organic Synthesis*, 8 vols., 1980.

Real variable

A variable whose range is a subset of the real numbers. By extension the term is also used to refer to the theory of functions of one or more real variables. This theory has to do with properties of broad classes of functions, such as continuity, types of discontinuities, differentiability of functions, oscillation and variation of functions, and the various kinds of integrals. *See* INTEGRATION.

Real numbers. Real numbers are those commonly used in the geometric theory of measurement. The integers and fractions, also called rational numbers, are included among the real numbers. In practice an irrational number x is specified by telling which rational numbers are less than x and which are greater than x. Such a division of the rational numbers into two classes was used by J. W. R. Dedekind as the formal definition of a real number and is called a Dedekind cut.

The system of real numbers has the familiar algebraic properties and is also ordered. This order is related to the algebraic operations of addition and multiplication by the following properties:

If $a < b$, then $a + c < b + c$ for every number c.

If $a < b$, then $ac < bc$ for every number $c > 0$.

If $a > 0$ and $b > 0$, there exists an integer n such that $a < nb$.

Various subsets of the real numbers may be defined by means of inequalities. The set of numbers x satisfying the inequalities $a < x < b$ is called an open interval and denoted by (a,b). Here a and b are the ends of the interval but are not included in it. When the left end a is included, the notation is changed to $[a,b)$. The interval $[a,b]$, including both ends, is called a closed interval. A more complicated set, called the Cantor discontinuum, which is very frequently of use in the theory of functions, may be constructed as follows. From the interval $[0,1]$ remove the open intervals $(1/3, 2/3)$, $(1/9, 2/9)$, $(2/3 + 1/9, 2/3 + 2/9)$, and in fact all intervals $(c - 1/3^n, c)$, where $c = d_1/3 + d_2/3^2 + \cdots + d_{n-1}/3^{n-1} + 2/3^n$ and each $d_i = 0$ or 2. It may be proved that the remaining set contains just as many numbers as the original interval $[0,1]$. Another set to which reference will be made consists of all the rational numbers x satisfying the inequality $s^2 < 2$. *See* INFINITY.

An upper bound for a set S of numbers is a number U such that $x \leq U$ for every x in S, and a lower bound is a number L such that $L \leq x$ for every x in S. A set need not have an upper bound, for example, the set

of all rational numbers. A least upper bound for S is an upper bound U such that V is not an upper bound when $V < U$. There cannot be more than one least upper bound for a set of real numbers. A distinguishing property of the real number system states that every set having an upper bound also has a least upper bound in the system. This property fails for the system consisting only of the rational numbers, as is shown by the fact that the least upper bound of the set consisting of the rational numbers x such that $x^2 < 2$ is the irrational number $\sqrt{2}$.

Functions. Consider a function $f(x)$ defined on an interval $a \leqq x \leqq b$. By the oscillation of lf on a subinterval $[c,d]$ of $[a,b]$ is meant the number $\omega(c,d) =$ least upper bound of $|f(x_1) - f(x_2)|$ for x_1 and x_2 in $[c,d]$. The function f is continuous at a point x of $[a,b]$ in case $\omega(c,d)$ tends to zero when c and d approach x, and f is continuous on $[a,b]$ when it is continuous at every point of $[a,b]$. A function f continuous on a closed interval $[a,b]$ attains a greatest and a least value on $[a,b]$. Moreover, it takes every value between $f(a)$ and $f(b)$. If f has a derivative f' on the interval $[a,b]$, then f' also takes every value between each pair of its values, although f' may fail to be continuous. However, a continuous function f may fail to have a derivative at any point whatever. This suggests a useful classification of functions according to the number of continuous derivatives they possess.

The variation of a function f on an interval $[a,b]$ is defined to be the least upper bound V (a,b) of the sums shown in expression (1) for

$$\sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \qquad (1)$$

all finite sets of points $x_0, \ldots, x_n$ satisfying $x_0 = a$, $x_{i-1} < x_i$, $x_n = b$. One can write $V(a,b) = +\infty$ when no upper bound exists, and in all other cases one says that f has bounded variation. A continuous function f may fail to have bounded variation—for example, relation (2) on the interval $[0,1]$. On

$$f(x) = x \sin\left(\frac{1}{x}\right) \qquad 0 < x \leq 1$$
$$f(0) = 0 \qquad (2)$$

the other hand, a function of bounded variation may be discontinuous; however, it always has right-hand and left-hand limits at each point, so all its discontinuities are jump discontinuities. This follows readily from the fact that a function of bounded variation always is equal to the difference of two nondecreasing functions. A very remarkable theorem states that every function f of bounded variation has a derivative $f'(x)$, except at the points of a set which can be enclosed in a sequence of intervals the sum of whose lengths is arbitrarily small.

If S is a set composed of nonoverlapping intervals $[c_k,d_k]$, the variation $V(S)$ of f over S may be defined as the sum of the variations $V(c_k,d_k)$. If $V(S)$ tends to zero with the sum of the lengths of the intervals, the function f is said to be absolutely continuous on $[a,b]$. An absolutely continuous function f admits a

generalization of the fundamental theorem of integral calculus, as shown in Eq. (3), where the integral is a Lebesgue integral.

$$f(b) - f(a) = \int_a^b f'(x)\, dx \qquad (3)$$

Continuous functions may be characterized in a way different from the definition given above. Thus a function f defined on a finite closed interval $[a,b]$ is continuous on $[a,b]$ if and only if for every positive number ϵ there exists a polynomial function $p(x)$ which approximates $f(x)$ with an error less than ϵ on the entire interval $[a,b]$, that is, $|f(x) - p(x)| < \epsilon$ for $a \leqq x \leqq b$. A variation of this condition (which may seem minor) yields a much more restricted class of functions than the continuous functions—that is, if one requires the approximating polynomials to be of the form shown in Eq. (4), where the coefficients c_k

$$p(x) = \sum_{k=0}^{n} c_k(x - x_0)^k \qquad (4)$$

belong to a fixed infinite sequence $c_0, c_1, c_2, \ldots$, and x_0 is a point of $[a,b]$, the function f will have continuous derivatives of all orders and will have many other properties which are best understood when the variable x is allowed to be complex. On the other hand, if it is required only that there exist a sequence of polynomials p_n such that Eq. (5) holds for each

$$\lim_{n \to \infty} p_n(x) = f(x) \qquad (5)$$

x on $[a,b]$, then $f(x)$ may be discontinuous. Such a function f is said to belong to Baire's class 1. A sequence f_n of functions in Baire's class 1 may converge to a discontinuous function not in the class 1. This is the beginning of an infinite collection of classes of discontinuous functions. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.   Lawrence Murray Graves

Bibliography. S. K. Berberian, *A First Course in Real Analysis*, 1994; S. A. Lang, *Real Analysis*, 3d ed., 1993; M. H. Protter and C. B. Morrey, *A First Course in Real Analysis*, 2d ed., 1993; H. L. Royden, *Real Analysis*, 3d ed., 1988; W. Rudin, *Real and Complex Analysis*, 3d ed., 1987; A. Torchinsky, *Real Variables*, 1988.

Real-time systems

Computer systems in which the computer is required to perform its tasks within the time restraints of some process or simultaneously with the system it is assisting. Usually the computer must operate faster than the system assisted in order to be ready to intervene appropriately.

Types of systems. Real-time computer systems and applications span a number of different types. Reactive real-time systems have an ongoing interaction with their external environment. Embedded real-time systems control specialized tasks in the system in which they are incorporated. Synchronous real-time systems control events at predictable time

intervals, whereas asynchronous real-time systems react to their environment to control or respond to unpredictable events requiring something from the real-time system.

Examples. At the race track in the last few minutes before a race, most bets are placed; the asynchronous real-time computer system must record the bets and update the betting odds for up to tens of thousands of bets per minute. In storm weather tracking, the real-time system must obtain data from up to hundreds of thousands of sensors per local area and process them to arrive at multiple weather forecasts per minute. Factory just-in-time real-time scheduling of machines and transportation occurs when trucks arrive with materials and subassemblies which need to be assembled into final products, packaged, palletized and loaded onto trucks, without intermediate warehousing, ready for delivery to end customers. Personal computer users require that their keystrokes and mouse actions are handled immediately (in the real time of human thinking and action). Weapons fire control requires microsecond real-time control. At the supermarket, real-time inventory restocking is synchronized by taking into account sales (as registered from bar code readers at the check-out counter), just-in-time ordering and delivery (time-lag scheduling), and shelf stocking timing. New jet engine design testing requires testing a strapped-down prototype engine in the factory by running it to its extreme limits; the real-time system must process thousands of sensor data inputs per microsecond to control the experiment and to shut down the engine just before it is about to fly to pieces. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING; MICROCOMPUTER; STORM DETECTION.

Real-time control and real-time process control. In these applications the computer is required to process systems data (inputs) from sensors for the purpose of monitoring and computing system control parameters (outputs) required for the correct operation of a system or process. The type of monitoring and control functions provided by the computer for subsystem units ranges over a wide variety of tasks, such as turn-on and turn-off signals to switches; feedback signals to controllers (such as motors, servos, and potentiometers) to provide adjustments or corrections; steering signals; alarms; monitoring, evaluation, supervision, and management calculations; error detection, and out-of-tolerance and critical parameter detection operations; and processing of displays and outputs. *See* CONTROL SYSTEMS; PROCESS CONTROL.

Real-time assistance. Here the computer is required to do its work fast enough to keep up with a person interacting with it (usually at a computer terminal device of some sort, for example, a screen and keyboard). These are people-amplifier-type real-time computer systems. The computer supports the person or persons interacting with it and provides access, retrieval, and storage functions, usually through some sort of database management system, as well as data processing and computational power. System access allows the individual to intervene (control, adjust, supply parameters, direct, and so forth) in the system's operation. The real-time computer also often provides monitoring or display information, or both. *See* DATABASE MANAGEMENT SYSTEM; HUMAN-COMPUTER INTERACTION; MULTIACCESS COMPUTER.

Real-time robotics. In this case the computer is a part of a robotic or self-contained machine. Often the computer is embedded in the machine, which then becomes a smart machine. If the smart machine also has access to, or has embedded within it, artificial intelligence functions (for example, a knowledge base and knowledge processing in an expert system fashion), it becomes an intelligent machine. *See* ARTIFICIAL INTELLIGENCE; EMBEDDED SYSTEMS; EXPERT SYSTEMS; ROBOTICS.

Evolution. Real-time computer systems have been evolving constantly since the interrupt function, which allowed the computer to be synchronized with the external world, was invented. There are five primary paths along which real-time systems continue to advance:

First-generation real-time control systems comprise process control (such as an oil refinery); guidance and control (such as antiballistic missile and intercontinental ballistic missile systems); numerical control (such as factory machine operations); dedicated (mini) computer systems; and store-and-forward message switching. *See* DATA COMMUNICATIONS; GUIDANCE SYSTEMS.

Second-generation real-time computer systems comprise time-shared and multiprocessor computing; interactive computing; smart and intelligent terminals; and computer networks (distributed computers, distributed smart machines, and distributed intelligent machines). *See* MULTIPROCESSING.

Third-generation real-time assistance systems comprise operating systems; CAD (computer-aided design), CAM (computer-aided manufacturing), CAI (computer-assisted instruction), MIS (management information systems), and DSS (decision support systems); personal computers; word processing and work stations; and artificial intelligence expert systems. *See* DECISION SUPPORT SYSTEM; INFORMATION SYSTEMS ENGINEERING; OPERATING SYSTEM; WORD PROCESSING.

Fourth-generation real-time machines comprise smart machines with embedded computers; intelligent machines; and robots (dumb, smart, and intelligent).

Fifth-generation real-time integrated systems comprise the factory-of-the-future (totally automated factories); just-in-time (JIT) systems (in the factory and in distribution); computer utilities and knowledge utilities; and knowledge inference processing systems. *See* COMPUTER-INTEGRATED MANUFACTURING.

Artificial intelligence and expert systems. Artificial intelligence (AI) is rapidly advancing beyond the research stage into practical use by scientists, management, and many other areas of business and society. One branch is directed toward the development of real-time expert systems. Thus, widespread use of artificial intelligence, knowledge bases, expert

systems, and people-amplifiers is expected to develop, and decision making will be amplified with real-time and intelligent computer systems. *See* ARTIFICIAL INTELLIGENCE; EXPERT SYSTEMS.

Real-time simulation. The real-time computer can serve as a tool allowing simulation of models of the real world. By coupling this tool with artificial-intelligence knowledge-base expertise, scientific experimentation becomes possible without cumbersome laboratory equipment and procedures. Scientists can play serious real-time experimental and mathematical games with the object of their research without first needing to learn sophisticated laboratory techniques or mathematics. The detailed mathematical and discipline-oriented experimental skills and procedures are embedded within the computerized model. Such robot simulators allow the scientist to concentrate upon the investigation, rather than the scientist becoming buried within the relevant mathematics and discipline crafts. However, the scientist must first learn his or her field plus the computer simulation/modeling language. Thus, robot simulator assistants increasingly do for the scientist what the calculator does for the average person: they remove the need to perform bulky, precise, and rote skill functions, allowing the researcher to get more quickly and easily to the core of scientific investigation.

With robot simulators the researcher can ask questions of a computer modeling system, and have simulated experiments performed that are otherwise nearly impossible or too time-consuming and costly. For example, the researcher can ask "What happens if . . . ?" After performing the simulated experiment, the computer gives an answer, while the experimenter views the progress of the computerized experiment and intervenes when desired. Then the experimenter can ask "What happens if something else is done instead?" to arrive at a different comparable answer from the robot simulator. Through such real-time interactive simulations the researcher becomes directly involved in the experiment as a surrogate participant.

Future robot simulators are expected to take the form of advanced hand-held calculators with voice dialog and artificial intelligence capabilities, and to contain specialized expert knowledge making them capable of general decision making. Such smart robot simulators will be used as people-amplifiers or electronic assistants by managers, programmers, doctors, politicians, voters, and others. *See* COMPUTER; DIGITAL COMPUTER; SIMULATION.   Earl C. Joseph

Bibliography. S. T. Allworth and R. N. Zobel, *Introduction to Real-Time Software Design*, 2d ed., 1990; J.-P. Calvez, *Embedded Real-Time Systems: A Specification and Design Methodology*, 1993; S. Goldsmith, *A Practical Guide to Real-Time Systems Development*, 1993; P. A. Laplante, *Real-Time Systems Design and Analysis: An Engineer's Handbook*, 2d ed., 1997; M. Schiebe (ed.) *Real-Time Systems Engineering and Applications*, 1992; C. Vickery, *Real-Time Systems Programming for PCs*, 1993.
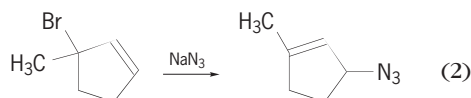
Rearrangement reaction

A reaction in which an atom or bond moves or migrates, having been initially located at one site in a reactant molecule and ultimately located at a different site in a product molecule. A rearrangement reaction may involve several steps, but the key feature defining it as a rearrangement is that a bond shifts from one site of attachment to another. The simplest examples of rearrangement reactions are intramolecular, that is, reactions in which the product is simply a structural isomer of the reactant [reaction (1)].

$$O \longrightarrow O \qquad (1)$$
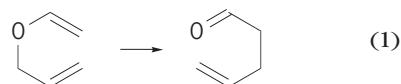
See MOLECULAR ISOMERISM.

More complex rearrangement reactions occur when the rearrangement is accompanied by another reaction, for example, a substitution reaction (2).
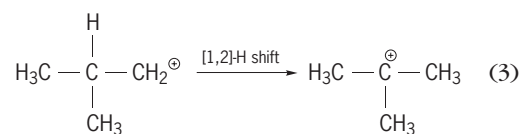
$$\underset{H_3C}{\overset{Br}{\diagdown}} \xrightarrow{NaN_3} \underset{H_3C}{\overset{H_3C}{\diagdown}} N_3 \qquad (2)$$

Classification. Rearrangement reactions are classified and named on the basis of the group that migrates and the initial and final location of the migrating bond. The initial bond location is designated as position 1, and the final location as position i, where the number of atoms is simply counted along the connection from 1 to i. Such a migration is called a [1,i] rearrangement or [1,i] shift. If the migrating group also reattaches itself at a different site from the one to which it had originally been attached, then both shifts are indicated, as in [i,j] shift. Reaction (1) is an example of a [3,3] rearrangement, because the initial carbon-oxygen (C-O) bond that breaks designates the 1 position for each component, and both components then rearrange and form a new C-C bond by reattaching at position 3 for each component. Reaction (2) is an example of a [1,3] rearrangement with substitution, commonly called an S_N2' reaction.
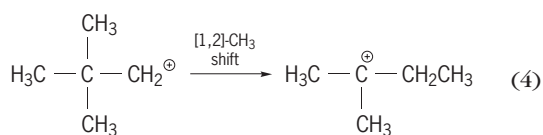
In addition, classification can be based on how many electrons move with the migrating group. Of the two electrons in the initial bond that breaks, the migrating group may bring with it both electrons (nucleophilic or anionotropic), one electron (radical), or no electrons (electrophilic or cationotropic). If the rearrangement is a concerted reaction in which there is a cyclic delocalized transition state that results in shifts of pi bonds as well as sigma bonds, the reaction is called a sigmatropic rearrangement [for example, reaction (1)]. *See* DELOCALIZATION; PERICYCLIC REACTION.

Carbocation rearrangements. The great majority of rearrangement reactions occur when a molecule develops a severely electron-deficient site, such as a carbocation. Shift of a nearby atom or group, with its pair of electrons, can serve to satisfy the electron deficiency at the original site, although it typically leaves behind another site with electron deficiency. As long as the final site can bear the electron deficiency better than the original site, the

rearrangement will be favorable. The stability of carbocations relies on electron donation from attached substituents and nearby groups. For simple alkyl groups and other electron-donating substituents, carbocation stability increases with greater numbers of substituents. Thus a simple hydrogen shift ($[1,2]$-H) can convert a less stable carbocation into a more stable one [reaction (3)].

$$H_3C - \overset{\overset{H}{|}}{\underset{\underset{CH_3}{|}}{C}} - CH_2^{\oplus} \xrightarrow{\text{[1,2]-H shift}} H_3C - \overset{\oplus}{\underset{\underset{CH_3}{|}}{C}} - CH_3 \quad (3)$$

In such a reaction, the hydrogen migrates with a pair of electrons, so the reaction is sometimes called a hydride shift (H:$^-$ is the hydride anion). Hydrogen is the atom most commonly observed to rearrange. If an adjacent hydrogen is not available for a [1,2] shift, alkyl groups [for example, the methyl group (CH_3)] may also migrate [reaction (4)].

$$H_3C - \overset{\overset{CH_3}{|}}{\underset{\underset{CH_3}{|}}{C}} - CH_2^{\oplus} \xrightarrow[\text{shift}]{\text{[1,2]-CH}_3} H_3C - \overset{\oplus}{\underset{\underset{CH_3}{|}}{C}} - CH_2CH_3 \quad (4)$$

Rearrangements in carbocations can create subtle changes or major changes in the carbon skeleton of a molecule. The phenomenon of carbon skeletal rearrangements was first recognized by pioneers in the area of terpene chemistry, for example, in the conversion of isoborneol (1) to camphene (2) [reaction (5)].

$$(5)$$

(1)   **(2)**
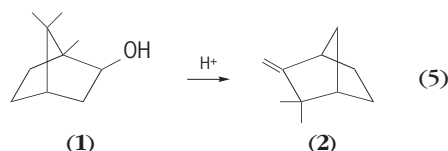
See TERPENE.

The acid-catalyzed dehydration of alcohols proceeds through carbocation intermediates, and carbon skeletal rearrangements occurring during the dehydration are known as Wagner-Meerwein rearrangements. The specific rearrangement in the conversion of isoborneol to camphene is simply a [1,2]-carbon shift, well camouflaged. Without the methyl substituents as markers, detection of the rearrangement can be even more subtle, as in the conversion of exo-2-norbornyl derivatives, which are chiral, into their enantiomeric form. Enantiomers are mirror-image isomers, and the overall reaction results in racemization and loss of optical activity. *See* OPTICAL ACTIVITY; RACEMIZATION; REACTIVE INTERMEDIATES.

Migratory aptitude. The pinacol rearrangement [reaction (6) where Ph represents a phenyl group], represents a case in which there are two different groups that may migrate; the major product indicates that the phenyl group (C_6H_5) prefers to shift rather than the methyl group. The ability of a group to shift with a pair of electrons is called migratory aptitude, and it is determined by the ability of the migrating group to

$$(6)$$
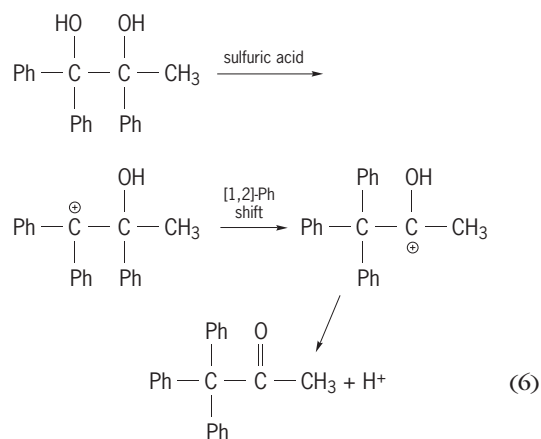
stabilize the electron deficiency. In general, aromatic groups have a greater migratory aptitude than alkyl groups.

Other rearrangements to electron-deficient carbon. The benzilic acid rearrangement [reaction (7)] rep-

$$(7)$$
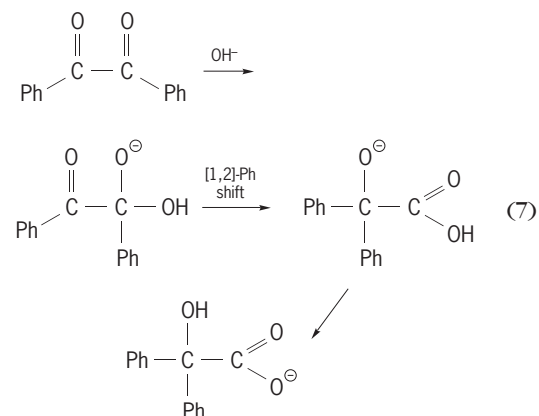
resents the migration of a phenyl group to a carbonyl carbon (C=O).

Rearrangements to electron-deficient nitrogen or oxygen. A number of related rearrangements occur in which an electron deficiency develops at a heteroatom (any atom other than hydrogen or carbon) and a carbon group shifts in response. The Beckmann rearrangement [reaction (8)] converts an

$$(8)$$
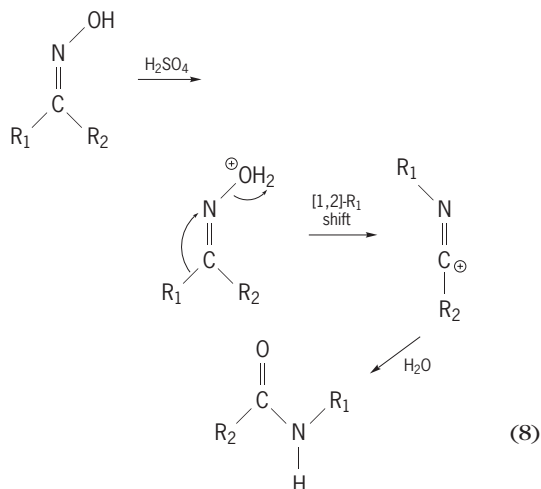
oxime (which is readily prepared from a ketone) to an amide. Of the two possible groups that may migrate to nitrogen (N), only the group anti (opposite) to the group that leaves from nitrogen migrates. This is taken as evidence that the migration is concerted (simultaneous) with the departure of the leaving group.

The Beckmann rearrangement, used in the synthesis of caprolactam (the monomer in the manufacture of Nylon 6), involves a ring expansion reaction. Rearrangement to electron-deficient oxygen occurs in the Baeyer-Villiger rearrangement, which converts a ketone to an ester. A related reaction involving peroxides is an important industrial reaction for the production of phenol and acetone.

Free-radical rearrangements. Free radicals undergo rearrangements much less readily than carbocations, an effect attributed to the molecular orbital requirements for migration. The transition state for a typical [1,2] migration involves bonding of three atoms simultaneously. One delocalized molecular orbital can accommodate two electrons, but any additional electrons, as in a radical migration, would have to be accommodated in higher-energy orbitals. Radical rearrangements are observed when the odd electron can be stabilized, for example, in migration of aryl groups where the odd electron can be delocalized into the pi system [reaction (9)].

(9)
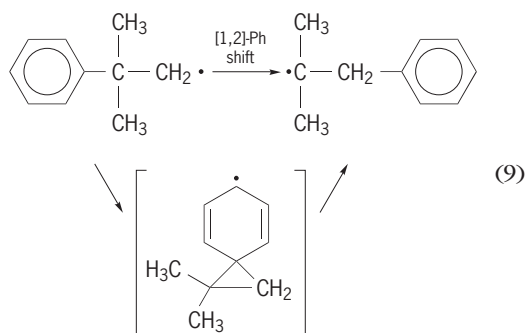
See FREE RADICAL; MOLECULAR ORBITAL THEORY.

Sigmatropic rearrangements. Concerted rearrangements that fall under the general category of pericyclic reactions are called sigmatropic rearrangements and are covered by the Woodward-Hoffmann rules. Such reactions have a cyclic delocalized transition state, and the rearrangement involves shifts of both sigma and pi bonds. reaction (1) is an example of a [3,3] sigmatropic rearrangement, specifically a Cope rearrangement. A [1,5] sigmatropic rearrangement of hydrogen [reaction (10)] is an example of a degen-

(10)
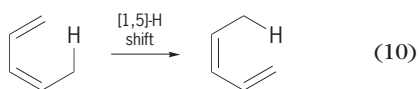
erate rearrangement, that is, the product is identical to the reactant. Detection of degenerate rearrangements requires subtle techniques, of which the most common and useful is nuclear magnetic resonance (NMR) spectroscopy, which can distinguish the specific environment of all of the hydrogen or carbon atoms in a molecule. Nuclear magnetic resonance

spectra indicate that all of the carbon atoms are equivalent in bullvalene (3) [reaction (11)]. This is

etc. (11)
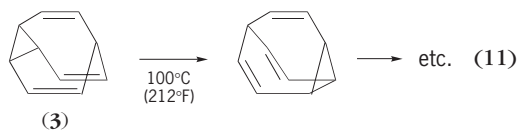
attributed to multiple Cope rearrangements, which allow every carbon atom to take its turn at any of the various sites in the molecule; in all, there are over 1 million possible forms of rearranging the 10 carbon atoms in bullvalene, and the molecule appears to sample all of them. *See* NUCLEAR MAGNETIC RESONANCE (NMR).

Application of the Woodward-Hoffmann rules to sigmatropic rearrangements allows the prediction of what sizes of rearrangements are favorable (or allowed). Cyclic delocalized transition states containing $(4n + 2)$ electrons are most favorable, since they are predicted to proceed with the optimum stereochemistry; that is, all bonding can occur on the same side of the molecule (suprafacial). Thus the Cope rearrangement, a [3,3] rearrangement, involves a total of six electrons and is favorable. *See* WOODWARD-HOFFMANN RULE.

The knowledge of the favored stereochemistry also allows specific predictions regarding the correlation of reactant and product stereochemistry. A specific example of a [1,5]-H shift [reaction (12)] illus-

(12)
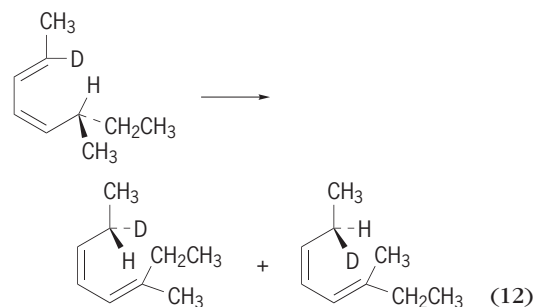
trates that the stereochemistry at the saturated centers that are the origin and terminus of the migration can be correlated with the stereochemistry of the original and final double bonds. The correlation indicates that the migration of H occurred specifically on the same side of the molecule, that is, suprafacially, part of the time along the top face and part of the time along the bottom face. *See* CHEMICAL BONDING; DEUTERIUM; ORGANIC SYNTHESIS; STEREOCHEMISTRY.   Carl C. Wamser

Bibliography. P. de Mayo, *Molecular Rearrangements*, 2 vols., 1963; P. de Mayo, *Rearrangements in Ground and Excited States*, 3 vols., 1980; J. March, *Advanced Organic Chemistry*, 5th ed., 2001.

Reciprocating aircraft engine

A fuel-burning internal combustion piston engine specially designed and built for minimum fuel consumption and light weight in proportion to developed shaft power. The rotating output shaft of the

Fig. 1. Continental A-40 engine. (*Teledyne Continental Motors*)

engine may be connected to a propeller, ducted fan, or helicopter rotor that, in turn, accelerates the surrounding air, imparting an equal and opposite thrust force for propulsion of the aircraft.

Reciprocating aircraft engines are used in about 75% of all powered aircraft flying in the United States. Most of the aircraft powered by these engines belong to the general aviation segment of the domestic aviation fleet, which includes all aircraft except those flown by scheduled airlines and the military. There are over 200,000 active aircraft in the United States general aviation fleet. The reciprocating aircraft engine is used to power single-engine and multiengine airplanes, helicopters, and airships. It is the principal
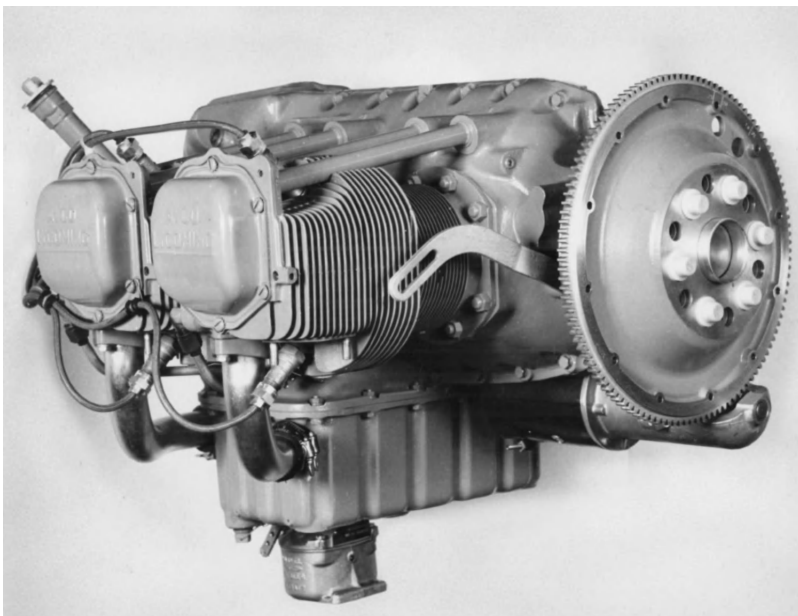
Fig. 2. Lycoming O-235 engine. (*AVCO Lycoming*)

engine used in aircraft for air taxi, pilot training, business, personal, and sport flying as well as aerial application of seed, fertilizer, herbicides, and pesticides for farming. *See* AGRICULTURAL AIRCRAFT; GENERAL AVIATION.

Predominantly, reciprocating aircraft engines operate on a four-stroke cycle, where each piston travels from one end of its stroke to the other four times in two crankshaft revolutions to complete one cycle. The cycle is composed of four distinguishable events called intake, compression, expansion (or power), and exhaust, with ignition taking place late in the compression stroke and combustion of the fuel-air charge occurring early in the expansion stroke. These spark-ignition engines burn specially formulated aviation gasolines and produce shaft power by the force of combustion gas pressure on pistons acting on connecting rods turning a crankshaft. Major parts are the crankcase, crankshaft, connecting rods, pistons, cylinders with intake and exhaust valves, camshafts, and auxiliary operating systems such as ignition, fuel injection or carburetion, and fuel and oil pumps. *See* INTERNAL COMBUSTION ENGINE.

Four-stroke-cycle compression-ignition (diesel) reciprocating engines are currently certified in a limited segment of the general aviation market. The two-stroke-cycle spark-ignition engine has found widespread use in powered ultralight airplanes where horsepower requirements are generally less than 50 hp (37.5 kW). *See* ULTRALIGHT AIRCRAFT.

The reciprocating engine powered all aircraft for the first 40 years of heavier-than-air flight, including all military, commercial, and general aviation types. The advent of the turbojet engine near the end of World War II started a rapid conversion to turbine power for military aircraft. This conversion was accelerated by the Cold War and the Korean War. The emphasis on government financing of turbine engine development resulted in a virtual halt in further development of reciprocating engines after the end of World War II. The major swing to turbine power by commercial airlines began in 1958, and the design of smaller turbojet and turboprop engines started the gradual influx of business jets into general aviation. About 7% of all general aviation aircraft are turbine-powered, a proportion not expected to change drastically in the near future.

Most modern aircraft using engines with up to 450 hp (336 kW) output are powered by air-cooled, horizontally opposed, reciprocating engines. The smaller turbine engines that fall within this power range are used in helicopters (turboshaft engines), and the few turboprop engine designs below 450 hp have failed to meet the low cost and fuel economy necessary to compete with existing reciprocating engines. The gap in power between the largest reciprocating engine and the smallest popular turboprop has been filled by the rapidly dwindling supplies of post-World War II radial engines which are no longer manufactured, but for which spare parts are still available. This gap will probably disappear altogether as lower-cost, more efficient turboprop engines are developed to fill the needs of a growing general

aviation aircraft fleet. The trend in modern reciprocating engine development is toward lower engine weight and improved fuel economy rather than increased power.

Early development. Through World War I, aircraft had relatively low velocity head available for cooling. Velocity head is the term used to indicate the air pressure due to the forward velocity of the airplane in flight. A limited knowledge of air cooling of cylinders also tended to limit use of air-cooled fixed cylinders. Major combat aircraft were therefore mostly powered by water-cooled 6-cylinder in-line, 8-cylinder V, and 12-cylinder V types. There was, however, considerable use by the French of air-cooled rotary radial engines for fighter or pursuit aircraft and some use by the British of air-cooled V engines.

World War I demonstrated the great importance of aircraft as a military weapon. This resulted in the financing and development of a great variety of engine types and forms, particularly those promising higher power and improved aircraft performance.

During the years between the two world wars, with advancements in design and materials, the variety of engine types diminished to a few well-proved designs. These were the single-row 9-cylinder and two-row 14-cylinder air-cooled radials and the liquid-cooled V-12 engines for high-speed military aircraft. During the late 1930s, larger and much higher-powered two-row radials of both 14 and 18 cylinders were developed. Design work also emphasized higher outputs and high temperature cooling on liquid-cooled 12-cylinder V engines.

This period also saw the emergence of small, low-power engines which were to become the power plants of the general aviation industry. Before 1930, there were few reliable small aircraft engines. Manufacturers of small civil aircraft used surplus World War I engines like the Curtiss OX-5 liquid-cooled V-8, or modified automotive power plants. Two former automobile engine manufacturing companies turned to the light aircraft engine business in 1928 when it became apparent that the successful car manufacturers were interested in manufacturing their own engines. The first two engines were the Continental Model A-70, a 160-hp (119-kW) 7-cylinder air-cooled radial, and the Lycoming R-680, a 215-hp (160-kW) 9-cylinder radial. These two engines and their derivatives went on to power many civil and military aircraft, including the Boeing PT-13 Stearman biplane.

There was a demand for even lower-powered engines for the emerging civil aviation market. In 1931 a small 4-cylinder, horizontally opposed engine designated the A-40 (**Fig. 1**) was manufactured. It was first installed in the Taylorcraft E-2 airplane. The engine produced 37 hp (28 kW) at 2500 revolutions per minute and had a recommended major overhaul period of 500 h. The development of the horizontally opposed reciprocating aircraft engine is considered the single most important factor in the growth of the general aviation industry. The A-40 was followed in rapid succession by the A-50, A-65, A-75, and the Lycoming O-145. The O-145 engine powered the first successful helicopter, built and flown by Igor Sikorsky. These engines continued to be produced as power plants for aircraft manufactured by Aeronca, Luscombe, Piper, Taylorcraft and many others, up until 1942 when production was halted due to World War II. *See* HELICOPTER.
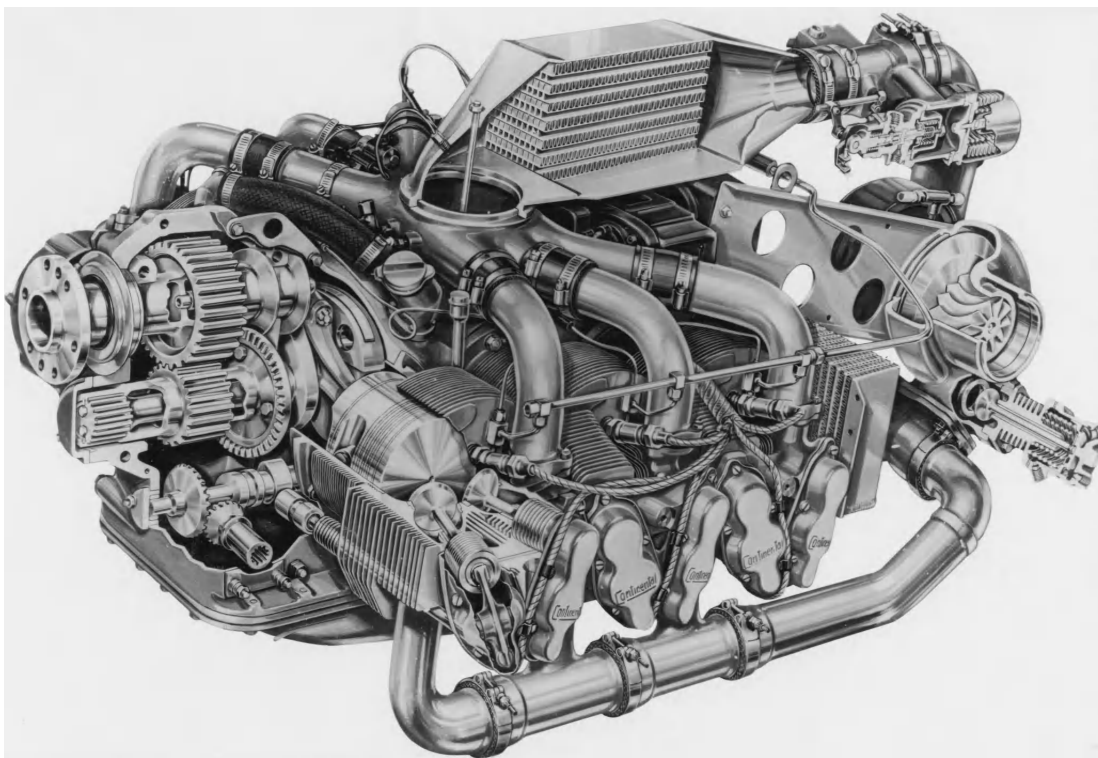

Fig. 3. Continental GTSIO-520 engine. (*Teledyne Continental Motors*)

Modern engines. Modern reciprocating engines are purely commercial products meeting the needs of the general aviation aircraft market. Modern aircraft reflect greatly improved economy, durability, and versatility in service. This is due in no small part to the advances made in small reciprocating engines since their inception in the early 1930s. Following the early 4-cylinder, horizontally opposed engines, production expanded after World War II to include 6-cylinder and 8-cylinder engines of the same format. Gradually these engines adopted mechanically driven superchargers, turbochargers with aftercooling, oil coolers to cope with the increasing power per unit displacement, fuel injection, propeller governors for constant speed props, gear reduction from the crankshaft to the propeller shaft, and sonic venturis for aircraft cabin pressurization.

The engines in production range from the 4-cylinder 115-hp (86-kW) Lycoming O-235 (**Fig. 2**) to the 6-cylinder 435-hp (324-kW) Continental GTSIO-520 (**Fig. 3**). These engines are used in over a hundred different aircraft designs, from small, two-seat

Fig. 4. Bombardier/Rotax Type 447 two-cylinder, in-line engine. (**BRP-Rotax GmbH**)

Fig. 5. Thielert Centurion 1.7 diesel aircraft engine. (**Thielert Aircraft Engines GmbH**)

Fig. 6. IOL-200, four-cylinder, liquid-cooled engine used in the *Voyager* airplane during its nonstop, unrefueled circumnavigation of the globe. (**Teledyne Continental Motors**)

training airplanes to sophisticated 18-seat commuter airliners. Some of the airplanes powered by the larger turbocharged engines fly at maximum true airspeeds of up to 260 knots (482 km/h) and altitudes of 30,000 ft (9144 m).

On the smaller end of the scale, the demand for low-cost recreational flying has opened another growth area for the reciprocating aircraft engine market. Powered ultralight airplane manufacturers are increasingly taking advantage of new stronger and lighter materials, advanced structural design techniques, and improved aerodynamics to produce a class of airplanes which is not required to be certified by the U.S. Federal Aviation Administration (FAA) and for which no pilot's license is needed. Most of the engines used to power these ultralights are 1-cylinder or 2-cylinder, two-stroke-cycle, and fan-cooled, and operate on automobile gasoline. These engines are adapted from existing uses (for example, lawnmower, snowmobile) and drive pusher propellers through a multiple-sheave-belt reduction drive. The power output of these engines is usually less than 50 hp (37.5 kW). The Bombardier/Rotax Type 447 engine (**Fig. 4**) is typical of the 2-cylinder, in-line engines.

Thielert Aircraft Engines GmbH of Germany has certified a 135-hp (99-kW), four-stroke-cycle, four-cylinder, turbocharged diesel engine (**Fig. 5**), which operates on either automotive diesel or jet fuel. As of 2005, there were over 500 general aviation airplanes flying with these Centurion 1.7 engines. The engine incorporates a full-authority digital engine control (FADEC) system which reduces the pilot's workload by controlling both the fuel flow and the rotational speed of the propeller.

Voyager engines. In December 1986, the experimental aircraft *Voyager*, piloted by D. Rutan and J. Yeager, completed the first nonstop, unrefueled circumnavigation of the globe by an airplane. *Voyager* was powered by two reciprocating aircraft engines and carried over 7000 lb (3175 kg) of fuel.

Voyager's front engine was a conventional, air-cooled, 4-cylinder O-240 model that had been in production for 18 years. Rated at 130 hp (97 kW), it

was operated in conjunction with the rear engine for the initial takeoff and for various periods during the flight to climb to higher altitudes to avoid mountains and bad weather.

The main, liquid-cooled rear engine, an experimental design designated IOL-200 (**Fig. 6**), was a 4-cylinder, horizontally opposed design producing 110 hp (82 kW) at 2750 revolutions per minute. Specially designed for low oil and fuel consumption, it had a high-turbulence combustion chamber and a compression ratio of 11.4:1. The rear engine ran for the entire 9 days of the flight, except for a 5-min period when it shut down because of a fuel-transfer problem. At the end of the flight, with only 106 lb (48 kg) of fuel remaining, the *Voyager* had averaged a fuel consumption of 22 mi/gal (9.35 km/liter).

Kenneth J. Stuckas

Bibliography. B. Gunston, *The Development of Piston Aero Engines*, 2d ed., Patrick Stevens, Ltd., 1999; IAP Staff (ed.), *Aircraft Reciprocating Engines*, 1985; P. Jackson (ed.), *Jane's All the World's Aircraft*, Jane's Information Group, annually J. L. Kerrebrock, *Aircraft Engines and Gas Turbines*, 2d ed., MIT Press, 1992; M. Kroes and T. W. Wild, *Aircraft Powerplants*, 7th ed., McGraw Hill, 1994; H. Smith, *A History of Aircraft Piston Engines*, Sunflower University Press, 1993; P. H. Wilkenson, *Aircraft Engines of the World*, 1966–1967, 1964–1965, 1944–1962, 1941.

Reciprocity principle

In the scientific sense, a theory that expresses various reciprocal relations for the behavior of some physical systems. Reciprocity applies to a physical system whose input and output can be interchanged without altering the response of the system to a given excitation. Optical, acoustical, electrical, and mechanical devices that operate equally well in either direction are reciprocal systems, whereas unidirectional devices violate reciprocity. The theory of reciprocity facilitates the evaluation of the performance of a physical system. If a system must operate equally well in two directions, there is no need to consider any nonreciprocal components when designing it.

Examples of reciprocal systems. Some systems that obey the reciprocity principle are any electrical network composed of resistances, inductances, capacitances, and ideal transformers; systems of antennas, with restrictions according to Eq. (2); mechanical gear systems; and light sources, lenses, and reflectors.

Devices that violate the theory of reciprocity are transistors, vacuum tubes, gyrators, and gyroscopic couplers. Any system that contains the above devices as components must also violate the reciprocity theory. The gyrator differs from the transistor and vacuum tube in that it is linear and passive, as opposed to the active and nonlinear character of the other two devices. *See* FARADAY EFFECT.

Rayleigh's theorem of reciprocity. Reciprocity is concisely expressed by a theorem originally proposed by Lord Rayleigh for acoustic systems and later generalized by J. R. Carson to include electromagnetic systems. Both mathematical expressions of the theory of reciprocity are closely related to the mathematical theorem known as Green's theorem. The acoustical reciprocity theorem of Lord Rayleigh is as follows: In an acoustic system consisting of a fluid medium having boundary surfaces $s_1, s_2, \ldots, s_k$ and subject to no impressed body forces, surface integral (1) holds. Here p_1 and p_2 are the pressure

$$\int_s (p_1 v_{2n} - p_2 v_{1n})\, ds = 0 \qquad (1)$$

fields produced respectively by the components of the fluid velocities v_{1n} and v_{2n} normal to the boundary surfaces $s_1, s_2, \ldots, s_k$. The integral is evaluated over all boundary surfaces. *See* GREEN'S THEOREM.

For a region containing only one simple source H. L. F. Helmholtz has shown that the theorem can be expressed as follows: A simple source at A produces the same sound pressure at B as would have been produced at A had the source been located at B. In other words, the response of a human ear at B due to a vibrating tuning fork at A is the same as the response of the ear at A due to the same tuning fork when located at B. The human ear, the tuning fork, and the intervening acoustical media constitute a physical system that obeys the theory of reciprocity.

Electromagnetic systems. The generalization of Lord Rayleigh's theorem to electromagnetic systems can be mathematically expressed by volume integral (2) where $\mathbf{E}_1$ and $\mathbf{H}_1$ are the electric and mag-

$$\int_v \nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1)\, dv = 0 \qquad (2)$$

netic field vectors describing a state due to one electromagnetic sound and $\mathbf{E}_2$ and $\mathbf{H}_2$ describe another state due to a second source. The above relation is valid as long as the medium is isotropic and the field vectors are finite and continuous, and vary according to a linear law (thus excluding ferromagnetic materials, electronic space charges, and ionized gas phenomena).

By means of Maxwell's equations, relation (2) can be expressed in another form when restricted to systems of conduction current only where $\mathbf{J}_1$ and $\mathbf{J}_2$ are the conduction current densities in an electromagnetic system due to the action of the external electric fields $\mathbf{E}_1$ and $\mathbf{E}_2$, respectively.

Equation (3) is readily applied to antennas and radiation.

$$\int_v (\mathbf{E}_1 \cdot \mathbf{J}_2 - \mathbf{E}_2 \cdot \mathbf{J}_1)\, dv = 0 \qquad (3)$$

diation. If, in **Fig. 1**, $\mathbf{J}_1$ is the resulting current density in antenna B due to an electric field $\mathbf{E}_1$ established by antenna A, and $\mathbf{J}_2$ is the current density in antenna A due to electric field $\mathbf{E}_2$ established by antenna B, then $\mathbf{J}_1 = \mathbf{J}_2$, provided $\mathbf{E}_1 = \mathbf{E}_2$. The two emfs need not be applied at the same instant of time. The integral in Eq. (3) over all space reduces to an integral over the two antennas since $\mathbf{J}_1$ and $\mathbf{J}_2$ are zero elsewhere. From this particular application of the

reciprocity theorem it is seen that the transmitting and receiving patterns of an antenna are the same.

Equation (3), when evaluated over an N-mesh electrical network, reduces to Eq. (4), where a and b are

$$\sum_{j=1}^{N} V_{aj}i_{bj} = \sum_{j=1}^{N} V_{bj}i_{aj} \qquad (4)$$

two different states of the network and the j subscript denotes in which of the N meshes the voltage and current are measured. For the two-mesh network in **Fig. 2**, Eq. (4) gives Eq. (5). Expressed in

$$V_{a1}i_{b1} = V_{b2}i_{a2} \qquad (5)$$

words: If an emf source of magnitude V and zero internal impedance, when applied to terminals 1–1, produces a current I at terminals 2–2, then the same current I will be measured at terminals 1–1 when the emf V is applied to terminals 2–2. This statement, that is, Eq. (5), is probably the most familiar form of the theorem of reciprocity.

Electrostatic systems. The statement of reciprocity for electrostatics is given by Eq. (6), where V_1 and

$$\int_{v} \rho_1 V_2 dv = \int_{v} \rho_2 V_1 dv \qquad (6)$$

V_2 are the electric potentials produced at some arbitrary point due to the volume charge distributions ρ_1 and ρ_2, respectively. When the integral expression in Eq. (6) is applied to the electrostatic system of two charged conductors in **Fig. 3**, it becomes Eq. (7). Here V_a is the potential on conductor a due

$$V_a q_a = V_b q_b \qquad (7)$$

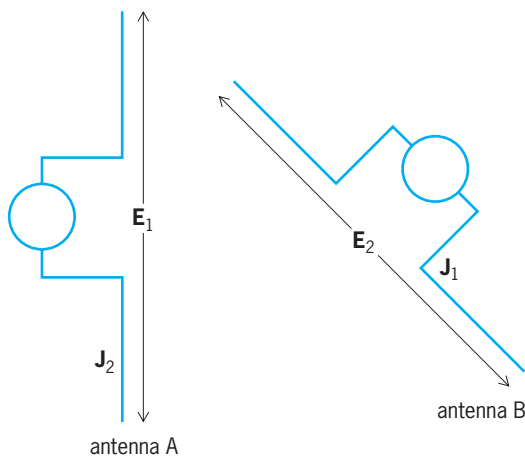to charge q_b on conductor b; the remaining quantities are similarly defined. In other words, if a charge
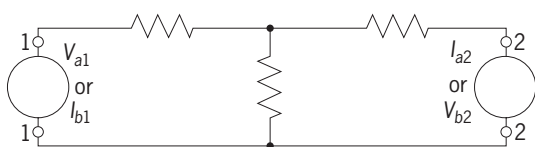

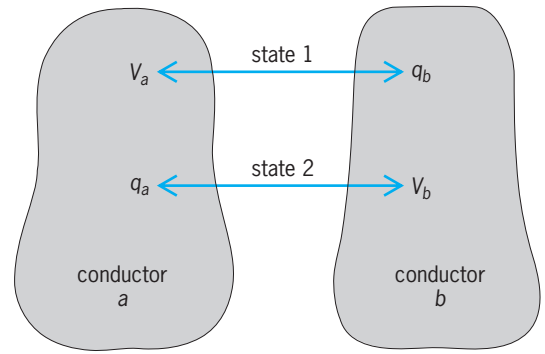Fig. 1. Antenna system.

Fig. 2. Two-mesh network.

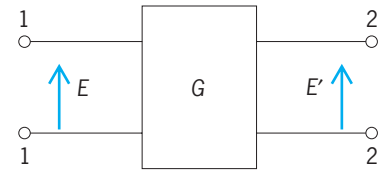**Fig. 3.** Charged conducting bodies.

Fig. 4. Four-terminal, black-box network.

q_b on conductor b raises the potential of conductor a to V, then the same charge on conductor a raises the potential of conductor b to V.

Electrical networks. A somewhat different approach to reciprocity is the so-called black box, or two-terminal pair, method illustrated in **Fig. 4**. The box might contain a mechanical, acoustical, optical, or electrical system. The applied excitation or cause is E and the response or effect is E'. The ratio of E/E' (or E'/E) is the transfer function G for the system within the black box. Using the subscript notation of G_{12} when E is impressed at terminals 1–1 and E' is measured at terminals 2–2, then G_{21} represents a response measured at 1–1 for an excitation at 2–2. Mathematically the general behavior of the box to excitations at both sets of terminals can be expressed by Eqs. (8), as long as the response bears a linear

$$E_1 = G_{11}E'_1 + G_{12}E'_2$$
$$E_2 = G_{21}E'_1 + G_{22}E'_2 \qquad (8)$$

relation to the excitation.

If, in addition to its linear characteristic, the system satisfies Eq. (9), the principle of reciprocity is

$$G_{12} = G_{21} \qquad (9)$$

obeyed, and the device will operate equally in either direction. Whenever $G_{12} \neq G_{21}$, the system violates the theory of reciprocity, with the result that the response in one direction is different from that obtained in the other direction. *See* NETWORK THEORY.

<div align="right">Hugh S. Landes</div>

Bibliography. B. Bleaney and B. I. Bleaney, *Electricity and Magnetism*, 3d ed., 1976, reprint 1989; D. E. Gray (ed.), *American Institute of Physics Handbook*, 3d ed., 1972; P. Hammond and J. K. Sykulski, *Engineering Electromagnetism: Physical Processes and Computation*, 1994; J. D. Kraus, *Antennas*, 2d ed., 1988; M. Valkenburg (ed.), *Reference Data*

for Engineers: Radio, Electronics, Computers, and Communications, 8th ed., 1992; J. Vanderlinde, *Classical Electromagnetic Theory*, 1993.

Recombination (genetics)

The formation of new genetic sequences by piecing together segments of previously existing ones. Classical genetics is based on the transmission of traits, attributable to particular genetic determinants or genes, through sexual generations. The genes are predominantly present in chromosomes, the threadlike structures that fill the cell nucleus and divide as the cells divide. The key component of the chromosomes is deoxyribonucleic acid (DNA); each chromosome has a single coiled and folded double-stranded DNA molecule running along its length, and the genes of the chromosomes occupy segments of this DNA. *See* CHROMOSOME; DEOXYRIBONUCLEIC ACID (DNA).

Recombination often follows DNA transfer in bacteria and, in higher organisms, is a regular feature of sexual reproduction. The basic feature of sexual reproduction is the cyclical alternation between the diploid state, with a double set of chromosomes (two of each kind), and the haploid state, with a single set. Animals and most flowering plants are diploid, and their germ cells (eggs and sperm, or eggs and pollen grains) are haploid. In most fungi, the organism is normally haploid, but sexual fertilization creates a transient diploid cell that immediately divides to form a tetrad of four haploid spores. *See* REPRODUCTION (ANIMAL); REPRODUCTION (PLANT).

In all sexually reproducing organisms, the transition from the diploid to the haploid state is brought about by the process of meiosis—two rounds of cell division accompanied by only one division of the chromosomes. The two chromosomes of each kind present in the diploid meiotic mother cell are necessarily segregated into different haploid meiotic products. *See* MEIOSIS.

Free reassortment and linkage. When the diploid cell undergoing meiosis is heterozygous, that is, carries two alternative forms (alleles) of the same gene inherited from different parents, the meiotic segregation of chromosome pairs will ensure that half of the meiotic products (germ cells) will carry one alternative and half the other. When the diploid is doubly heterozygous (*A/a B/b*), the two pairs of alleles segregate independently of one another at meiosis, provided the two gene differences are associated with different chromosome pairs. Consequently, four possible products—*AB*, *Ab*, *aB*, and *ab*—will occur with statistically equal frequency. *See* ALLELE; MENDELISM.

If, in a double heterozygote *Aa Cc*, the two genes are in the same chromosome and they are not very far apart, they show linkage, with the parental combinations (*AC* and *ac*, for example) tending to appear in the meiotic products more frequently than the recombinant types (*Ac* and *aC*). The fact that recombinants types *Ac* and *aC* occur at all is due to reciprocal exchanges between chromosomes (cross-
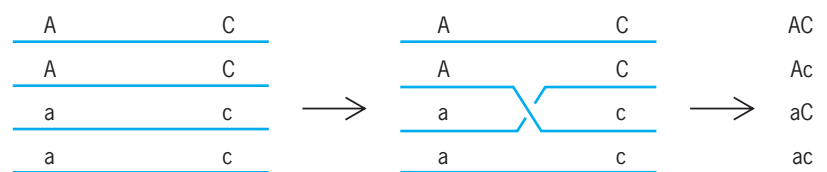

Fig. 1. Crossing-over in meiotic tetrads. Only two of the four products usually show crossing-over of linked markers.

ing over) that take place in the first meiotic division. The reciprocal nature of these events is shown most clearly in fungi such as yeast and the filamentous fungus *Neurospora crassa*, where the four products of a single meiosis can be recovered together as a tetrad of haploid cells (or, in the case of *Neurospora*, of spore pairs). In such organisms, one recombinant type (*Ac*) is generally accompanied in the same tetrad by the reciprocal product (*aC*). *See* CROSSING-OVER (GENETICS); LINKAGE (GENETICS).

Only two of four tetrad members are generally recombinant. Tetrads of constitution *AC, Ac, aC, ac* are far commoner than those containing four recombinant products, the inference being that crossing-over between chromosome pairs occurs after the chromosomes have replicated and that each crossover involves only half of each divided chromosome (chromatid) [**Fig. 1**]. If more than one crossover occurs within the linkage group, it is a matter of chance which chromatid of each chromosome is involved in each chiasma.

Chromosome maps. All the segregating genetic markers found in a given organism can be assigned to any of a number of linkage groups. Markers in different groups reassort independently, whereas markers in the same group are linked. The number of linkage groups is the same as the haploid number of chromosomes, and through correlations of changed linkage relationships with specific chromosome aberrations, each linkage group can be assigned to a specific chromosome (**Fig. 2**).

The measure of linkage is recombination frequency [$(Ac + aC)/(AC + Ac + aC + AC)$], which is often expressed as a percentage, and it must be significantly less than 50% to be distinguishable from random reassortment. The map distance, in map units or centimorgans, between two markers is 100 times the mean number of crossovers per chromatid, or 50 times the mean number per meiotic tetrad. If there is never more than one crossover between the markers, recombination frequency equals map distance. However, if there is a significant number of double or multiple crossovers, recombination frequency is an underestimate of map distance because each additional crossover within a marked interval has as much chance of canceling recombination as of causing it. Above a certain distance, recombination frequency approaches a limiting value of 50%. In practice, the linkage map of each chromosome is built up from the summing of a series of intervals, each too short for double crossovers to have a significant effect. A corollary of the theory of chromosome mapping is that the total map length of a
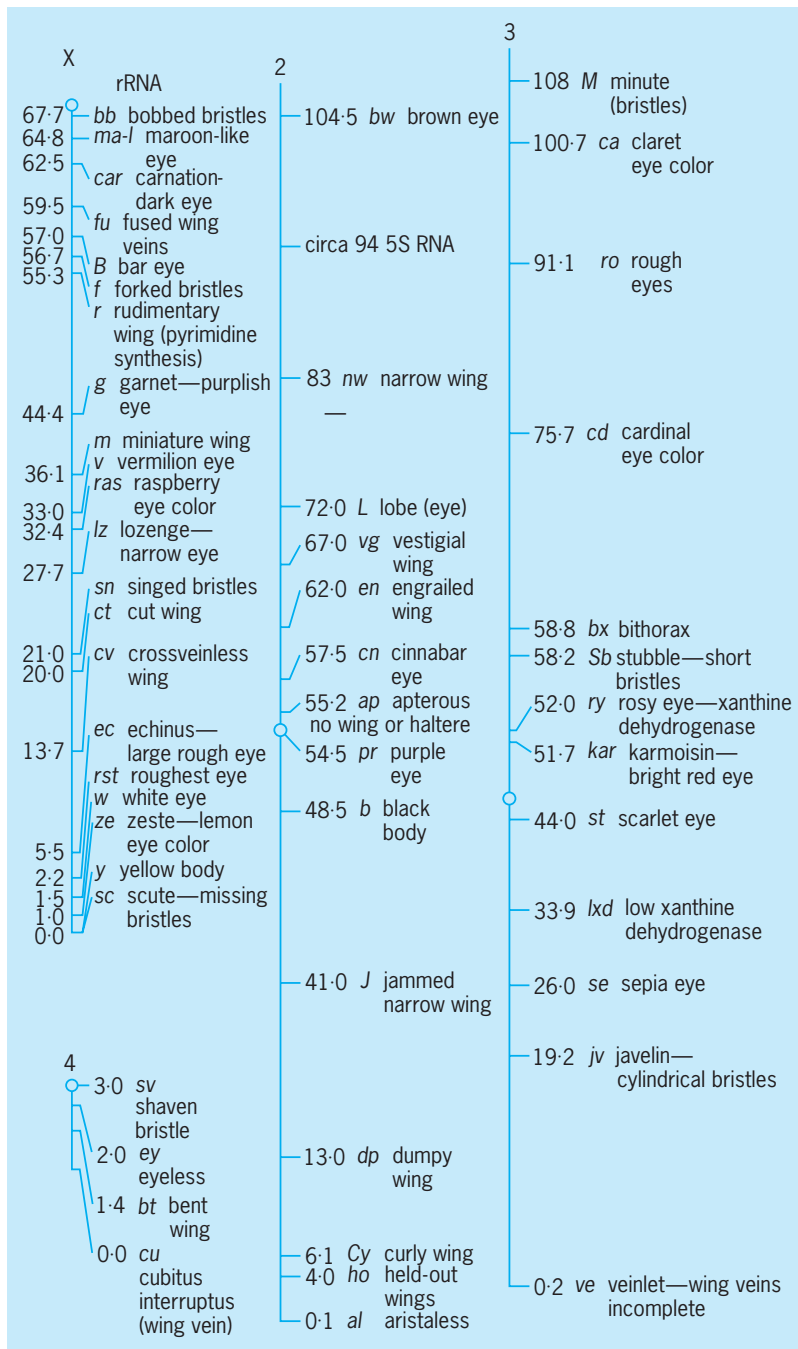
X

rRNA
- 67.7 — bb bobbed bristles
- 64.8 — ma-l maroon-like eye
- 62.5 — car carnation-dark eye
- 59.5 — fu fused wing veins
- 57.0
- 56.7 — B bar eye
- 55.3 — f forked bristles
- r rudimentary wing (pyrimidine synthesis)
- g garnet—purplish eye
- 44.4
- m miniature wing
- v vermilion eye
- 36.1 — ras raspberry eye color
- 33.0
- 32.4 — lz lozenge—narrow eye
- 27.7
- sn singed bristles
- ct cut wing
- 21.0 — cv crossveinless wing
- 20.0
- ec echinus—large rough eye
- 13.7 — rst roughest eye
- w white eye
- ze zeste—lemon eye color
- 5.5 — y yellow body
- 2.2 — sc scute—missing bristles
- 1.5
- 1.0
- 0.0

2
- 104.5 — bw brown eye
- circa 94 5S RNA
- 83 — nw narrow wing
- —
- 72.0 — L lobe (eye)
- 67.0 — vg vestigial wing
- 62.0 — en engrailed wing
- 57.5 — cn cinnabar eye
- 55.2 — ap apterous no wing or haltere
- 54.5 — pr purple eye
- 48.5 — b black body
- 41.0 — J jammed narrow wing
- 13.0 — dp dumpy wing
- 6.1 — Cy curly wing
- 4.0 — ho held-out wings
- 0.1 — al aristaless

3
- 108 — M minute (bristles)
- 100.7 — ca claret eye color
- 91.1 — ro rough eyes
- 75.7 — cd cardinal eye color
- 58.8 — bx bithorax
- 58.2 — Sb stubble—short bristles
- 52.0 — ry rosy eye—xanthine dehydrogenase
- 51.7 — kar karmoisin—bright red eye
- 44.0 — st scarlet eye
- 33.9 — lxd low xanthine dehydrogenase
- 26.0 — se sepia eye
- 19.2 — jv javelin—cylindrical bristles
- 0.2 — ve veinlet—wing veins incomplete

4
- 3.0 — sv shaven bristle
- 2.0 — ey eyeless
- 1.4 — bt bent wing
- 0.0 — cu cubitus interruptus (wing vein)

Fig. 2. Linkage maps for chromosomes X, 2, and 3 of *Drosophila melanogaster* at the metaphase state of mitosis. Numbers indicate the distance from the upper end of the chromosome of important genes (in map units). (*From J. R. S. Fincham, Genetics, Jones and Bartlett, 1983*)

linkage group should be 50 times the mean chiasma frequency of the corresponding chromosome pair at meiosis. This has been confirmed in organisms, such as corn (*Zea mays*), where both relatively complete linkage maps and accurate chiasma counts are available. *See* GENETIC MAPPING.

The average amount of DNA corresponding to 1% recombination (one map unit) varies enormously between species, from less than 10,000 DNA base pairs in the yeast *Saccharomyces cerevisiae* to about a million base pairs in humans and even higher values in some amphibia and flowering plants. Furthermore, the value varies from one part of the genome to another within a species, since not all chromosome regions have the same probability of crossing-over.

A standard method of determining the sequence of three linked markers is the three-point test cross. This measures the frequencies of the eight possible kinds of meiotic product formed by a diploid segregating with respect to three linked markers, say *ACD/acd*. If the markers are in the order written, recombinant meiotic products of constitution *AcD* or *aCd* will be relatively infrequent compared with *Acd*, *aCD*, *ACd*, and *acD*, since, unlike the latter classes, which arise from single crossovers, they require relatively rare double crossovers for their formation. The method would be seriously undermined if crossovers tended to occur in clusters such that doubles and multiples were as probable as singles. In fact, the occurrence of a crossover in one chromosome interval tends to reduce the likelihood of a crossover in an adjacent interval, an effect called crossover or chiasma interference. A decreased probability of occurrence of crossovers beyond the first has also been inferred from chiasma counts under the microscope. Whereas the chance of at least one chiasma is nearly 100% (very few chromosome pairs are left unjoined by chiasmata at the first meiotic division), shorter chromosomes commonly have no more than one and even the longest may never have more than three or four.

Conversion. Recombination was once thought to occur only between genes, never within them. Indeed, the supposed indivisibility of the gene was regarded as one of its defining features, the other being that it was a single unit of function. The functional criterion was failure of alleles to complement one another. If two defective mutants of independent origin failed to compensate for each other's defects to determine a more normal phenotype when present together in a heterozygote, then neither was thought capable of supplying the normal function that was defective in the other; hence, both must be defective in the same gene, defined as a unit of function. The two definitions agreed to the extent that mutations placed in the same gene by the complementation criterion were always at least very closely linked. However, examination of very large progenies shows that, in all organisms studied, nearly all functionally allelic mutations of independent origin can recombine with each other to give nonmutant products, generally at frequencies ranging from a few percent (the exceptionally high frequency found in *Saccharomyces*) down to 0.001% or less. The indivisible unit of recombination is no larger than a single DNA base-pair.

Mutations within a gene can be mapped in a linear order either by using the criterion of recombination frequency or, more accurately, by deletion mapping, which uses deletions of extended segments of the gene. Overlapping deletion mutations cannot recombine to give normal recombinants since neither has all of the sequence in which the other is defective; for the same reason, a deletion mutation gives no recombinants with any single-base-pair mutation
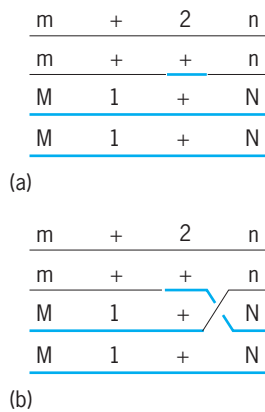
m	+	2	n
m	+	+	n
M	1	+	N
M	1	+	N

(a)

m	+	2	n
m	+	+	n
M	1	+	N
M	1	+	N

(b)

Fig. 3. Typical patterns of intragenic recombination in meiotic tetrads: 1 and 2 are mutant sites within the gene (+ is the wild-type site); *M,m* and *N,n* are flanking markers. (*a*) Conversion of site 1 to wild type, to give a wild-type recombinant gene, is accompanied by adjacent crossing-over in this tetrad but (*b*) not in the other tetrad. Contributions of the two recombining chromosomes are distinguished by black and colored lines.

falling within the region defined by the deletion. Given a series of overlapping deletions, any point mutation can be positioned within one or any other of the segments defined by the deletions and their overlaps.

Tetrad analysis, especially in *S. cerevisiae*, shows that recombination within genes is most frequently nonreciprocal. In a cross between two allelic mutants, both nonmutant and double-mutant products are formed, but seldom in the same tetrad (**Fig. 3**). This phenomenon is termed gene conversion, and must represent the nonreciprocal transfer of DNA sequence between chromatids of paired chromosomes. Gene conversion is also found in crosses involving only a single mutational difference. In addition to the expected 2:2 tetrad segregation pattern, up to several percent 3:1 and 1:3 ratios can be found in yeast; frequency in other experimental organisms is generally very much lower. Both in *Saccharomyces* and other fungi, genes often show a gradient of conversion frequency from one end of a gene to the other, an effect called conversion polarity.

Conversion sometimes affects only half a chromatid, equivalent to a single DNA strand, rather than both strands of the double helix. In this case, formation of a completely converted gene is delayed until after the first replication of the DNA following meiosis. When donor and recipient genes differ in more than one mutational marker, the two markers are often transferred together (coconverted). The frequency of coconversion, compared with single-marker conversion, diminishes as the distance between the markers increases; the relationship between frequency and distance indicates that conversion involves the transfer of segments of DNA commonly of the order of several hundred to a few thousand base-pairs in length. Coconversion does not result in any marker recombination; in order to recombine two markers within a gene, a conversion segment has to include one but not the other. *See* MUTATION.

Unifying theory. Recombination between and within genes may seem to be two distinct processes involving reciprocal crossing-over and nonreciprocal conversion, respectively. It is, however, attractive to postulate that both follow from the same kind of initiating event. Conversion within a gene is often (with a frequency of 40–50%) accompanied by a closely adjacent crossover, detectable through the reciprocal recombination of markers flanking the gene undergoing conversion (Fig. 3). Both conversions and crossovers are probably formed from precursor structures involving local unilateral transfer of DNA segments between chromatids. Such structures, which are called conversion tracts, will lead to observable gene conversion only in the rare event that the transferred segment happens to include a genetic marker. If the structure develops into a crossover (which is presumed to happen in yeast with a probability of 40–50%), it will always lead to reciprocal crossing-over between flanking markers, however distant they may be from the initiating event. In the case of recombination between extremely close markers (particularly those within the same gene), the initiating event is necessarily very close to the markers being recombined, which will thus have a high probability of being included in the conversion tract.

Molecular mechanisms. There are two prevailing hypotheses as to the nature of the event that initiates recombination. In the first model, a chromatid involved in meiotic pairing is cut in one strand of its DNA, probably at one of a large number of special sites distributed along the chromosome. A stretch of single-stranded DNA then unwinds from one side of the cut and invades the DNA of the other chromatid, displacing the corresponding single-strand segment from its duplex. In the simplest sequence of events, the invading single strand becomes sealed into the recipient duplex, creating a mismatch between the complementary strands at any point at which the recipient and donor duplexes differ. Such a mismatch can be left unrepaired, in which case the result will be half-chromatid conversion and postmeiotic segregation; or (more frequently, at least in yeast) it is repaired by excision and replacement of the mismatch in one strand either to restore the original arrangement or to convert the recipient whole chromatid to the type of the donor. To explain the association between conversion and crossing-over, it is supposed that the single-strand bridge can lead to a reciprocal single-strand exchange (a Holliday junction) and then to a reciprocal double-strand exchange, that is, a regular crossover. The chance of a single-strand transfer leading to a crossover is variable but seldom exceeds 50% (**Fig. 4***a*).

A radically different hypothesis proposes that the initiating event is a double-stranded break, which may be subsequently enlarged to a gap, in one chromatid. The gap is healed by copying of both strands of the undamaged partner chromatid. The repair process involves the formation of Holliday junctions on both sides of the gap and, depending on the way these are resolved, crossing-over may take place. The
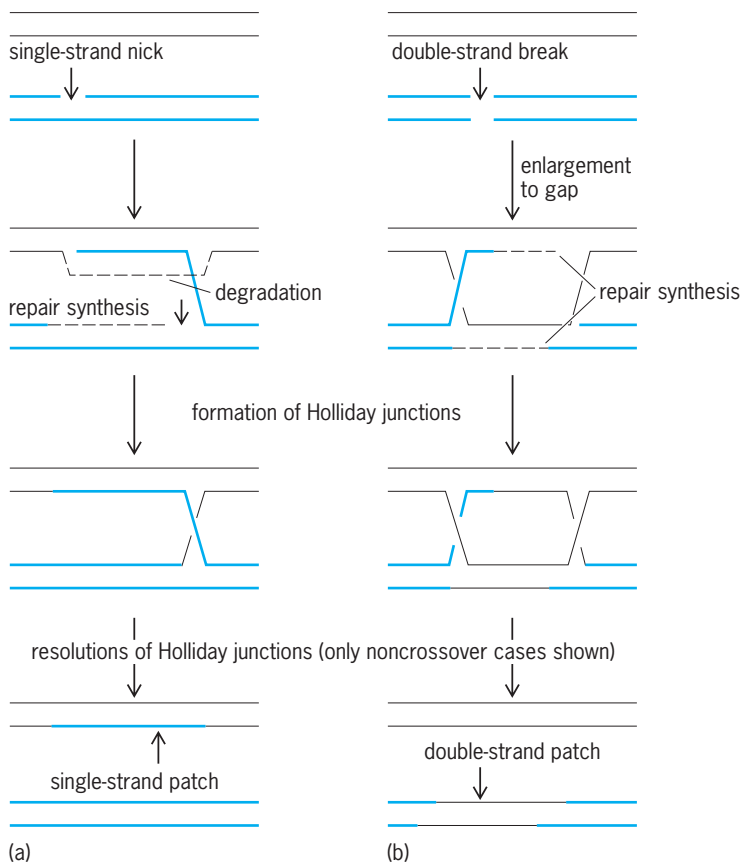
Fig. 4. Typical models of general recombination. (*a*) Single-strand assimilation model. (*b*) Double-strand gap repair model. In these models, conversion occurs without crossing-over, but other, supposedly equally probable, models of resolution of the Holliday junctions lead to crossing-over adjacent to the conversion event. Contributions of the recombining chromatids are differentiated by black and colored lines.

double-strand break-repair hypothesis accounts for conversion either with or without adjacent crossing-over, but it does not readily explain the occurrence of conversion events that affect only half a chromatid, leading to postmeiotic segregation. However, this too can be explained if the regions of heteroduplex that would be formed on each side of the gap (from annealing of single-strand ends into the undamaged duplex) are sufficiently extensive (Fig. 4*b*).

Both hypotheses provide for a scattering of conversion segments in all chromosome pairs, some of which are associated with crossovers. Both explain polarity in gene conversion as due to the initial DNA cuts, whether single- or double-stranded, occurring at preferred sites in the DNA; the probability of heteroduplex formation or gapping, and thus the probability of a marker undergoing conversion, should decrease with increasing distance from such a site. The double-strand break-repair model appears more probable, inasmuch as a site at the high-conversion end of at least one yeast gene appears to undergo double-strand breakage at meiosis. However, elements of both proposed mechanisms may have to be included in a more refined model.

Most speculation on recombination mechanisms has focused on recombining chromosomes as DNA molecules, which is an oversimplification. Little is yet known about how the higher-order DNA protein

structure of chromosomes affects their recombination. The role of the synaptonemal complex, across which the exchanged strands of DNA may need to pass, is also obscure. *See* MOLECULAR BIOLOGY.

Mitotic recombination. Crossing-over between homologous chromosome pairs can also occur during the prophase of mitotic nuclear division. The frequency is very much lower than in meiosis, presumably because the mitotic cell does not form the synaptic apparatus for efficient pairing of homologs. Mitotic crossing-over has been studied in the fruit fly *Drosophila melanogaster*, in the filamentous fungus *Aspergillus nidulans*, and in *Saccharomyces* yeast. In these species it is detected through the formation of homozygous clones of cells in an initially heterozygous diploid. There is a 50% chance of homozygosity in daughter cells whenever a crossover occurs between chromatids in the interval between the marker and the centromere, the chromosomal site of attachment to the mitotic spindle (**Fig. 5**). The frequency of mitotic crossing-over is greatly increased by radiation. *See* MITOSIS.

Integration of bacterial DNA fragments. Bacteria have no sexual reproduction in the true sense, but many or most of them are capable of transferring fragments of DNA from cell to cell by one of three mechanisms. (1) Fragments of the bacterial genome can become joined to plasmid DNA and transferred by cell conjugation by the same mechanism that secures the transfer of the DNA of transmissible plasmids. (2) Genomic fragments can be carried from cell to cell in the infective coats of bacterial viruses (phages), a process called transduction. (3) Many bacteria have the capacity to assimilate fragments of DNA from solution and so may acquire genes from disrupted cells.

Fragments of DNA acquired by any of these methods can be integrated into the DNA of the genome in place of homologous sequences previously present. If the incoming DNA has no homology with that of the recipient cell, it usually cannot be integrated and is lost for lack of ability to replicate autonomously. Homologous integration in bacteria is similar, in its nonreciprocal nature and perhaps also in its mechanism, to gene conversion in eukaryotic organisms. *See* BACTERIAL GENETICS; BACTERIOPHAGE; TRANSDUCTION (BACTERIA).

Site-specific recombination. Bacteriophages, plasmids, bacteria, and unicellular eukaryotes provide many examples of differentiation through controlled and site-specific recombination of DNA segments. In vertebrates, a controlled series of deletions leads to the generation of the great diversity of gene sequences encoding the antibodies and T-cell receptors necessary for immune defense against pathogens. All these processes depend upon interaction and recombination between specific DNA sequences, generally but not always with some sequence similarity, catalyzed by site-specific recombinase enzymes. The molecular mechanisms may have some similarities with those responsible for general meiotic recombination, except that the latter does not depend on any specific sequence, only on similarity (homology) of the sequence recombined.
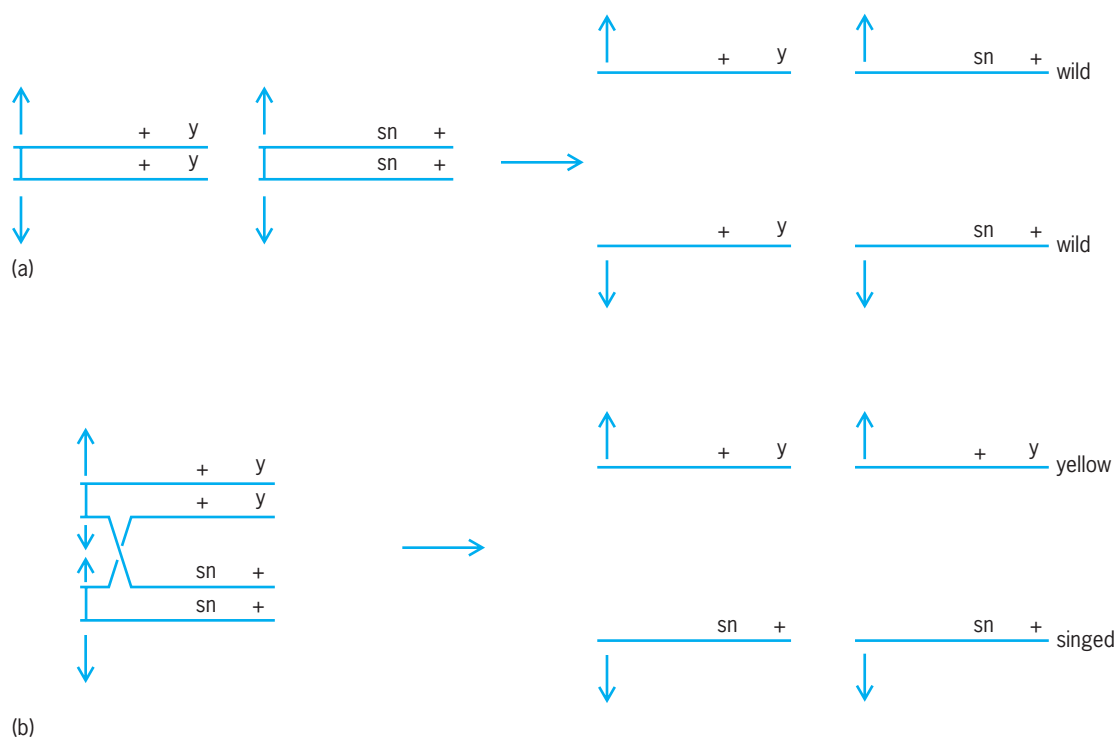
Fig. 5. Mitotic crossing-over, as demonstrated by one *Drosophila* X chromosome in the female carrying the cuticle marker *yellow* (y), and the other the linked marker *singed* (sn), affecting bristle morphology. (*a*) Normal mitosis, with heterozygosity of both markers maintained and a wild-type appearance of the fly, since both mutations are recessive. (*b*) The effect of a crossover in the position indicated during embryonic development, which gives (with 50% probability) adjacent yellow and singed patches on the surface of the adult fly.

Nonhomologous recombination. Techniques have been devised for the artificial transfer of DNA fragments from any source into cells of many different species, thus conferring new properties upon them (transformation). In bacteria and the yeast *S. cerevisiae*, integration of such DNA into the genome (on which the stability of transformation generally depends) requires substantial sequence similarity between incoming DNA and the recipient site. However, cells of other fungi, higher plants, and animals are able to integrate foreign DNA into their chromosomes with little or no sequence similarity. These organisms appear to have some unidentified system that recombines the free ends of DNA fragments into chromosomes regardless of their sequences. It may have something in common with the mechanism, equally obscure, whereby broken ends of chromosomes can heal by nonspecific mutual joining. *See* TRANSFORMATION (BACTERIA).

Recombinant DNA in genetic engineering. The science of genetics has been revolutionized by the development of techniques using isolated cells for specific cleaving and rejoining of DNA segments and the introduction of the reconstructed molecules into living cells. This artificial recombination depends on the use of site-specific endonucleases (restriction enzymes) and DNA ligase. *See* GENE; GENE ACTION; GENETIC ENGINEERING; GENETICS; RESTRICTION ENZYME.   J. R. S. Fincham

Bibliography. J. R. S. Fincham and P. Oliver, Initiation of recombination, *Nature*, 338:14–15, 1989; J. Ilan (ed.), *Translational Regulation of Gene Expression*, 2d ed., 1993; M. Karin (ed.), *Gene Expression: General and Cell-Type-Specific*, 1993; K. B. Low (ed.), *The Recombination of Genetic Material*, 1988; M. Meselson and C. M. Radding, A general model for recombination, *Proc. Nat. Acad. Sci. USA,* 72:821–828, 1982; A. Prokop, R. K. Bajpapi, and C. S. Ho, *Recombinant DNA Technology and Applications*, 1991; J. W. Szostak et al., The double-strand break repair model of recombination, *Cell,* 33:26–35, 1983.

Rectifier

A nonlinear circuit component that allows more current to flow in one direction than in the other. An ideal rectifier is one that allows current to flow in one (forward) direction unimpeded but allows no current to flow in the other (reverse) direction. Thus, ideal rectification might be thought of as a switching action, with the switch closed for current in one direction and open for current in the other direction. Rectifiers are used primarily for the conversion of alternating current (ac) to direct current (dc). *See* ELECTRONIC POWER SUPPLY.

A variety of rectifier elements are in use. The vacuum-tube rectifier can efficiently provide moderate power. Its resistance to current flow in the reverse direction is essentially infinite because the tube does not conduct when its plate is negative with respect to its cathode. In the forward direction, its resistance is small and almost constant. Gas tubes, used primarily for higher power requirements,

also have a high resistance in the reverse direction. The semiconductor rectifier has the advantage of not requiring a filament or heater supply. This type of rectifier has approximately constant forward and reverse resistances, with the forward resistance being much smaller. Mechanical rectifiers can also be used. The most common is the vibrator, but other devices are also used. *See* GAS TUBE; MECHANICAL RECTIFIER; SEMICONDUCTOR RECTIFIER.

A rectifying element can be illustrated by assuming a device having a forward resistance R_1 and a reverse resistance R_2, which is much greater than R_1. A sinusoidal alternating voltage $E_m \sin 2\pi ft$ is applied to the rectifier, where E_m is the maximum value of the applied voltage, f is the frequency of the voltage wave, and t is time. The magnitude of the current in the forward direction is $(E_m/R_1) \sin 2\pi ft$. This current flows from t equals 0 to $1/(2f)$, or for one-half the cycle of the alternating voltage wave. The forward current, averaged over one cycle, is $E_m/(\pi R_1)$. The reverse current has the magnitude $(E_m/R_2) \sin 2\pi ft$ and flows from t equals $1/(2f)$ to $1/f$, or for the other half-cycle. The reverse current, averaged over one cycle, is $E_m/(\pi R_2)$. The net forward average current is $E_m(R_2 - R_1)/(\pi R_1 R_2)$.

If reverse resistance R_2 is extremely large compared to R_1, the average current approaches $E_m/(\pi R_1)$. If the average current is subtracted from the current flowing in the rectifier, an alternating current results. This ripple current flowing through a load produces a ripple voltage which is often undesirable. Filter and regulator circuits are used to reduce it to as low a value as is required. *See* ELECTRIC FILTER; VOLTAGE REGULATOR.

Half-wave rectifier circuit. A half-wave rectifier circuit is shown in **Fig. 1**. The rectifier is a diode, which allows current to flow in the forward direction from A to B but allows practically no current to flow in the reverse direction from B to A. The ac input is applied to the primary of the transformer; secondary voltage e supplies the rectifier and load resistor R_L.
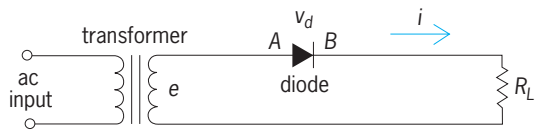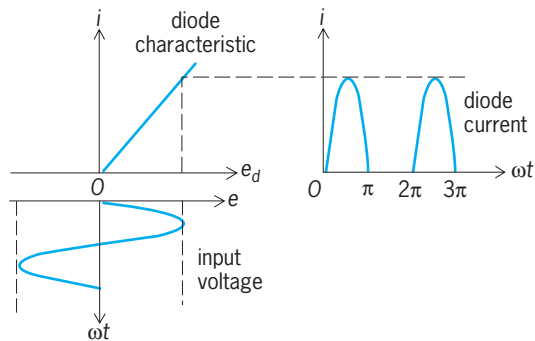

Fig. 1. Half-wave diode rectifier.

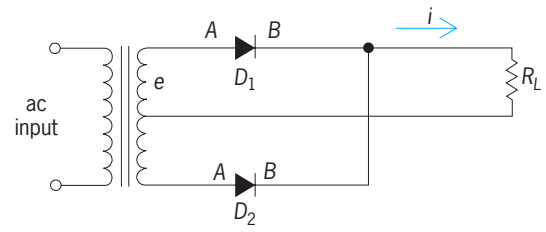**Fig. 2. Rectifying action of half-wave diode rectifier.**

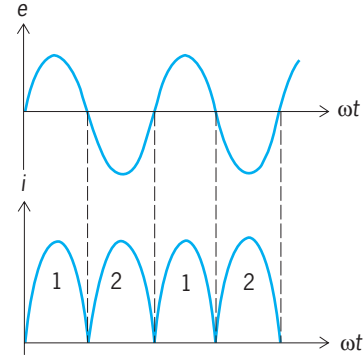**Fig. 3. Full-wave diode rectifier.**

Fig. 4. Applied voltage and output current of full-wave rectifier.

The rectifying action of the diode is shown in **Fig. 2**, in which the current i of the rectifier is plotted against the voltage e_d across the diode. The applied sinusoidal voltage from the transformer secondary is shown under the voltage axis; the resulting current i flowing through the diode is shown at the right to be half-sine loops. Averaging the value of these half-sine loops gives the direct current flowing; the ripple current is the variation of load current about the average value.

Full-wave rectifier circuit. A full-wave rectifier circuit is shown in **Fig. 3**. This circuit uses two separate diodes, D_1 and D_2. The center tap of the transformer is connected through the load resistance R_L to the B sides of both diodes. During one-half cycle of the ac input the A terminal of diode D_1 is positive with respect to the B terminal, and D_1 conducts current from A to B. This current passes through the load resistor R_L in the direction shown by i. During this time diode D_2 is not conducting because terminal A is negative with respect to B. When the ac potential goes through zero, A of diode D_1 becomes negative and the diode stops conducting. The potential on A of D_2 then becomes positive, and D_2 starts to conduct. The resulting current wave shape is shown in **Fig. 4**.

The effect of using two diodes instead of one is to produce a more continuous flow of direct current through load R_L because the first diode conducts for the positive half-cycle and the second diode conducts for the negative half-cycle, as shown in Fig. 4. Comparison of Figs. 2 and 4 indicates that a full-wave circuit will produce a more nearly uniform output than will the half-wave circuit.

Polyphase rectifier circuits. When high dc power is required by an electronic circuit, a polyphase rectifier circuit may be used. It is also desirable

when expensive filters must be used. This is particularly true of power supplies for the final radio-frequency and audio-frequency stages of large radio and television transmitters. The rectifier employed in polyphase circuits generally is a gas tube that has a low voltage drop in the forward direction and thus has a high efficiency. Semiconductor diodes are also used. The number of phases used in these circuits is most often 3, but 2, 4, 6, and 12 phases are used occasionally.

The simplest polyphase circuit is the three-phase half-wave circuit of **Fig. 5**. The primaries of the transformers are connected in delta to the three-phase ac line, and the secondaries are connected in wye with the common connection going to one end of the load resistor. The other end of the load resistor is attached to the B terminals of the three rectifier diodes required in the circuit. The A terminals are connected to the separate ends of the three transformer secondaries.

Operation of the circuit is such that diode D_1, connected to the first secondary, conducts for $120°$ of the ac cycle. As soon as the voltage on secondary 2 equals that of secondary 1, diode D_2 starts to conduct and diode D_1 stops conducting. Secondary voltages e_{o1}, e_{o2}, and e_{o3} are shown in **Fig. 6**, and diode currents are indicated as i_1, i_2, and i_3. The resulting load current i_L is also shown in Fig. 6. This current is much closer to a true direct current than is the current for the single-phase circuits of Figs. 1 and 3. Ripple voltage is much lower, and less elaborate filter circuits are needed to smooth the output wave. The diodes in Fig. 5 could be replaced by gas rectifier tubes or ignitrons if higher load currents were required.

Another common polyphase rectifier circuit is the three-phase full-wave or six-phase half-wave circuit shown in **Fig. 7**. Here the tubes conduct for $60°$ instead of $120°$, as in the circuit of Fig. 5. The ripple voltage for the full-wave circuit is much smaller.

Many other polyphase rectifier circuits are possible.

Bridge rectifier circuits. Bridge rectifier circuits are useful in both single-phase and polyphase applications in which a transformer must be used whose secondary has no center tap or in which dc voltages approximately equal to the total secondary voltage of the transformer must be obtained. Another use of the bridge circuit is in ac rectifier-type meters. The bridge circuit is shown in **Fig. 8**. Four separate half-wave rectifier diodes are used. When the left-hand side of the transformer secondary is positive, current flows through diode D_1, the load resistor R_L, and diode D_3. When the secondary voltage reverses, the current flows through diodes D_2 and D_4, passing through resistor R_L in the same direction as during the first half cycle.

Parallel rectifiers. If greater current is desired, two or more rectifiers can be connected in parallel. In such an arrangement small resistors or inductors are put in series with the rectifiers before they are connected in parallel.

Controlled rectifiers. Controlling the current delivered by a rectifier can be accomplished by varying
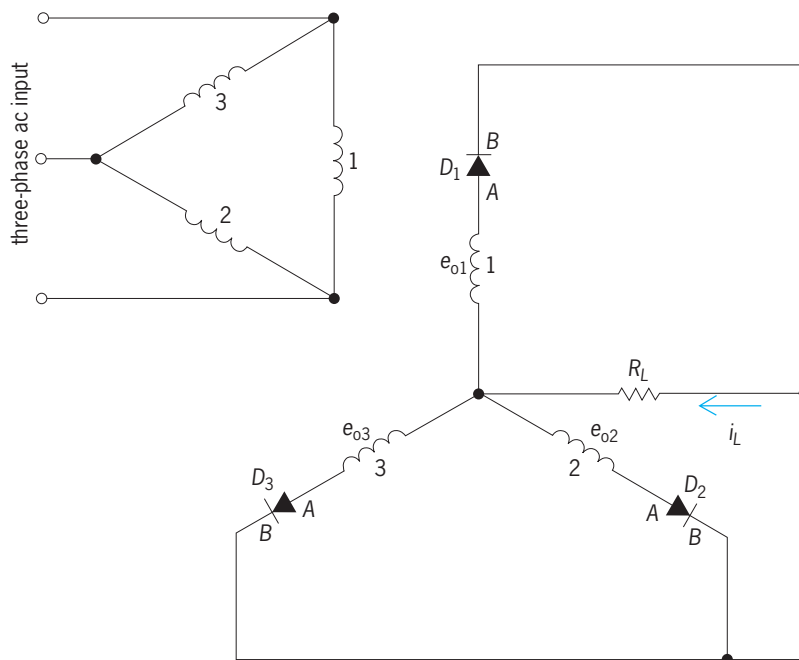

Fig. 5. **Three-phase half-wave rectifier.**

the primary voltage of the power transformer or by changing a resistance in series with the load resistor. The first technique has the disadvantage of being expensive; the second leads to poor efficiency. A more convenient and less expensive method is to control the angle at which the rectifier tube starts to conduct. Special gas tubes that accomplish this control are thyratrons and ignitrons. Silicon controlled rectifiers (SCR) may also be used for this purpose. Thyratrons are hot-cathode gas tubes with a large grid structure that prevents the arc from being ignited until the correct voltage is applied to the grid. An ignitron is a cold-cathode pool-type tube with an igniter grid actuated by an electrical pulse. The igniter of the ignitron
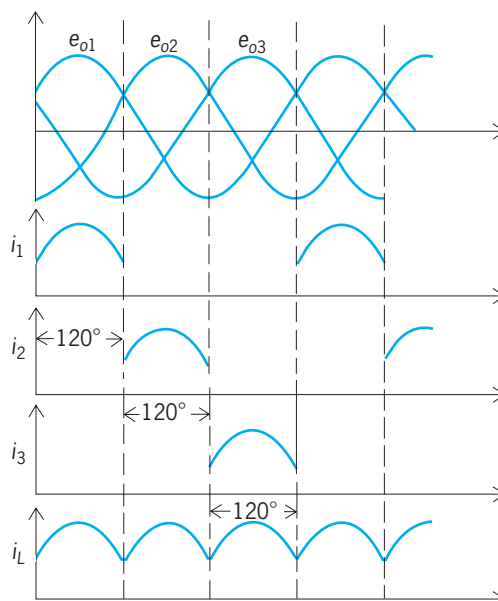

Fig. 6. **Transformer voltages, diode currents, and load current in a three-phase half-wave rectifier.**
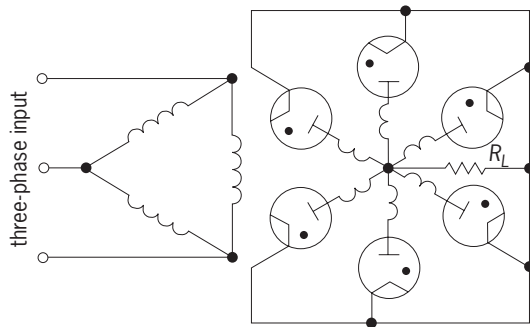
Fig. 7. Three-phase full-wave or six-phase half-wave rectifier, an example of polyphase rectifier circuit.
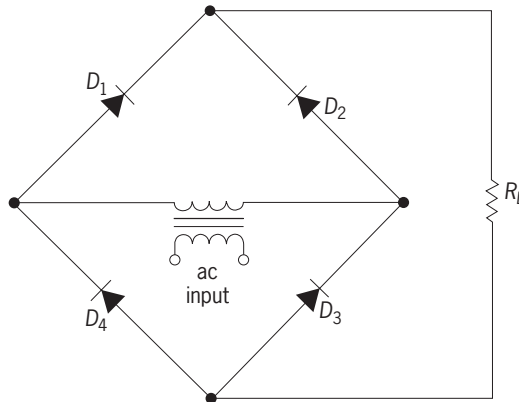

Fig. 8. Single-phase full-wave bridge circuit.

requires a substantial amount of power, which is usually supplied by an auxiliary thyratron in the control circuit. The SCR is a silicon semiconductor device which conducts when the trigger gate electrode is raised to the triggering potential. *See* CONTROLLED RECTIFIER; SEMICONDUCTOR RECTIFIER.

One circuit for an SCR is shown in **Fig. 9**, with control of the triggering point of the SCR possible over a full-half-cycle of 180 electrical degrees. When the upper terminal of the ac input is positive, capacitor C will charge to the triggering voltage of the SCR in a time determined by the RC time constant and the voltage across the SCR. Current then flows through load resistor R_L until the ac input voltage starts to reverse. When the lower terminal of the ac input is positive, capacitor C charges negatively through the
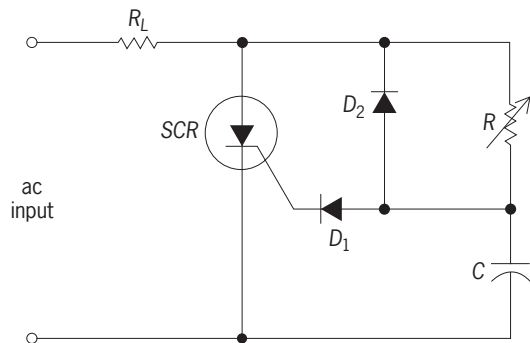
Fig. 9. One circuit for a silicon controlled rectifier with phase control triggering network.

diode D_2, and the cycle of conduction repeats. Diode D_1 prevents the negative peak voltage from appearing on the gate of the SCR. By varying the resistor R, the triggering point can be varied from zero degrees when the SCR will conduct the maximum current and to $180°$ when the SCR conducts zero current. Hence a continuous control from zero to maximum current is achieved. Other control circuits can also be used.

Inverse voltage. The inverse voltage of a rectifier is the voltage that the rectifier must withstand when it is not conducting or when it is conducting slightly in the reverse direction. As an example, in the full-wave rectifier circuit of Fig. 3, when diode D_2 is conducting, diode D_1 has impressed upon it the total secondary voltage of the power transformer minus the voltage drop in D_2. For a well-designed power supply the maximum value of the inverse voltage should not exceed the rated value of the rectifier specified by the manufacturer.

Current ratings. Another important rating for a rectifier is the average current through it. The average rectifier current of a half-wave single-phase rectifier is the same as the average load current. For a full-wave single-phase rectifier the average rectifier current is one-half the average load current. The maximum value of instantaneous current through the rectifier should not exceed the peak current rating of the rectifier. This is particularly true when capacitor-input filters are employed, because these filters generally produce high peak-to-average current ratios in the rectifier. For information on electronic circuits in general *see* CIRCUIT (ELECTRONICS)   Donald L. Waidelich

Bibliography. T. H. Barton, *Rectifiers, Cycloconverters, and AC Controllers*, 1994; A. A. Jaecklin (ed.), *Power Semiconductor Devices and Circuits*, 1993; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997.
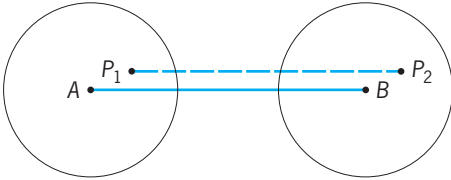
Rectilinear motion

Motion is defined as continuous change of position of a body. If the body moves so that every particle of the body follows a straight-line path, then the motion of the body is said to be rectilinear.

When a body moves from one position to another, the effect may be described in terms of motion of the center of mass of the body from a point A to a point B (see **illustration**). If the center of mass of the body moves along a straight line connecting the points A and B, then the motion of the center of mass of the body is rectilinear. If the body as a whole does not rotate while it is moving, then the path of every particle of which the body is composed is a straight line parallel to or coinciding with the path of the center of mass, and the body as a whole executes rectilinear motion. This is shown by the straight line connecting points P_1 and P_2 in the illustration. *See* CENTER OF MASS.

Rectilinear motion is an idealized form of motion which rarely, if ever, occurs in actual experience, but

it is the simplest imaginable type of motion and thus forms the basis for the analysis of more complicated motions. However, many actual motions are approximately rectilinear and may be treated as such without appreciable error. For example, a ball thrown directly upward may follow, for all practical purposes,

Rectilinear motion. All points move parallel to the center of mass.

a straight-line path. The motion of a high-speed rifle bullet fired horizontally may be essentially rectilinear for a short length of path, even though in its larger aspects the ideal path is a parabola. The motion of an automobile traveling over a straight section of roadway is essentially rectilinear if minor variations of path are neglected. The motion of a single wheel of the car is not rectilinear, although the motion of the center of mass of the wheel may be essentially so. *See* BALLISTICS; MOTION.   Rogers D. Rusk

Recursive function

A function that maps natural numbers to natural numbers and is special in that it must be calculable by using a precisely specified algorithm. The mathematical definitions of partial recursive functions and recursive functions were developed to give a precise mathematical characterization of those functions or operations on the natural numbers which are computable by using effective procedures.

Effective procedure. An effective procedure is a procedure or process determined by a finite list of precise instructions. There is no upper bound on the number of instructions in this list. These instructions can be carried out in a discrete one-step-at-a-time fashion. The only equipment needed in executing them are pencil and paper, and an unlimited supply of both is assumed to be available. No creativity is involved in applying these instructions, and no random devices such as flipping a coin are allowed in carrying them out. The instructions are very specific about all matters, in particular, as to when to terminate these instructions and when to give various outputs. Consequently, each time that these instructions are applied under the same circumstances, the calculation or computation proceeds in exactly the same way with exactly the same outcome. The computation may never end and, consequently, may go on forever. Even when the computation does terminate, there is no predetermined bound on the time required to carry it out.

While a computer program can be viewed as an effective procedure, in general, more than an ordinary computer would be needed to implement the computation specified by a program in the sense of effective procedure, since there would be no restrictions on the size of the program or on the available memory and running time. Effective procedures give a theoretical upper bound on the kind of computations that a computer or person could ever actually carry out.

Effectively computable function. An effective procedure may be used to calculate a function as follows: The calculator inputs to the effective procedure a natural number (0, 1, 2, 3, . . .) and allows the procedure to be carried out on this input. If an output is obtained, the function is said to be defined at this input and its value at this input is the output so obtained. Any function which can be computed in this way is called effectively computable. For example, addition and multiplication of natural numbers are effectively computable functions.

Turing machines. The notion of effective procedure is intuitive and vague. In the setting of an idealized computing device called a Turing machine, a more precise definition of certain effectively computable functions can be given. A Turing machine is a very simple machine which operates on a paper tape divided into blocks. Each block can contain either a B (for blank) or a 1. The paper tape is potentially infinitely long in both directions, and runs through the Turing machine, which can view or scan only a single block at a time (see **illus.**). The machine can
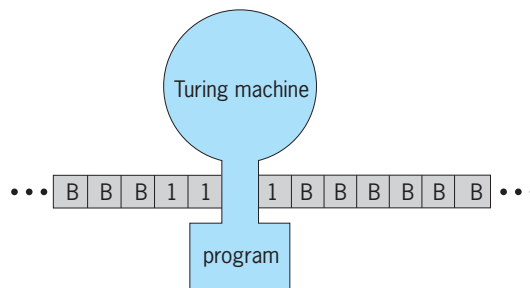
Turing machine scanning a single block on its tape.

perform only four simple actions on the tape: move the tape one block to the left (so that after this action it is scanning the block immediately to the right of the one previously scanned), move the tape one block to the right (with the corresponding effect), print a B (erase a 1), or print a 1 on the block currently being scanned. (These actions are labeled L, R, B, and 1 respectively.)

A program for a Turing machine is specified as follows: A finite set of states is used, denoted by the symbols $s_0, \ldots, s_k$. Any instruction for the machine is a quadruple of the form $(s_i; a; b; s_j)$, where $0 \leq i, j \leq k$; a is B or 1; and b is B, 1, L, or R. This instruction is executed as follows: If the machine is in state s_i and is scanning a block with the symbol a on it, it does the action specified by b and enters the

state s_j. A program P for a Turing machine is any finite set of quadruples of the above form such that there are never two different instructions in P applicable at the same time. *See* AUTOMATA THEORY.

Partial recursive functions. A Turing machine with such a program can be used to compute a unary function on certain natural numbers as follows: A number n is input by printing $n + 1$ consecutive symbols 1 on the tape (B's every place else), the tape is arranged so that the left-most 1 is the block being scanned by the machine, and the machine is started in initial state s_0. The computation proceeds according to the quadruples in its program and terminates when the machine encounters a situation for which no instruction is applicable. If the computation terminates, the output or value of the function is taken to be the total number of 1's appearing on the tape at that time. If the computation never terminates, the function is considered undefined at that input n. A k-ary function can be computed by using a Turing machine with the same kind of program by varying the input; for example, an input for a binary function consists of two strings of 1's separated by a B.

Any function that can be calculated in this way is called a partial recursive function. A recursive function (also called a general recursive function or a total recursive function) is a partial recursive function that is defined at all inputs.

Church's thesis. Besides the Turing-machine approach, many alternatives have been proposed for a more precise mathematical definition for the intuitive notion of effectively computable function, but they have all led to the class of partial recursive functions defined above. Also, all known examples of effectively computable functions on the natural numbers have been shown to be partial recursive functions. This gives evidence for Church's thesis, a widely accepted principle that any effectively computable function on the natural numbers must be partial recursive function. Church's thesis cannot be proven because it asserts that the intuitive notion of effectively computable function is precisely captured by the mathematical definition of partial recursive function; that is, that this is a good definition. Many results in mathematical logic take on a greater significance by assuming Church's thesis. For example, proving that a certain function is not recursive shows that an effective procedure could never be found for computing it. This makes it possible to prove that certain mathematical problems, whose solution would require that certain functions be effectively computable, do not have a solution or are unsolvable. *See* GÖDEL'S THEOREM; LOGIC.   George C. Nelson

Bibliography. J. Barwise (ed.), *Handbook of Mathematical Logic*, 1977; M. Davis, *Computability and Unsolvability*, 1982; H. B. Enderton, *A Mathematical Introduction to Logic*, 2d ed., 2000; S. C. Kleene, *Introduction to Metamathematics*, 1971; H. Rogers Jr., *Theory of Recursive Functions and Effective Computability*, 1987.

Recycling technology

Methods for reducing solid waste by reusing discarded materials. Essentially, recycling is a three-part process that uses waste materials to make new products. The three integral phases involve the collection of recyclable materials, manufacture or reprocessing of recyclable materials into new products, and purchase of recycled-content products. Various techniques have been developed to recycle plastics, glass, metals, paper, and wood.

Plastics

Plastic discards represent an estimated 9–10% by weight and up to 26% by volume of the municipal solid waste in the United States discarded after materials recovery. About 2% by weight of discarded plastics is recovered. Most of the effort to recycle postconsumer plastic involves high-density polyethylene (HDPE) and polyethylene terephthalate (PET) resins. Plastic waste made of these two resins is more easily identified and separated from other plastic waste as a result of industry standardization of materials for soft drink bottles (polyethylene terephthalate), base cups from these soft drink bottles (high-density polyethylene), and milk bottles (high-density polyethylene). Perhaps 26% of all plastic containers are recovered.

Plastic available for recycling. Approximately half of plastic waste consists of single-use convenience packaging and containers. Many manufacturers prefer plastic for packaging because it is lightweight, resists breakage and environmental deterioration, and can be processed to suit specific needs. Once plastics are discarded, these attractive physical properties become detriments.

Six resins are used for most commodity plastic products: low-density polyethylene (LDPE), high-density polyethylene, polyethylene terephthalate, polyvinyl chloride (PVC), polypropylene (PP), and polystyrene (PS). These are single-resin thermoplastics and are considered suitable for recycling because they become pliable when heated and can be remolded. Acrylonitrile butadiene styrene (ABS), also a thermoplastic, is used to make many computer casings. *See* POLYOLEFIN RESINS; POLYSTYRENE RESIN; POLYVINYL RESINS.

Plastics are also used with paper and metal as multicomponent films, mostly for packaging. They are more difficult to recycle than single-resin plastic material discards, and can be converted only to mixed-plastic products.

Recycled plastic products. Secondary materials are products recovered from the waste stream and reformulated into new physical forms serving end uses other than those of the original materials. A variety of secondary materials are produced from plastic waste. The outer plastic layer of disposable diapers has been recycled into garbage bags, flowerpots, and plastic lumber. Polyethylene terephthalate is recycled into scouring pads, fiber fill for jackets, paintbrushes, carpet fibers, and other products. Polyethylene terephthalate is also recycled to produce polyol for use in urethane foam and furniture, and

unsaturated polyester for boat hulls, pools, auto body parts, and appliance parts. High-density polyethylene is recycled into base cups for soft drink bottles, flowerpots, toys, and pallets. Projects to recycle bottles made from polyvinyl chloride resins have focused on products such as bottles for shampoo and vegetable oil.

Recycling postconsumer polystyrene foam has been targeted in certain pilot projects by industry for reuse as coat hangers, building insulation, office accessories, trash receptacles, and flowerpots. As much as 95% of clean preconsumer industrial plastic waste (plastic scrap from an industrial or manufacturing process) is recycled. Companies have also recycled unclean industrial plastic waste for use in inner layers, with virgin resins as outer layers, for products such as multilayer detergent bottles; however, the actual amount of unclean plastic industrial scrap that is recycled is unknown.

Research in recycling of plastics has expanded to include the development of lumber from a mix of all types of postconsumer plastic. Such a mix, commonly called commingled plastics, comprises various plastic resins, pigments, additives used in manufacturing, and nonplastic contaminants. Plastic lumber has a number of advantages over wood, such as resistance to rot, chemicals, water, and insects. Plastic lumber is being used for fence posts, poles, marine pilings and bulkheading, dock surfaces, park benches, landscape timbers, retaining walls, palettes, and parking space bumpers. Engineering tests have shown that it is better to make plastic lumber from a single polymer than from a commingled plastic.

Methods and technologies. Postconsumer plastic may be collected in a curbside recycling program or at designated drop-off centers. There are a number of problems associated with the collection of plastic waste: its high volume can burden an existing collection program; the material does not crush easily; and plastics made from different resins may be mixed together.

To assist the public and laborers working in material recovery facilities with identifying different plastics, the Society of the Plastics Industry sponsors a voluntary coding system. However, in the United States it is required by law in 39 states. The coding system consists of a triangular arrow stamp with a number in the center and letters underneath to identify the resin used in the container.

The collected plastic waste is usually separated manually from the waste stream, and often it is cleaned to remove adhesives or other contaminants. It is sorted further, based on different resins. Mechanical techniques to separate plastics are available, and they can be used to sort plastics based on unique physical or chemical properties. It is anticipated that more sophisticated methods under development will provide better systems for separating plastics made of different resins.

The technologies used to manufacture recycled plastic are virtually the same as those used to manufacture products from virgin plastic resins. The recycled products are melted; for some products, additives or virgin resins are used to improve the properties. The plastic is then extruded into specific products or pellets.

For secondary lumber-type materials made from mixed plastic waste, the technology generally used is known as Extruder Technology 1 (ET/1). ET/1 has three main components: an extruder, a molding unit, and an extraction unit. The extruder is used for melting the plastic at temperatures of 360–400°F (182–204°C). Plastic with higher melt temperatures, such as polyethylene terephthalate, become encapsulated within the melted plastic. The molding unit consists of linear molds that are water cooled. The product shrinks within the mold during cooling. It is possible in some units to make lengths that exceed those of typical wooden lumber. It is also possible to chemically break down the plastic polymers into monomers or oligomers. These are cleaned and can then be made back into the polymer. This process perhaps offers the best opportunity for reusing the plastic for the same product as was originally made from the virgin material.

Engineering and environmental issues. A significant problem in plastic recycling is the presence of contaminants such as dirt, glass, metals, chemicals from previous use, toxicants from metallic-based pigments, and other nonplastic materials that are part of or have adhered to the plastic products. The U.S. Food and Drug Administration (FDA) has expressed concern over potential risks in using recycled plastic products in contact with food. However, a redesigned chemical recycling process has been used to recycle PET into soft-drink and salad-dressing bottles. Other constraints involve inconsistencies in the amount of different plastic resins in commingled plastic wastes used for recycling, and engineering aspects of recycled plastic products, such as lessened chemical and impact resistance, strength, and stiffness, and the need for additional chemicals to counteract other types of degradation for reprocessing. There may be limitations to the number of times a particular plastic product can be effectively recycled as compared to steel, glass, or aluminum, which can be recycled many times with no loss of their properties and virtually no contamination.

The long-term engineering properties of recycled products are still largely unknown. The many products made from recycled plastic include gimmick toys, penholders, and paper trays; these may appear in the waste stream in a short time.

The recycling of plastics creates a waste stream of its own—contaminated wastewater and air emissions. Many additives used in processing and manufacturing plastics, such as colorants, flame retardants, lubricants, and ultraviolet stabilizers, are toxic and may be present in the waste stream.

Economic and social acceptability. Markets for recycled single-resin plastics have had more success than mixed-plastic products. The reason may be that bottle return laws lead to better collection of single-resin polyethylene terephthalate beverage containers and to focusing by the plastics industry on

recycling technologies for single-resin plastics. Marketing problems that arise with products made from commingled plastics include inconsistency in feedstock, lack of suitable engineering specifications, and unpredictable performance.

An important issue related to the marketability of recycled products is whether such products are, in fact, cheaper than the products they will be replacing. Plastic lumber, for example, is expensive compared to wooden lumber, concrete, and other materials for which it can be substituted. However, product lifetime must also be taken into consideration with the initial cost of a material. The average lifetime of plastic lumber and other products made from commingled plastics is undetermined. Other factors, such as plastic lumber's low maintenance costs, are also significant considerations.

For products made from commingled plastics, marketing problems due to a reduced esthetic appeal may arise: there is a lack of black and very dark colors, and impurities such as bits of paper or metal can be seen within the material.

A notable marketing advantage for recycled plastic products over items made from virgin plastics or other material is that they are appealing from an environmental standpoint to many consumers.

Casings for computers, often acrylonitrile butadiene styrene, are a growing part of the waste stream. These casings are made of an engineered thermoplastic and therefore can be recycled. Unfortunately, few of the component parts of computers are recycled, largely because of the cost of disassembling. Recycling plastics is an essential part of a national waste prevention and waste reduction strategy. Recycling plastic that may otherwise be diverted to landfills, incinerators, or roadside litter is a positive step in reducing the waste stream and should be incorporated in all programs for solid-waste management. Research, engineering specifications, government regulations, and continued interest are necessary for future success of products made from recycled post-consumer plastic.

R. L. Swanson; Vincent T. Breslin; Marci L. Bortman

Glass

Recycling of glass is an important aspect of solid-waste management. Glass containers are a usual ingredient in community recycling programs; they are 100% recyclable and can be recycled indefinitely. Moreover, the recycling process creates no additional waste or by-products.

In 1993 in the United States, glass containers, which constitute 6% of the solid-waste stream by weight, were recycled at a rate of 35%. This means that nearly one-third of the glass containers available for consumption in the United States were cycled back into glass containers and other useful items such as glasphalt, or were returned as refillable bottles.

Collection methods. There are three primary methods of collecting glass: drop-off centers, buy-back centers, and curbside collection. Drop-off centers and buy-back centers typically collect less material,

but they also have much lower capital and operating costs than curbside programs.

Drop-off centers are one of the simplest forms of glass recovery. These may be mobile collection stations or permanent sites maintained by local municipalities. In some areas, fiberglass domes known as igloos are used as around-the-clock collection points. In other places, a drop-off may be as simple as several 55-gal (208-liter) metal drums, each marked for the color of glass it is to hold.

Buy-back centers purchase glass and other recyclables from people who voluntarily transport the material to the site. The centers serve as a convenient market for recyclers who cannot sell directly to a glass plant.

Comprehensive curbside collection programs are increasingly popular. Curbside collection programs allow residents the opportunity to put their recyclable materials out at the curb, separated from their regular trash, for pick-up. Many programs offer commingled collection, which affords a resident the opportunity to put all recyclable material into one container or bag. Glass is an integral part of these programs. Often aluminum and glass provide the necessary revenue to sustain the community's program.

Processing specifics. The process of recycling glass is relatively straightforward. Used glass bottles and jars are mixed with silica sand, soda ash, and limestone in a melting furnace at temperatures up to 2800°F (1540°C). The molten glass is poured into a forming machine, where it is blown or pressed into shape. The new containers are gradually cooled, inspected, and shipped to the customer. Before glass can be recycled, however, it must be furnace-ready, that is, sorted by color and free of contaminants.

Color sorting. Sorting of green, amber, and flint (clear) glass containers is essential in order to maintain color consistency during the manufacturing process. Darker brown and green containers are used in many cases to help preserve the shelf life of a food product.

Different chemical agents are used to color glass. For example, chromium salts are generally used to create a green color, selenium salts are used to produce clear bottle glass, and iron oxides impart an amber finish. The coloring properties of these chemicals are so strong that a small amount of miscellaneous colors in a glass batch mix will adversely affect the result. Thus it is critical to deliver color-separated glass to manufacturing facilities in order to avoid costly problems involving equipment and production.

A considerable amount of research is focused on developing automated color-sorting equipment that can be used at glass processing and manufacturing plants. In addition, there is experimentation with the staining of flint glass bottles. The coating would be burned off in the furnace, making it unnecessary to separate by color.

Beneficiation plants. Cullet must meet a standard of quality similar to that of the raw material it replaces. Contamination from foreign material will result in the cullet being rejected by the plants, as it poses

a serious threat to the integrity and purity of the glass packaging being produced. Contaminants include metal caps, lids, stones, dirt, and ceramics. Paper labels do not need to be removed for recycling, as they burn off at high furnace temperatures.

Many glass manufacturers have invested in glass-processing equipment. These beneficiation units remove metals and other contaminants. For example, the New York State Energy Research and Development Authority, with funding and technical assistance support from several private entities including the Glass Packaging Institute, introduced a new glass beneficiation system that can optically sort out ceramics and separate nonferrous metals from the cullet. A nonferrous metal separator has been developed. This system removes aluminum, lead, and small contaminants, including the rings of residue in bottlenecks. Following this process, the cullet is moved to the optical ceramic sorter, which removes unwanted ceramics by releasing quick jets of air activated by an infrared light that detects the ceramic's opacity. Several other efforts are going on across the country to develop improved methods of processing furnace-ready cullet.

Energy and environment benefits. From a manufacturing standpoint, cullet can reduce wear and tear on furnaces, and maintenance expenses, since it can be melted at a lower temperature than that required to combine virgin materials. This reduced melting point can save as much as 25% of the energy used to make virgin glass, depending on what type of heating is used in the furnace. The energy saved by recycling a single glass bottle is estimated to be the equivalent of lighting a 100-W bulb for 4 h. Similarly, for every soft-drink bottle recycled, enough energy is saved to run a television set for $1^1/_2$ h.

Glass recycling also has environmental benefits, such as reductions in emissions to the atmosphere. It has been estimated that 27.8 lb of air pollution is produced for every ton of new glass produced (13.9 kg/metric ton), and recycling glass reduces this source off pollution by 14–20%. Recycling a ton of glass saves 1330 lb (599 kg) of sand, 433 lb (195 kg) of soda ash, 433 lb (195 kg) of limestone, and 151 lb (68 kg) of feldspar. Generally costs for cullet are less than for soda ash, which it replaces at a ratio of 3.5 to 1.

End markets. The primary recycling market for container cullet consists of the manufacturers of glass containers. However, over the years difficulties involving color sorting and transportation have spawned new applications for cullet. Thus cullet use in secondary materials has increased. There markets include fiberglass, glasphalt, roadbed, reflective beads, decorative glass, and drainage. One of the most visible secondary markets is manufacture of glasphalt, which is a type of asphalt mix that incorporates crushed glass. Glasphalt was developed as an application of commingled cullet. One useful characteristic of glasphalt is its ability to retain heat. Glass acts as an insulating agent, making glasphalt workable under much colder environments. Manufacture of fiberglass, predominantly used in the form of glass

wool for thermal and acoustical insulation, is another common secondary market for cullet.

Nathan Tyler; Natalie U. Roy

Metals

A modern society depends on a tremendous quantity of metals to provide for transportation, housing, communications, and other essentials of a healthy, comfortable existence. To meet these needs with minimal environmental damage, metals must be recycled to alleviate the need to mine more ore, to reduce energy consumption, to limit the dissemination of metals into the environment, and to reduce the cost of metals. In the United States a substantial portion of these needs are met by recycling metals.

The extensive recycling of materials is very important for three reasons: (1) The energy required to recycle a metal is considerably less in comparison to producing it from ore. (2) Extracting the metal from ore produces a tremendous amount of waste material. For instance, to produce 1 ton (0.9 metric ton) of copper metal requires the moving of 350 tons (315 metric tons) of overburden (the waste material above the ore body); 150 tons (135 metric tons) of ore, of which 145 tons (131 metric tons) ends in tailings (the waste material left after the metal is removed from the ore); plus using 120,000 gal (450,000 liters) of water and consuming 6 tons (5 metric tons) of air. In addition, 2.7 tons (2.4 metric tons) of sulfuric acid and 1.8 tons (1.6 metric tons) of slag are produced. The United States produces about 1.5×10^6 tons (1.4×10^6 metric tons) of copper from ore each year, resulting in the generation of large waste streams. (3) Metals that are not recycled eventually begin to be dissipated throughout the environment. Since many metals are toxic, this can result in the pollution of water and soil. Lead and cadmium are examples of metals that are considered toxic. Lead was used in paints and as a fuel additive and thus was widely disseminated throughout the environment. With the recognition of its toxic properties, it was virtually eliminated as a fuel additive and its use greatly reduced in paints. Cadmium is used in rechargeable batteries, and these batteries have become much more widely used. However, methods to recover cadmium from these batteries has not kept pace and, as a consequence, landfills are being contaminated with these batteries. *See* HAZARDOUS WASTE.

Challenges. While it is beneficial to recycle materials, there are at least two important problems that hinder recycling: collection and impurity buildup. When a metal becomes scrap and is a candidate for recycling, it must be collected at a cost that makes it attractive to recyclers. It is useful to divide scrap into three categories—home scrap, new or prompt scrap, and old or obsolete scrap—whose methods for collection differ significantly. Home scrap is waste produced during fabrication, and includes casting waste (for example, risers), shearings and trimmings, and material that has been rejected. This scrap is usually recycled within the plant, and therefore it is not recognized as recycled material in recycling

statistics. New or prompt scrap is waste generated by the user of semifinished material, that is, scrap from machining operations (such as turnings or borings), trimmings, and rejected material. This material is collected and sold to recyclers and, if properly labeled and segregated, it is easy to recycle and is valuable. Old or obsolete scrap is waste derived from products that have completed their life cycle, such as used beverage cans, old automobiles, and defunct batteries. The collection and the impurity buildup problems are most severe when considering old or obsolete scrap. The recycling of automobiles will be used to demonstrates these problems.

Automotive recycling. The recycling of automobiles illustrates the problems associated with recycling old or obsolete scrap. The first step in the recycle process is carried out by the dismantler. The dismantler's role is to remove the high-value parts, plus engine blocks and fuel tanks that have economic value but also must be removed for safety reasons. The auto hulk is then compressed into a flattened body. Next, the shredder takes the flattened bodies and tears them into fist-sized pieces by using hammer mills. By using techniques such as magnetic separation (to remove the ferrous material), air streams, or hand sorting, the ferrous metals, nonferrous metals (primarily aluminum), and automotive shredder residue are separated. The automotive shredder residue, which consists of polymers, glass, fibers, dirt, and other materials, goes primarily to landfills, since it is not economic to recycle. The ferrous scrap goes to minimills to make low-carbon steels used in fencing, reinforcing bars, and other items; and to integrated steel producers to be used in basic oxygen furnaces, where it is mixed with pig iron to make various grades of steel.

The nonferrous metal scrap is then separated into primarily three groups; aluminum, copper, and zinc. The lead has been separated previously by the dismantler by taking out the batteries, of which 97% are recycled. The aluminum generally goes to the secondary aluminum market, where it is used to produce aluminum foundry alloys for use in the auto industry. Copper must be separated from ferrous material, because it severely degrades the properties of steel. Zinc is becoming more of a problem because the steel sheet used in the body of the car is being coated with zinc to reduce rusting.

Major new problems have developed in auto recycling. The desire to produce autos of lighter weight for greater fuel economy has resulted in increased use of high-strength alloy steels, aluminum, and polymers. Little recycling of polymers occurs, and therefore more automotive shredder residue is produced that must be deposited in landfills. The ferrous materials retain some of the alloying elements plus some contamination from copper used in wiring, making the recycled steel lower in value. In the future it will be necessary to make automobiles easier to disassemble if their recycling is to remain economically viable.

Problems have developed because of the undesirable alloying elements in the recycled material.

There are hundreds of different alloys of steel and aluminum, each with very desirable properties for certain applications, but disastrous in other applications. For instance, small amounts of lead added to steel (free-machining steel) make it much easier to machine, but this same lead can make steel susceptible to unexpected brittle fracture when stressed at elevated temperatures. With more and different materials being mixed together in the waste streams, it is ever more likely that impurity contamination will be a problem.

Iron/steel. The tonnage of iron and steel used in the United States dwarfs all other metals combined, and therefore it is by far the most important to recycle.

A major change in the production of steel began in the 1960s, with the basic oxygen furnace and the electric-arc furnace replacing open-hearth steel making. About 70% of steel is made by the basic oxygen furnace process, and 30% by electric-arc furnaces. The principal feedstock for electric-arc furnaces is recycled steel, with a significant portion being recycled automobiles. Some scrap metal is also used in the basic oxygen furnace, but it is significantly less in comparison to that used in the old open-hearth method. Therefore, the two processes are complementary in balancing the use of virgin and recycled metal.

Before 1965, the electric-arc melting furnace was used primarily to produce specialty steels (such as stainless or tool), but it is now a major process for producing low-carbon steel for wire and rod for making nails, netting and mesh, tire cord, spring wire, fencing, and re-bar in minimills. *See* ARC HEATING; IRON METALLURGY; STEEL MANUFACTURE.

Aluminum. Recycling of aluminum is done by independent secondary aluminum smelters in addition to the large integrated aluminum producers. The secondary smelters use gas- or oil-fired reverberatory furnaces of 30,000–100,000 lb (13,500–45,000 kg) capacity to remelt and refine scrap aluminum, with much of the product going into die castings for the automotive and appliance industries. Old scrap represents about 60% of the 2.7×10^6 tons (2.4×10^6 metric tons) of aluminum produced from scrap, and of this about 50% is used in beverage cans. *See* ALUMINUM.

Copper. Since copper is a relatively expensive material, it is recycled extensively. However, it is used in products that have a long lifetime, for example, electrical machinery (30-year life). Consequently, less than half of the production of copper comes from scrap, of which slightly more than 50% is new scrap. *See* COPPER METALLURGY.

Lead. The uses of lead have changed dramatically. Before 1972, lead was used extensively in gasoline additives, pigments, ammunition, and chemicals; however, with the concern about lead toxicity these uses were greatly reduced. Now the primary use of lead is in batteries, and very effective means of recycling have been established (primarily the buy-back of used batteries by retailers). Used lead batteries are sent to a secondary lead smelter, where they are

shredded and material is produced that is sold back to the battery makers.

Zinc. About 75% of the zinc produced is used in the metal form, with the remaining 25% used to produce chemical compounds that have many applications in agriculture and manufacturing. The compounds are generally not recycled; about 90% of the elemental metal is used in the galvanizing of steel and its alloys. Consequently, only about 10% of the production of zinc is from old scrap. Many techniques are used to recycle zinc, with the recycling of zinc-coated steel from automobiles growing in importance. *See* ZINC.   Dale F. Stein

Paper

Paper and paperboard used for making recycled paper products come from a variety of sources, including offices, retail businesses, converters, printers, and households. The accepted term for such is "recovered paper." "Wastepaper" is applied to paper stock that is not recovered for recycling and is disposed of in landfills or incinerated. Recovered paper products that have been distributed, purchased, and in some way have served their intended primary purpose are classified as postconsumer paper. Examples are used newspapers and magazines, corrugated containers, writing paper, and copy paper. Recovered paper from paper manufacturing (mill broke) and converting operations (trimmings), which is of excellent quality and is often used to make recycled paper, is not considered postconsumer paper. It is sometimes referred to as preconsumer paper. "Recycled paper" is the paper made, at least in part, from recovered paper. Recycled paper often contains a combination of postconsumer paper fiber, preconsumer paper fiber, and virgin fiber. The percentge of recycled fiber (and often the percentage of postconsumer fiber) is specified for the product.

Recovered paper grades. There are over 50 grades of recovered paper and several broad classifications. Five categories that are often used throughout the industry are (1) old corrugated containers, which account for nearly half of the paper and paperboard recovered in the United States and include used corrugated shipping packaging; (2) old newspapers, the second most recovered grade of paper; (3) mixed papers (a catch-all category), including old mail, magazines, office paper, and computer paper; (4) high-grade deinking, which are mostly printed and unprinted white papers collected from converting operations, printing plants, and offices; and (5) pulp substitutes, which are a general classification of high-quality recovered paper grades that can be used directly in paper manufacture without cleaning, bleaching, or deinking. Over 95% of pulp substitutes have been recycled for many years because they are a clean, high-quality material.

Each category comprises different types of papermaking fibers that are useful for manufacturing different types of recycled paper products. The efficiencies in the collection and transportation to recycling mills, as well as advances in fiber processing technology, also have an impact on the economies of using

one grade over another. Although not all grades of paper can be recycled because of their components or prior use (for example, medical papers or tissues), recycling will increase until a practical limit is reached. This limit is considered to be 50–60%. In 2000, the paper recovery rate (total paper recovered for recycling as a percentage of paper consumed) was 48%.

Recovered paper fibers are used in the manufacture of many recycled paper products such as paperboard, corrugated containers, tissue products, newspapers, and printing and writing papers. They can also be used in other products such as insulation, packing materials, egg cartons, and flowerpots. *See* PRINTING.

Processing. Before paper can be recovered and used again, it must be collected, sorted, and cleaned. The first step in recycling, collection occurs in curbside programs, in drop-off centers, in paper drives, and more recently in commercial collection programs run in conjunction with waste (garbage) collection for landfill or incineration. Many factors have an impact on collection programs, including legislation, landfill availability, proximity to recycling mills, and levels of community commitment and participation.

Reprocessing begins by sorting recovered papers by grade and level of cleanliness. This can be done at the source, during collection, or by a broker. At the recycling plant, the recovered paper is processed to remove contaminants and produce a pulp that is sufficiently clean and strong to be used for papermaking. An important difference between processing of virgin pulp and recovered pulp is that recovered paper requires the removal of a much wider variety and often a much higher level of contaminants. Many of these materials are inherent in the production, converting, and end use of paper products. Some common contaminants are inks, plastic, polystyrene foam, metals, glass, dirt or sand, and adhesives. There are also a number of nonfibrous materials used in making paper, such as fillers (clay, calcium carbonate, titanium dioxide), starch, latex, and resin, that may need to be separated from the fibers. Some contaminants have an adverse affect not only on the quality of the paper being made but also on the papermaking process. If they are not removed in the recycling process, contaminants such as adhesives, resins, waxes, hot melts, and ink binders can build up on the papermaking equipment as sticky deposits, which require expensive downtime for cleanup. In general, the amount of contaminants removed depends on the type present, the grade of recovered paper being used, and the requirements of the paper being made. For example, grades such as boxboard and corrugated containers are less demanding and require a simpler cleaning process than printing and writing grades.

The basic operations used at a recycling mill to remove contaminants include (1) pulping to break down the recovered paper to a fibrous state and begin separation of contaminants from the fibers; (2) screening to mechanically separate large

contaminants from the pulp; (3) centrifugal cleaning to separate by specific gravity the contaminants that are heavier and lighter than the fibers; (4) flotation to separate and remove ink particles; and (5) washing to remove very small particles not eliminated in the previous steps.

In the pulping step, the recovered paper (usually in the form of bales) is mixed with a large quantity of water in a slusher or pulper to produce a fiber-and-water slurry called pulp. Chemicals may be added to the pulper to decolorize the paper and separate the contaminants from the fibers for removal in later stages. After the paper has been sufficiently agitated, the pulp is withdrawn from the bottom of the pulper through a screen plate to remove large contaminants (greater than about 5 mm or 0.2 in.).

The pulp is then diluted with water and pumped through a screening operation that separates contaminants that differ in size and shape from the fibers. Both coarse screens with holes (often 1.6 mm or 0.062 in. in diameter) and fine screens with slots (0.25–0.10 mm or 0.01–0.004 in. in width) may be used. Flakes, plastics, hot melts, labels, and films are removed in screening.

Screening is usually followed by centrifugal cleaning which subjects the pulp to a vortex in a tapered cone and separates particles that differ in specific gravity from the fibers. Specially designed cleaners are used to separate high-specific-gravity materials such as sand, metal, and fillers, and low-specific-gravity materials such as Styrofoam™, waxes, polyethylene, and some plastics.

If the pulp requires deinking, a flotation step is included in the recycling process. There are various designs of flotation cells, but the basic process involves injecting air into the pulp. Ink particles (and some small adhesive particles) attach to the air bubbles and float to the surface, where they are removed as an inky foam.

The smallest particles, often 30 micrometers (0.001 in.) or less, are dislodged from the pulp in a subsequent washing stage. Various systems are used that dilute the pulp with water, loosen the particles, and wash them from the pulp by gravity, vacuum, or pressure.

Multiple screening, cleaning, and flotation stages may be used to ensure an adequate level of contaminant removal. If materials are still present that cause objectionable specks or operational problems, some mills add a dispersion step. By using mechanical energy or heat and pressure, dispersion reduces these contaminants to a minute size, minimizing their negative effects. If the recycled pulp is to be used to make white paper, a bleaching or color-stripping step may be added to brighten the pulp and remove color.

Trends and challenges. Continuing issues include the management and control of fiber from recovered paper, the need for expanded recycling programs (supply), and environmental considerations.

As with virgin pulp, the purpose of managing recovered fiber is to ensure that the right mix of fibers, at an economical cost, is available for the paper products being made. A mill making paper from virgin pulp will control the species of trees and the type of market pulp it uses, while a recycled paper mill has to control the grades of recovered paper (and the associated level and type of contaminants) it accepts. The need by a mill for a consistent supply of uniform-quality recovered paper grades requires that paper procurement practices be carefully coordinated. This may involve working closely with community curbside collection programs as well as with the many brokers and dealers who serve as intermediaries for buying and selling recovered paper.

The amount of recovered paper used is steadily increasing. Currently, about 81% of the paper recovered in the United States is recycled into new products in the nation's paper mills. About 16% is exported, and the rest is used domestically in nonpaper products such as cellulose insulation, animal bedding, and mulch. Requirements by the federal government regarding the procurement of paper with a minimum recycled fiber content have helped spur the production and use of recycled paper. With this increased consumption comes the challenge of continued supply. Already certain grades of recovered paper, such as old corrugated containers and old newspapers, are approaching their practical recovery limit. This means that more efficient collection systems have to be developed for these grades, and underutilized grades and sources have to be used. These underutilized grades often have a higher level of contaminants and contain materials that are difficult to handle. A notable example is the pressure-sensitive adhesives on labels and stamps. Unless removed in the recycling process, these materials can build up on papermaking equipment as sticky deposits. Paper printed with certain inks can also present problems. The need to use these less desirable papers will continue to drive the development of more efficient recycling technologies. At the same time, more attention is being given to product life cycle management and the use of "environmentally benign" materials in paper and packaging that are easily removed in the recycling process. Another trend is the increased use of nonchlorine bleaching processes for color removal from recovered paper.

With the use of paper having more contaminants comes the challenge of finding ways to use the residue that the mills have to discard. Often this material goes to landfill, but some has been incorporated into products such as concrete and peat pots. Other research challenges include the need to develop standardized methods to measure the quality of both recovered and recycled fiber, and continued improvement in the properties of fibers used to make recycled papers. *See* PAPER.   Paul W. Resler

Wood

Waste is generated at every stage of the process by which a forest tree is turned into consumer and industrial products. Additional waste is generated in the disposal of those products. Trimmings and limbs are left in the forest. Historically, this remainder was simply piled in the forest and burned as slash. Wastes from the sawmill or from other forest-products

manufacturing facilities, for example, furniture and millwork companies, are called mill residues. Crates and pallets used for packaging are eventually disposed of as waste. Wood waste is also produced by the homeowner and by large and small businesses and is generated from landscaping and agricultural operations such as pruning and tree removal. The United States generates about 1.5 lb (0.7 kg) of wood waste per person per year, and wood waste accounted for 6% of total municipal solid waste (see **illustration**).

For each type of wood residue, there exists a recycling technology. Lumber and flooring can be reclaimed during the renovation and demolition of residential and commercial buildings. In addition, much of the wood residue generated is burned for energy recovery or is composted. While these processes do not strictly return wood to the economy in its original form, they have the effect of diverting wood residues from the landfill, and thus they may be included under a broadly interpreted definition of recycling. Several types of waste have been recycled for many years. For example, the particleboard industry originally developed in response to the availability of inexpensive sawdust. Nonetheless, much wood waste ends up in landfills. Diverting this wood from the landfill has become more important as disposal costs have risen. Municipal solid waste contains 5–20% solid wood and solid wood products such as plywood and particleboard, depending upon the season and geographic area.

Mill residues. These wastes consist of bark, sawdust, edgings, clippings, slabs, shavings, or chips produced by primary industries in the production of lumber, timbers, poles, plywood, and other wood products. Substantially all of this waste is recovered by the primary industries. Most of it is used for fuel, but the remainder is recycled into other products. It is considered a clean waste in comparison to that obtained from municipal solid waste. The secondary wood products industry also produces some clean waste wood. These are firms that use the products of the primary industry to produce more complex products, for example manufacturers of furniture, fencing, and windows.

Industrial and municipal wastes. The solid wood in municipal solid waste comes from several sources: wood left over from the construction, remodeling, or demolition of houses and other buildings (known as construction and demolition waste); wood from discarded pallets, crates, and other types of packaging; wood used in railroad crossties, utility poles, and other types of wood that have been treated with chemicals (treated wood).

Construction and demolition waste. This type of wood waste includes items such as siding removed from old barns, which is often reused in new buildings and valued for its weathered look. It also includes wood that has been treated with paints or stains, glues, binders, or resins (found in plywood, particleboard, and laminated veneer lumber), and wood composites (such as vinyl- or melamine-laminated wood for countertops). If construction or
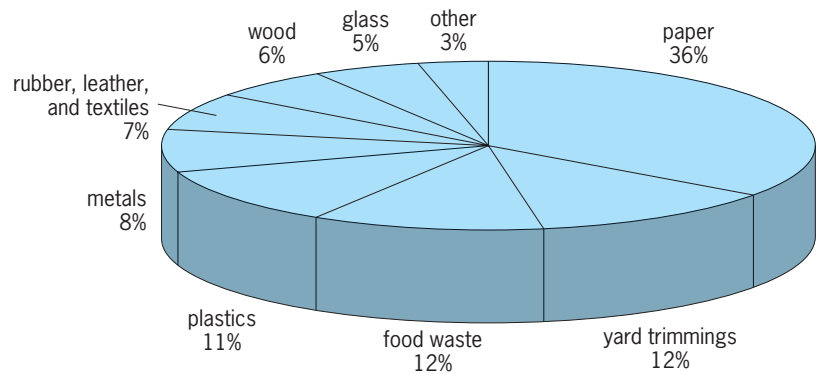
Municipal solid waste. (*U.S. Environmental Protection Agency*)

demolition waste is segregated while the structure is being built or demolished, the wood and some other materials can be recycled. The solid wood from the demolition waste stream is either removed from the building before it is demolished or separated from the other wastes. If not, it is invariably mixed with other materials such as cement, gypsum board, or glass. Then it must be separated before it can be recycled, and this separation must be done by hand. Therefore the value of the wood recovered must justify the expense of the separation. Only a small proportion of construction and demolition waste wood is directly recycled and used again as lumber. Most of this wood is processed into other products.

Packaging waste. Waste wood from packaging consists primarily of pallets. Since the mid-1970s the pallet industry has redesigned and standardized the pallet for recycling. This has led to a rapid increase in pallet recycling. Over 90% of the wood in recycled pallets is used to make new pallets. The remaining wood is used either as a fuel or in various products. An unknown number of pallets are also diverted by landfill operators and firms other than pallet manufacturers. These pallets are probably ground and used primarily for fuel or raw material for composite-board products.

Treated wood. A unique class of waste wood is treated wood, which has been injected with chemicals to prevent its decay. Disposal of these items presents problems due to the toxicity of the chemicals in the wood. In the United States, laws in most states allows treated wood to be disposed of in municipal landfills. Wood treated with creosote can be burned for fuel in permitted facilities. However, other types of treatment, such as chromated copper arsenate, yield wood containing heavy metals that cannot be safely burned. The preferred method for dealing with treated wood removed from service is recycling. For example, railroad ties that have been removed from service often continue to serve a useful purpose as landscape timbers. However, the disposal of treated wood and the contamination of the solid waste stream by treated wood are problematic.

Processing. Wood recycling begins with wood separation from the waste stream. Recovered

materials can then be processed into various products, including fuel, raw material furnish for the production of particleboard or other wood-composite panel products, compost, landscaping mulch, animal bedding, landfill cover, amendments for municipal solid waste and sludge compost, artificial firewood, wood-plastic composite lumber and other composite products, charcoal, industrial oil absorbents, insulation, and specialty concrete.

Most of these products require that the wood be ground into small particles. A typical grinder used for this application is a hammermill, although a variety of grinders are used. The size of the particles produced is determined by the end use of the wood; sizes smaller than about 20 mesh are called wood flour (the particles passing the 20-mesh screen are usually less than 0.8 mm in size). Though not produced by a saw, wood flour has the appearance of sawdust. Wood may then be passed over an electromagnet which removes items made of ferrous metals such as nails or staples. If additional processing is performed, it is typically to separate wood particles by size. This is accomplished in two ways: the particles can be passed through a series of screens of different mesh size and the various sized particles collected from the screens; or the particles may be separated in a tower with air blown in the bottom and out the top. The particles distribute themselves in the tower with the smaller, lighter particles on the top and the larger, denser ones at the bottom.

Wood that has been ground for use as fuel is known as hog fuel. The type of burner or furnace in which the hog fuel is used determines the applicable particle size. The range of particle sizes used in wood fuel is very large, from fine powders in fluidized-bed burners to fireplace logs. An important consideration in this use is moisture content. As the moisture increases, the amount of energy that can be recovered decreases, because some of the heat of combustion is used to convert the contained moisture into steam. About 30% of waste wood from all sources is used for fuel. However, this proportion varies from perhaps 15 to 50%, depending upon geographic region. *See* WOOD PRODUCTS. John Simonsen

Bibliography. M. J. Coleman (ed.), *Recycling Paper: From Fiber to Finished Product*, vols. 1 and 2, 1990; Earthworks Group, *The Recyclers Handbook*, 1990; F. R. Field III and J. P. Clark, Automobile recycling: Environmental policymaking in a constrained marketplace, *J. Metals*, pp. 17–21, April 1994; Glass Packaging Institute, *Glass Recycling: Why? How?*, 1991; T. Laufenberg, R. Rowell, and S. Sobozinski (eds.), *Materials Interactions Relevant to Recycling of Wood-Based Materials*, 1992; H. F. Lund, *The McGraw-Hill Recycling Handbook*, 2d ed., 2000; R. McKinney, *The Technology of Paper Recycling*, 1993; R. K. Miller and M. E. Rupnow, *Paper Recycling*, 1991; G. A. Smook, *Handbook for Pulp and Paper Technologists*, 2d ed., 1997; R. J. Spagenberg, *Secondary Fiber Recycling*, 1993; C. G. Thompson, *Recycled Papers: The Essential Guide*, 1992; Y. Virtanen and S. Nilsson, *Environmental Impacts of Waste Paper Recycling*, 1993.

Red dwarf star

A low-mass main-sequence star of spectral classes M and L. Red dwarf stars range from about 0.6 solar mass at class M0 down to 0.08 solar mass in cool M and warm L, below which the proton-proton chain cannot run. Lower-mass bodies are termed brown dwarfs. At the transition, within class L, the two are difficult to distinguish. Effective temperatures range from 3800 K (6400°F) at class M0 down to 2000 K (3100°F) at class L0, and absolute visual magnitudes from +9 to +20. Downward along the main sequence, red dwarfs produce progressively more radiation in the infrared. Bolometric corrections (which account for invisible radiation) range from −1.2 magnitudes to around −6, so luminosities range from about 0.05 down to 2×10^{-4} the solar luminosity. Radii range from about 0.5 to down 0.1 solar. Spectra become increasingly complex, with titanium oxide and neutral metals appearing at the warm end of class M, these and vanadium oxide appearing toward the cool end, and neutral metals and metal hydrides appearing in class L. *See* BROWN DWARF; MAGNITUDE (ASTRONOMY); PROTON-PROTON CHAIN.

Red dwarfs constitute over 70% of all stars, and as a result constitute about half the visible mass of the Milky Way Galaxy. The nearest stars are also dominantly red dwarfs, with the closest, Proxima Centauri (class M5), only 1.3 parsecs (4.2 light-years) away. Yet their intrinsic faintness renders them all invisible to the unaided eye; Proxima Centauri is of the 11th magnitude. Their lifetimes are so long that none has ever evolved in the Milky Way Galaxy's lifetime. *See* MILKY WAY GALAXY.

Warmer red dwarfs have (like the Sun) convective envelopes and radiative cores, but below about 0.3 solar mass, around class M4, they become convective throughout. Rotation and convection produce strong magnetic dynamos and complex fields that in turn produce powerful flares, especially common between M0 and M5. In a typical flare, the star increases in brightness by several tenths of a magnitude in less than a minute, the flare producing ultraviolet light and x-rays as well. *See* DWARF STAR; SPECTRAL TYPE; STAR; STELLAR EVOLUTION. James B. Kaler
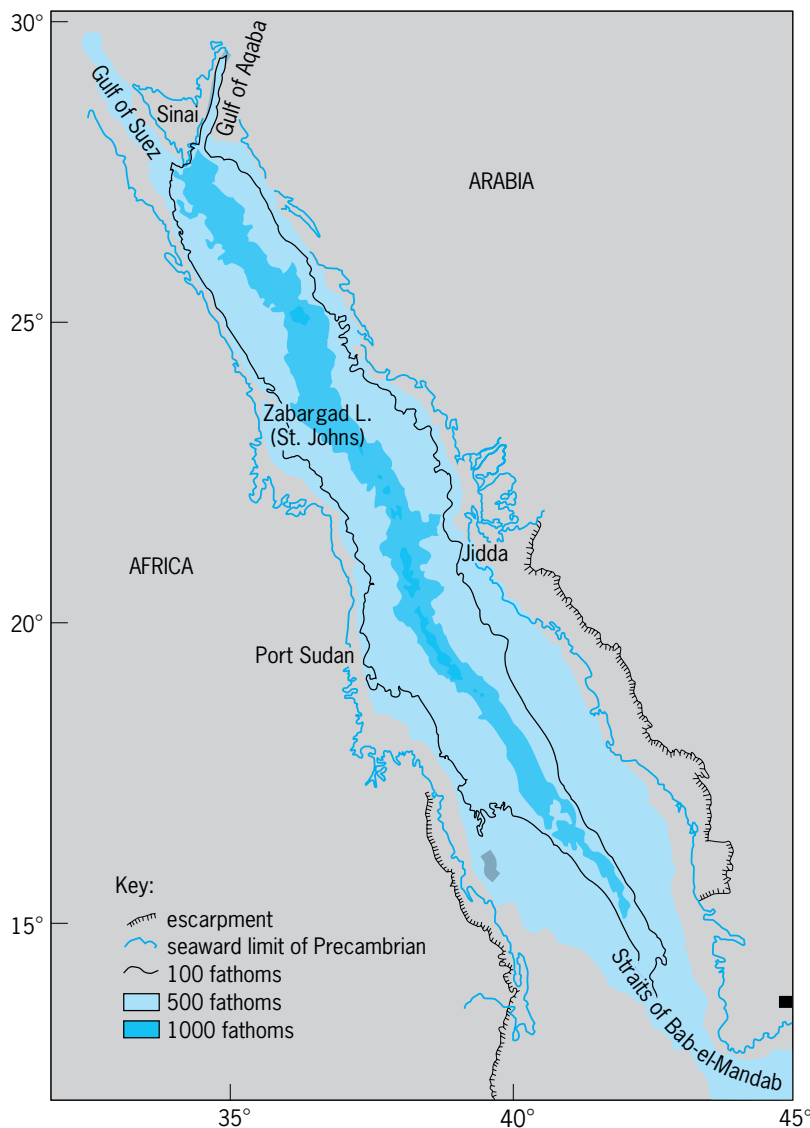
Red Sea

A body of water that separates northeastern Africa from the Arabian Peninsula. The Red Sea forms part of the African Rift System, which also includes the Gulf of Aden and a complex series of continental rifts in East Africa extending as far south as Malawi. The Red Sea extends for 1920 km (1190 mi) from Ras (Cape) Muhammed at the southern tip of the Sinai Peninsula to the Straits of Bab el Mandab at the entrance to the Gulf of Aden. At Sinai the Red Sea splits into the Gulf of Suez, which extends for an additional 300 km (180 mi) along the northwest trend of the Red Sea and the nearly northward-trending Gulf of Aqaba. The 175-km-long (109-mi) Gulf of Aqaba forms the southern end of the Levant transform, a

primarily strike-slip fault system extending north into southern Turkey. The Levant transform also includes the Dead Sea and Sea of Galilee and forms the northwestern boundary of the Arabian plate. *See* ESCARPMENT; FAULT AND FAULT STRUCTURES.

Physiography. The coastlines of the Red Sea are remarkably straight and parallel for nearly 500 km (310 mi) south of the Gulf of Suez to about 24°N with a separation of 190 km (120 mi). South of there, the coastlines become sinuous, and the sea gradually widens to a maximum width of 350 km (210 mi) between 16°N and 17°N before narrowing to about 40 km (25 mi) at the Straits of Bab el Mandab. Irregular, eroded escarpments face seaward from the uplifted rift shoulders that flank the Red Sea. The rift shoulders average 1000–3000 m (3300–9900 ft) in elevation and expose Late Precambrian granitic, metamorphic, and volcanic rocks. The total width of the rift occupied by the Red Sea increases from about 250 km (150 mi) in the north to 500 km (300 mi) in the south. *See* PRECAMBRIAN.

The Red Sea consists of narrow marginal shelves and coastal plains and a broad main trough with depths ranging from about 400 to 1200 m (1300 to 3900 ft). The main trough is generally bounded by narrow marginal shelves and coastal plains. However, south of 20°N on the eastern side and 17°N on the western side, shallow reefs and carbonate banks nearly fill the entire main trough. Carbonate reefs are found along the entire length of the Red Sea, but are confined to nearshore areas further north. From 15°N to 20°N, the main trough is bisected by a narrow (<60 km or 37 mi wide) axial trough with a very rough bottom morphology and depths of greater than 2000 m (6600 ft). The maximum recorded depth is 2920 m (9580 ft). *See* REEF.

Oceanography. Water circulation in the Red Sea is driven by monsoonal wind patterns and changes in water density due to evaporation. Evaporation in the Red Sea is sufficient to lower the sea level by over 2 m (6.6 ft) per year. No permanent rivers flow into the sea, and there is very little rainfall. As a result, there must be a net inflow of water from the Gulf of Aden to compensate for evaporative losses. During the winter monsoon, prevailing winds in the Red Sea are from the south, and there is a surface current from the Gulf of Aden into the Red Sea. Evaporation increases the salinity of surface water, causing it to become more dense and to sink, particularly in the north where surface temperatures are lower. This dense water spreads southward along the bottom of the Red Sea. Some Red Sea deep water spills over the sill in the Straits of Bab el Mandab and can be traced as a distinct water mass at intermediate depths through the Gulf of Aden and into the Indian Ocean. During the summer monsoon, the wind in the Red Sea blows strongly from the north, causing a surface current out of the Red Sea. Gulf of Aden water flows into the Red Sea under the surface current. The southerly surface flow causes upwelling in the northern Red Sea and a northward bottom current. Warm, high-salinity water then spreads to the south under the northward-flowing Gulf of Aden water, and some of

Division of the Red Sea into regions representing different stages in the formation of an ocean basin. (*After F. Martinez and J. R. Cochran, Structure and tectonics of the northern Red Sea: Catching a continental margin between rifting and drifting, Tectonophysics, 153:1–32, 1988*)

this dense water spills over into the Gulf of Aden. *See* MONSOON METEOROLOGY.

Geologic development. The rifting that has created the Red Sea began about 30 million years ago (Ma) [see **illus.**]. During the early development of the sea, the Gulf of Aqaba did not exist, and the Gulf of Suez formed the northern continuation of the Red Sea rift. The Gulf of Suez was cut off by the beginning of motion on the Levant transform in the middle Miocene (about 12–15 Ma). A total of about 140 km (87 mi) of motion has occurred between Africa and Arabia in the northern Red Sea and about 300 km (186 mi) of motion in the southern Red Sea. The present rates of motion between Africa and Arabia range from 0.75 cm/year (0.29 in./year) in the northern Red Sea to 1.5 cm/year (0.59 in./year) in the southern Red Sea.

Faunal evidence from sediments, particularly from the Gulf of Suez where there has been extensive

drilling for oil, indicates that during the early part of its development the Red Sea communicated with the Mediterranean. This connection was cut off or became extremely restricted in the middle Miocene, perhaps by uplift in Sinai related to the beginning of motion on the Levant transform. High evaporation rates caused sea level to fall rapidly, and the Red Sea became a shallow, hypersaline evaporitic basin precipitating primarily gypsum, anhydrite, and halite. The sea apparently did not completely dry out, but received either a small, steady supply or periodic influxes of water through the Gulf of Suez that maintained the salinity at a level where that mixture of evaporites was produced. These conditions persisted for perhaps as long as 10 million years, resulting in the accumulation of very thick (several kilometers) evaporitic deposits. About 5 Ma, tectonic activity in the southern Red Sea and Gulf of Aden created a connection with the Indian Ocean, ending evaporite deposition and reestablishing a more normal marine environment. *See* INDIAN OCEAN; MIOCENE; PLATE TECTONICS; SALINE EVAPORITES.

The axial trough in the southern Red Sea is associated with large-amplitude, lineated "sea-floor-spreading" magnetic anomalies resulting from the creation of new crust at a mid-ocean ridge spreading center. This interpretation is supported by the recovery of fresh oceanic basalt from the axial trough. The spreading center appears to have nucleated near 17°N at 4–5 Ma and to have propagated north and south to 19.5°N and 15°N, where the magnetic anomalies can be correlated to the geomagnetic time scale only to about 2 Ma. *See* GEOMAGNETISM; MID-OCEANIC RIDGE; PALEOMAGNETISM.

The axial trough becomes discontinuous north of 19.5°N and forms a series of isolated deeps spaced at 50–70-km (31–43-mi) intervals. These deeps are 25–40 km (16–25 mi) long, 10–25 km (6–16 mi) wide, and over 2000 m (6562 ft) deep, and are floored by oceanic basalt. They are associated with lineated sea-floor-spreading magnetic anomalies that can be identified to less than 2 Ma. The deeps are separated by intertrough zones that are shallower and broader, and are covered by several kilometers of sediment. These sediments are highly faulted and disturbed. The central Red Sea deeps appear to represent cells of organized sea-floor spreading that have not yet coalesced to form a continuous axis.

The main trough outside the axial trough has low-amplitude magnetic anomalies and appears not to have been formed by sea-floor spreading. The thick evaporitic sediments prevent investigation of the nature of the crust beneath the main trough by using seismic reflection techniques. Rifting in the southern Red Sea area has been accompanied by extensive volcanism. Large quantities of volcanic rocks are exposed onshore in Yemen, Ethiopia, and southwestern Saudi Arabia; and much of the crust along the margins of the Red Sea in these regions is made up of igneous intrusions. This pattern probably continues offshore, and it appears that extension between Africa and the Arabian Peninsula in the southern Red Sea has largely been accommodated by the injection of igneous material. These intrusions did not, however, become focused at a well-defined spreading center until 5 Ma. In contrast to the southern Red Sea, there has been very little volcanic activity accompanying the rifting in the northern half. Gravity measurements show a regular pattern of linear highs and lows that have been interpreted as resulting from a series of large rotated fault blocks. It thus appears that most of the extension in the northern Red Sea has been accommodated mechanically through faulting and thinning of the continental crust. *See* VOLCANOLOGY.

There is no evidence of organized sea-floor spreading in the northern 500 km (310 mi) of the Red Sea. The morphology of the northern Red Sea consists of a series of terraces stepping down to an axial region at a depth of about 1200 m (3937 ft) characterized by disturbed and deformed sediments. The axial region of the northern Red Sea is very similar in depth and appearance to the intertrough zones of the central Red Sea. Small deeps, generally associated with large magnetic anomalies, are spaced at roughly 60-km (36-mi) intervals along the northern Red Sea axis. These deeps are much smaller (a few kilometers across) and shallower (1400–1600 m or 4593–5250 ft deep) than the large central Red Sea deeps; and they are generally floored by thick sediments. The magnetic anomalies are all dipolar, indicating localized igneous intrusions, rather than lineated as are the sea-floor-spreading anomalies to the south. They also show that all of the intrusions are positively magnetized, implying that they were emplaced during the past 700,000 years since the last reversal of the Earth's magnetic field. It appears that the northern Red Sea deeps represent the first stage in the establishment of an oceanic spreading center within a rift.

Hot brines. Some small, enclosed basins within the central Red Sea deeps trap pockets of very dense, salty hot water. The salt content of the brine pools is greater than 25% by weight compared to normal Red Sea salinity of about 4%. Temperatures of the hot brines can be in excess of 50°C or 122°F (compared with average Red Sea bottom temperatures of 22°C or 72°F and normal deep-ocean bottom temperatures of 1–2°C (34–36°F)). The density contrast between the hot brines and the overlying seawater is great enough to cause a reflection of the sound waves used by echo sounders. The hot water results from hydrothermal circulation through the highly fractured, young oceanic crust, while the salt comes from the thick layer of evaporites. The hydrothermal circulation, in addition to heating the water, results in leaching of heavy minerals from the igneous rocks of the upper crust. These minerals become concentrated in the brine pools and precipitate out, resulting in the formation of sediments with very high concentrations of iron, manganese, zinc, lead, copper, and silver. Various schemes have been proposed to exploit these mineral deposits, but none have proved feasible. *See* AFRICA; SEAWATER.   James R. Cochran

Bibliography. R. G. Coleman, *Geologic Evolution of the Red Sea*, 1993; E. T. Degens and D. A. Ross (eds.),

Hot Brines and Recent Heavy Metal Deposits in the Red Sea, 1969; A. L. Edwards and S. M. Head (eds.), *Key Environments: Red Sea*, 1987; I. I. Gass, *Crustal and Mantle Processes: Red Sea Case Study*, 1980; X. LePichon and J. Cochran (eds.), The Gulf of Suez and Red Sea rifting, *Tectonophysics*, vol. 153, special issue, 1988; S. Locke and R. C. Thunell, Paleoceanographic record of the last glacial/interglacial cycle in the Red Sea and Gulf of Aden, *Palaeogeog. Palaeoclimatol. Palaeoecol.*, 64:163–187, 1988; F. Martinez and J. R. Cochran, Structure and tectonics of the northern Red Sea: Catching a continental margin between rifting and drifting, *Tectonophysics*, 150:1–32, 1988.

Redbeds

Clastic sediments and sedimentary rocks that are pigmented by red ferric oxide which coats grains, fills pores as cement, or is dispersed as a muddy matrix. These conspicuously colored rocks commonly constitute thick sequences of nonmarine, paralic (marginal marine), and less commonly shallow marine deposits. Clastic redbeds accumulated in many parts of the globe during the past 10^9 years of Earth history. They were among the first sedimentary deposits to be considered climatic indicators, because their color was assumed to reflect a unique, identifiable condition of deposition. On the contrary, development of red pigment in most redbeds was complex and is difficult to decipher in detail. Ferric oxides also pigment marine chert, limestone, and cherty iron formations and ooidal ironstones, but these chemical deposits are not usually included among redbeds.

Some redbeds contain abundant grains of sedimentary and low-grade metamorphic rocks and relatively few grains of iron-bearing minerals. Most of them, however, contain feldspar and relatively abundant grains of opaque black oxides derived from igneous and high-grade metamorphic source rocks. In older redbeds the black grains are predominantly specular hematite replacing magnetite; in younger ones magnetite and ilmenite are more common. Dark grains of iron-bearing silicate minerals are rather rare in red sandstones; but they are more common in early-cemented concretions and in modern analogs, pointing to their postdepositional destruction which released pigmenting ferric oxide. Clay minerals in older redbeds, as in most other ancient clastic deposits, are predominantly illite and chlorite, thus providing no specific clue to the climate in the source area or at the place of deposition. *See* ARKOSE.

In many of the younger redbeds the pigmenting ferric oxide mineral cannot be identified specifically because of its poor crystallinity. In most of the older ones, however, hematite is the pigment. As seen under the scanning electron microscope, the hematite is in the form of hexagonal crystals scattered over the surface of grains and clay mineral platelets. In red mudstones most of the pigment is associated with the clay fraction.

Redbeds do not contain significantly more total iron than nonred sedimentary rocks. Normally, iron increases with decreasing grain size of redbeds. Moreover, the amount of iron in the grain-coating pigments is small compared with that in opaque oxides, dark silicates, and clay minerals. These facts demonstrate that chemical and mineralogical data cannot differentiate redbeds formed in moist climates from those formed in deserts.

Source areas. Actively eroded source areas of most redbeds supplied relatively abundant grains of magnetite and iron-bearing silicate minerals, whether in humid or dry climates. Accordingly, the common grains of specular hematite in redbeds must have been produced by an oxidizing burial regime. Similarly, the sparse crop of dark silicates commonly rimmed with red pigment resulted from postdepositional alteration, emphasizing its role in pigmentation and as a source of some of the clay mineral matrix as well.

Regardless of climate or color of their soils, source areas generally deliver brown sediments. Accordingly, the inherited brown muddy matrix that contained amorphous to very poorly crystalline hydrated ferric oxide (ferrihydrite) must have transformed slowly to red hematite pigment after deposition. This process is favored by exposure during accumulation and by persisting oxygenating conditions of burial in nonmarine to paralic environments. Postdepositional conversion of brown hydrated ferric oxide to red pigments may have also operated in shallow- to deeper-water marine deposits. In all of these red deposits, the final color is no direct clue to the climate in the source area.

Most redbeds were made from rather ordinary alluvium that accumulated in nonmarine to shallow-marine environments, under a pervasive oxygenating condition of burial within a common range of water chemistry. None was red at the time of deposition, and for each the environment during and after accumulation was crucial for postdeposition dehydration (aging) of ferrihydrate to hematite. Criteria such as associated fauna, flora, eolian sands, evaporites, or coal measures provide the most reliable evidence about the climate.

Mode of deposition. Redbeds occur in some alluvial fans, floodplains, deltas, lakes, and deserts. The fan and braided deposits are generally uniformly red; the floodplain and delta plain deposits are red-banded; and the desert alluvial, lake, and eolian deposits are generally uniformly red. In any case, diagenesis is, in a real sense, the control of redbeds. The processes that create red hematite are dehydration of brown ferric oxide, derivation from iron silicates, derivation from ilmenite and magnetite, and replacement of iron-bearing grains by dolomite cement.

Redbeds common in late Precambrian and Phanerozoic nonmarine to paralic deposits comprise several quite different facies produced by different combinations of factors. Some accumulated in basins whose adjacent highlands had essentially the same climate; others had relatively remote source areas where the climate may have differed from that of the

basin. Sediments deposited in a hot, dry climate were derived either from adjacent arid uplands or from more distant sources that may have been moister. The most convincing analogs of these desert redbeds are forming diagenetically today in nonmarine Cenozoic deposits in the arid and semiarid regions of the southwestern United States and northwestern Mexico. Deposits of this sort are composed of sandy alluvium, dune sand, and fanglomerate with relatively little mudstone. They were pigmented by postdepositional destruction (hydrolysis) of iron-bearing grains. Examples include late Paleozoic redbeds of Colorado and eastern Canada, as well as the Permian upper Rotliegendes of central Europe and Penrith eolian sandstone of Great Britain. *See* DIAGENESIS.

Extensive, well-sorted, fine-grained redbeds derived from distant sources accumulated on evaporitic tidal flats and in saline lagoons. They were also pigmented largely by postdepositional destruction of iron-bearing minerals. Thick and extensive red claystone in some of these successions suggests derivation from distant soil-mantled uplands with a moister climate, implying a reddening of the mud by postdepositional dehydration of inherited brown pigment. This facies includes the Permian redbeds of Texas, Oklahoma, New Mexico, and Arizona; Triassic redbeds of Wyoming and South Dakota; and Jurassic redbeds of the Colorado Plateau. *See* GYPSUM; HALITE; SALINE EVAPORITES.

Sediments that accumulated in savannas were derived either from adjacent uplands with seasonally humid climate or from more distant sources that may have been even moister. Savanna redbeds were supplied with detritus that included both fresh grains of iron-bearing minerals and a soil-derived clay fraction pigmented with brown hydrated ferric oxide that dehydrated after deposition to red hematite. Commonly, savanna redbeds comprise interbedded red and drab layers and lenses. Red-banded (mottled) deposits are usually reddish brown, maroon, and lavender as well as greenish gray, yellowish gray, and yellowish orange. Most of these red-banded deposits are fining-upward fluvial sequences with lower drab channel sandstone that accumulated under reducing conditions below the water table, and upper oxidized reddish-brown overbank (floodplain) mudstone that retained its inherited free ferric oxide during deposition. In long interflood periods of exposure, paleosols commonly developed in the upper part of the floodplain and tidal-flat mud. These horizons are marked by extensive burrowing, brown calcareous nodules, and root-tube fillings. Paleosols of this sort point to subhumid to semiarid climate with seasonal rainfall and low water table, yet with sufficient moisture to maintain an immature soil profile in the slowly aggrading floodplain mud. Similarly varicolored and "violet" paleosols in red-banded deposits have been interpreted as evidence of subtropical, humid to semiarid climate. In addition, some savanna deposits bear reliable faunal and floral evidence of a warm, moist climate, as in the redbeds of the late Paleozoic coal measures of Great Britain and eastern Canada, the lower Rotliegendes of cen-

tral Europe, and the early Cenozoic deposits in the Rocky Mountain region. Other deposits are suggestive of some-what drier climate, for example, the Devonian Catskill redbeds of New York and Pennsylvania, the early Carboniferous Mauch Chunk Formation of Pennsylvania, and the Permian New Red Sandstone of Scotland. *See* PALEOSOL; SAVANNA.

Some of the red mudstones and siltstones that formed before the development of advanced land plants in mid-Paleozoic time accumulated in offshore marine environments, as in the Silurian marine redbeds of Great Britain and the United States, either on quite deep shelves or in basins of rapid sedimentation and burial. Moreover, some deepwater marine redbeds are resedimented nearshore deposits that were swept into deep basins by turbidity currents, as in the early Paleozoic of eastern Canada. *See* NEARSHORE PROCESSES; SEDIMENTARY ROCKS; TURBIDITY CURRENT.

Tectonics and continental drift. A genetic association of major redbeds with tectonic activity has long been recognized. As one tectonic end member, extensive uniform redbed-evaporite sequences accumulated on stable cratons, whereas some thick wedges of desert redbeds with border conglomerates accumulated in rift valleys. Overall, the tectonic background most commonly associated with redbeds is a late to post orogenic framework, as noted in the late Hercynian deposits of Europe and northwestern Africa, the late Alpine deposits of southern Europe, and the late Andean deposits of South America. Repeated episodes of deformation along the Appalachian belt produced redbeds at the close of three Paleozoic orogenies.

On a global scale, paleomagnetic evidence of the distribution of redbeds relative to their pole position corroborates paleogeographic data, suggesting that most redbeds, evaporites, and eolian sandstones accumulated less than 30° north and south of a paleo equator where hot, dry climate generally prevailed. But diagenetic development of red hematite may be acquired long after deposition. Moreover, continental drift reconstructions reveal that the most widespread redbeds in the geologic record developed near the Equator in late Paleozoic and early Mesozoic time when the continents were assembled in a great landmass, Pangaea. *See* CONTINENTS, EVOLUTION OF; PALEOGEOGRAPHY; PALEOMAGNETISM.   Franklyn B. Van Houten

Bibliography. R. H. Blodgett, J. P. Crabaugh, and E. F. McBride, The color of red beds: A geologic perspective, *Soil Sci. Soc. Amer. Spec. Publ.*, 31:127–159, 1993; T. M. Bown and M. J. Kraus, Lower Eocene alluvial paleosols (Willwood Formation, northwest Wyoming, U.S.A.) and their significance for paleoecology, paleoclimatology, and basin analysis, *Paleogeogr. Paleoclimat. Paleoecol.*, 34:1–30, 1986; H. Falke (ed.), *The Continental Permian in Central, West, and South Europe*, pp. 240–282, 1976; K. W. Glennie, Desert sedimentary environments, *Dev. Sedimentol.*, 14:173–193, 1970; J. C. Lorenz, *Triassic-Jurassic Rift-Basin Sedimentology*, 1988; E. F. McBride, Significance of color in red, green,

purple, olive, brown and gray beds of Difunta group, northeastern Mexico, *J. Sediment. Petrol.*, 44:760–773, 1974; J. G. McPherson, Calcrete (caliche) paleosols in fluvial redbeds of the Aztec Siltstone (Upper Devonian), southern Victoria Land, Antarctica, *Sediment. Geol.*, 22:267–285, 1979; P. Turner, *Continental Red Beds*, 1980; F. B. Van Houten, Origin of red beds: A review—1961-1972, *Annu. Rev. Earth Planet. Sci.*, 1:36-61, 1973; D. L. Woodrow, F. W. Fletcher, and W. F. Ahrnsbrah, Paleography and paleoclimate at the deposition sites of the Devonian Catskill and Old Red facies, *Geol. Soc. Amer. Bull.*, 84:3051–3064, 1973; A. M. Ziegler and W. S. McKerrow, Silurian marine red beds, *Amer. J. Sci.*, 275:31–56, 1975.

Redshift

A systematic displacement toward longer wavelengths of lines in the spectra of distant galaxies, and also of the continuous part of the spectrum. First studied systematically by E. Hubble, redshift is central to observational cosmology, in which it provides the basis for the modern picture of an expanding universe. There are two fundamental properties of redshifts.

First, the fractional redshift $\Delta\lambda/\lambda$ is independent of wavelength. ($\Delta\lambda$ is the shift in wavelength of radiation of wavelength λ.) This rule has been verified from 21 cm (radio radiation from neutral hydrogen atoms) to about 6×10^{-5} cm (the visible region of the electromagnetic spectrum) and leads to the interpretation of redshift as resulting from a recession of distant galaxies. Though this interpretation has been questioned, no other mechanism is known that would explain the observed effect.

Second, redshift is correlated with apparent magnitude in such a way that when redshift is translated into recession speed and apparent magnitude into distance, the recession speed is found to be nearly proportional to the distance. This rule was formulated by Hubble in 1929, and the constant of proportionality bears his name. Based on the results of the Hubble Key Project and the *Wilinkson Microwave Anisotropy Probe* (*WMAP*), Hubble's constant is currently estimated to lie between 67 and 75 km per second per megaparsec or 2.2 and 2.4×10^{-18} s^{-1}. *See* HUBBLE CONSTANT; MAGNITUDE (ASTRONOMY); WILKINSON MICROWAVE ANISOTROPY PROBE.

Until 1975 the largest redshifts that could be routinely measured for optical galaxies were around $\Delta\lambda/\lambda = 0.2$. Progressively higher redshifts have been measured as instruments have improved. Modern telescopes such as the Hubble Space Telescope and the Keck telescope have measured galaxy redshifts up to 6.68 and quasar redshifts up to 6.4. These objects were discovered in large surveys using photometric redshifts; this technique compares the predicted brightness of an object in various colors to the brightnesses observed, yielding an approximate redshift that can later be measured precisely using spectroscopy. *See* QUASAR; TELESCOPE.

The recession speed indicated by the redshift in the spectrum of a given galaxy is not the current value for that galaxy but the value appropriate to the epoch when the light now reaching the Earth was emitted. Consequently, the observed relation between redshift and apparent magnitude contains information about past values of Hubble's constant, as well as about the present value. If this information could be extracted from the record, it would enable astronomers to choose among various model universes that have been proposed by cosmologists. Current efforts to do this focus on type Ia supernovae, and the unexpected dimness of distant supernovae has been interpreted as evidence for acceleration; but the dimming may also plausibly be explained by supernova evolution or intergalactic dust. *See* SUPERNOVA.

Since redshifts of galaxies are nearly proportional to their distances, they can be used in conjunction with measured galaxy locations on the sky to construct very large scale three-dimensional maps of the galaxy distribution. When this was first done in 1988 for 1100 galaxies, the maps revealed structures in the galaxy distribution with sizes up to 100 megaparsecs (3×10^8 light-years). Subsequent redshift surveys (with up to 26,000 galaxies) have verified this result, and given statistical information about the large-scale distribution of cosmic matter. Redshift surveys now in progress, such as the Sloan Digital Sky Survey, will measure millions of additional galaxies. *See* COSMOLOGY; GALAXY, EXTERNAL; SLOAN DIGITAL SKY SURVEY; UNIVERSE.

David Layzer; Anthony Aguirre

Bibliography. E. J. Chaisson, *The Hubble Wars: Astrophysics Meets Astropolitics in the Two-Billion-Dollar Struggle over the Hubble Space Telescope*, Harvard University Press, 1998; P. J. E. Peebles, *Principles of Physical Cosmology*, Princeton University Press, 1993; S. Perlmutter et al., Discovery of a supernova explosion at half the age of the universe, *Nature*, 391:51, 1998; B. P. Schmidt et al., The high-Z supernova search: Measuring cosmic deceleration and global curvature of the universe using type Ia supernovae, *Astrophys. J.*, 507:46–63, 1998.

Redtop grass

One of the bent grasses, *Agrostis alba* and its relatives, which occur in cooler, more humid regions of the United States on a wide variety of soils. Redtop tolerates both wet and dry lands and acid and infertile soils, and it is used where other species of grasses do not thrive. Redtop is a perennial, spreads slowly by rootstocks, and makes a coarse, loose turf. Top growth is 2–3 ft (0.6–0.9 m) tall, with moderately leafy, wiry stems. The inflorescence is a reddish open panicle. Redtop is used for pasture and hay and is fairly nutritious if harvested promptly when heading occurs. It is effective in preventing erosion by holding banks of drainage ditches, waterways, and terrace channels. *See* CYPERALES; GRASS CROPS.

Howard B. Sprague

Redwood

A member of the pine family, *Sequoia semper-virens*, is the tallest tree in the Americas, attaining a height of 350 ft (107 m) and a diameter of 27 ft (8.2 m). Its range is limited to a strip along the Pacific Coast about 35 mi (55 km) wide and 500 mi (800 km) long, extending from southwest Oregon to about 100 mi (160 km) south of San Francisco, California. The leaves are evergreen, sharply pointed, small (0.3–1 in. or 8–25 mm long), distichous (disposed in two vertical rows) on short branches, and scalelike on the main stem (see **illus.**). The cones are egg-shaped, about 1 in.
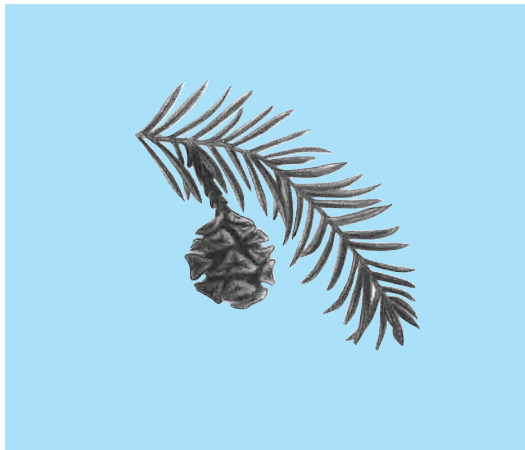
Redwood (*Sequoia sempervirens*) leaves and cones.

(25 mm) long and 0.5 in. (13 mm) broad. The bark is a dull red-brown, on old trees sometimes 1 ft (0.3 m) thick, densely fibrous, and highly resistant to fire. The tree gets its common name from the color of the bark as well as that of the heartwood. The wood is moderately light in weight and strength and low in warp and shrinkage, and is not difficult to work. The heartwood is especially valuable, being highly resistant to decay. The wood holds paint well. It is used for bridge timbers, tanks, flumes, silos, posts, shingles, paneling, doors, caskets, furniture, siding, and many other building purposes. *See* FOREST AND FORESTRY; PINALES; PINE; TREE.

Arthur H. Graves; Kenneth P. Davis

Reef

A rigid and wave-resistant marine structure that stands above its surroundings. Biologists and geologists specify that reefs are constructed by organisms that secrete calcium carbonate skeletons. To navigators, a reef is any rocky structure that poses a threat to navigation.

Coral reefs. Modern reef-builders include algae, mollusks, bryozoans, worms, and sponges, but the most important are corals.

Distribution. Coral reefs are most common in warm water (typically $\geq 20^{\circ}C$) in which calcium carbonate precipitates more easily. Most shallow-water corals are hermatypic, and they host symbiotic zooxanthellae (algal-like dinoflagellates) that provide metabolites through photosynthesis. This provides energy for the coral while removing carbon dioxide that inhibits calcification.

While coral reefs are more common in the shallow tropics, they can be found in both deeper and colder water, where they do not have zooanthellae. According to the National Oceanic and Atmospheric Administration, reefs dominated by the coral *Lophelia pertusa* occur off the coast of Florida at depths of 500–850 m; similar features built by *Oculina varicosa* are found at water depths of 70–100 m in the Gulf of Mexico and the Straits of Florida. Off the coast of Norway, *Lophelia* reefs occur at depths of 75–155 m, at temperatures ranging from 4 to 13°C. Coral diversity is much lower in these deep-water reefs, but they can harbor a variety of organisms. Between 350 and 400 species of mollusks, crustaceans, and other fauna have been reportedly found within the branches of *Oculina* off the eastern coast of Florida. *See* SCLERACTINIA.

Living corals and reefs. Most corals are colonial, with each individual polyp measuring from several millimeters to a few centimeters across (**Fig. 1***a*). Each polyp secretes an external calcium carbonate skeleton and sits within a cuplike calyx (Fig. 1*b*) that is part of the larger colony (Fig. 1*c*). At night, the coral polap extends its tentacles to feed. During the day, the coral retreats into its "skeletal cup" and relies on zooxanthellae for nutrition.

Reef types. Generally, coral reefs can be divided into three categories: fringing, barrier, and atoll. This classification is based on Darwin's original observations of mid-Pacific reefs.

Volcanoes form as a moving oceanic plate passes over a rising plume of hot mantle (**Fig. 2**). As the sea floor moves away from this hot spot, the volcanic lava (basalt) cools and subsides. Coral reefs initially form along the steep slopes of a new volcano in shallow waters close to shore. As the volcano subsides (moves away and sinks), reefs build vertically to stay close to sea level and the light needed for photosynthesis; the lagoon behind the reef progressively widens, and the influence of land decreases. Fringing reefs, which are closest to shore, occur around newly formed and thus high volcanic islands. As the volcanic island slowly subsides, reefs transition to barriers, and eventually to atolls, once the volcano has sunk from sight beneath the surface waters. *See* ATOLL; OCEANIC ISLANDS; VOLCANO.

Coral zonation. The species and shapes of corals on a reef are controlled primarily by light availability, wave energy, and sedimentation, resulting in predictable reef zones (**Fig. 3**). Branching corals, such as *Acropora palmata* in the Caribbean Sea (Fig. 1*c*; background), occur along the reef crest, where abundant light drives the photosynthesis needed to support rapid growth. Vigorous wave energy clears sediment from the branching species that have no
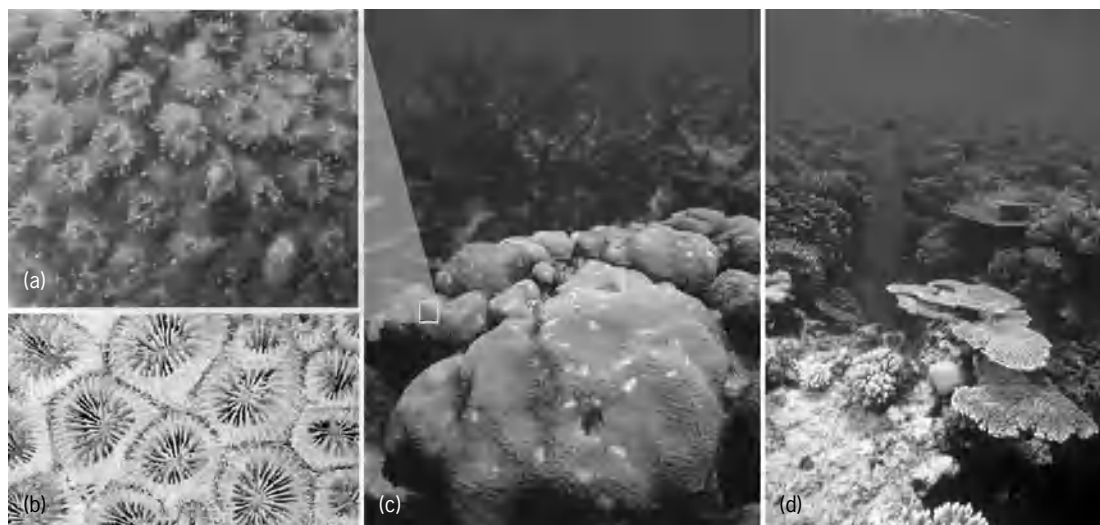
Fig. 1. Photographs illustrating the relationship between (*a*) individual coral polyps, (*b*) the skeletal structures they form, and (*c*) the colony in which they live. The hemispherical colony of *Montastraea faveolata* (*c*; foreground) sits on a sloping Caribbean forereef in 10 m of water. Branching colonies of *Acropora palmata* (*c*; background) occupy shallower water near the reef crest. (*d*) The branching/tablelike corals near the crest are responding to similar conditions and give way to more massive forms in deeper water along the Great Barrier Reef of Australia.

biological means of sediment removal. Although the species of branching and table-like corals vary by ocean, these types of corals usually occur near the reef crest, in response to similar processes (Fig. 1*d*). Further down the reef front, slower-growing hemispherical corals (Fig. 1*c*, foreground; Mixed/Massive Zone in **Fig. 3**) are better suited for removing sedi-

ment coming down the slope. In even deeper water, platelike colonies (Fig. 3) respond to continually decreasing light levels. This shape places all the polyps on upward-facing surfaces, optimizing the colony's ability to gather light—much like solar panels.

Reef building. Reefs have traditionally been thought of as structures built by intact and interlocking corals
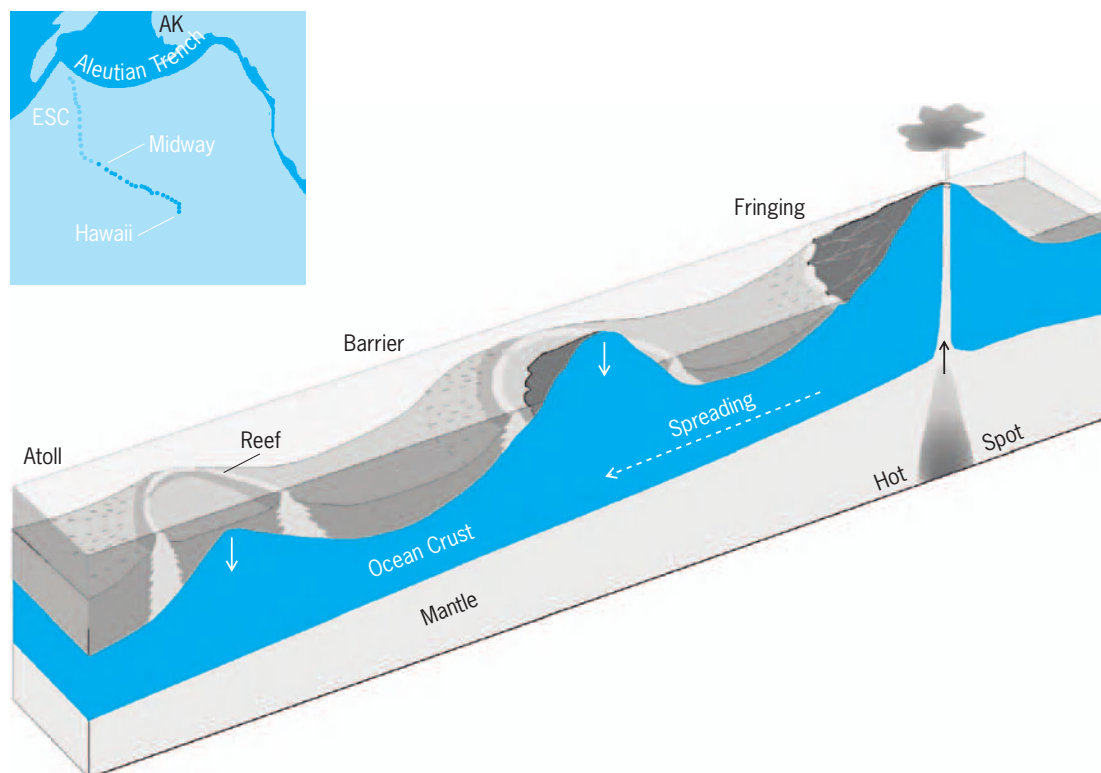
Fig. 2. Diagram of the evolution of fringing reefs into barrier reefs and atolls as subsiding volcanoes move away from a hot spot. The inset shows the trend of volcanic islands (dark color) and seamounts (sunken volcanoes, light color) associated with the hot spot near Hawaii. The volcanoes get older away from Hawaii.
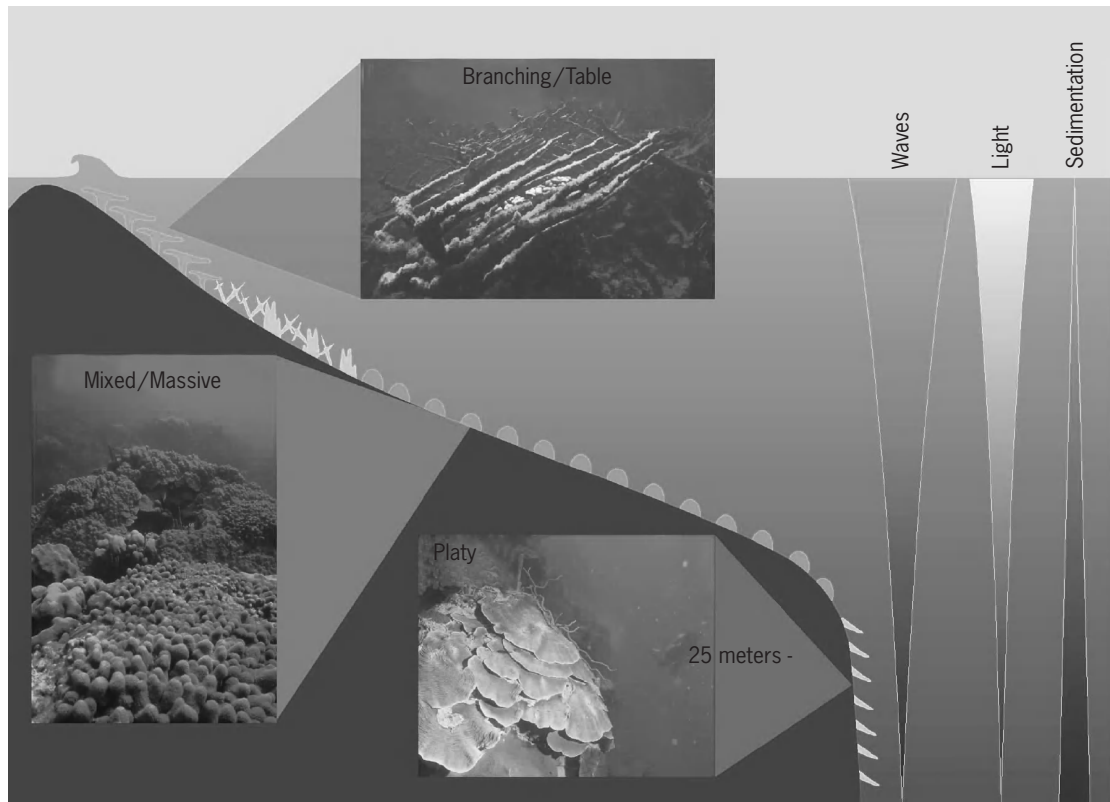
Fig. 3. Zonation on a typical Caribbean reef. The shift from branching corals to mixed/massive and platy corals down the reef front is a response to decreasing light and wave energy, accompanied by increasing sedimentation with depth.
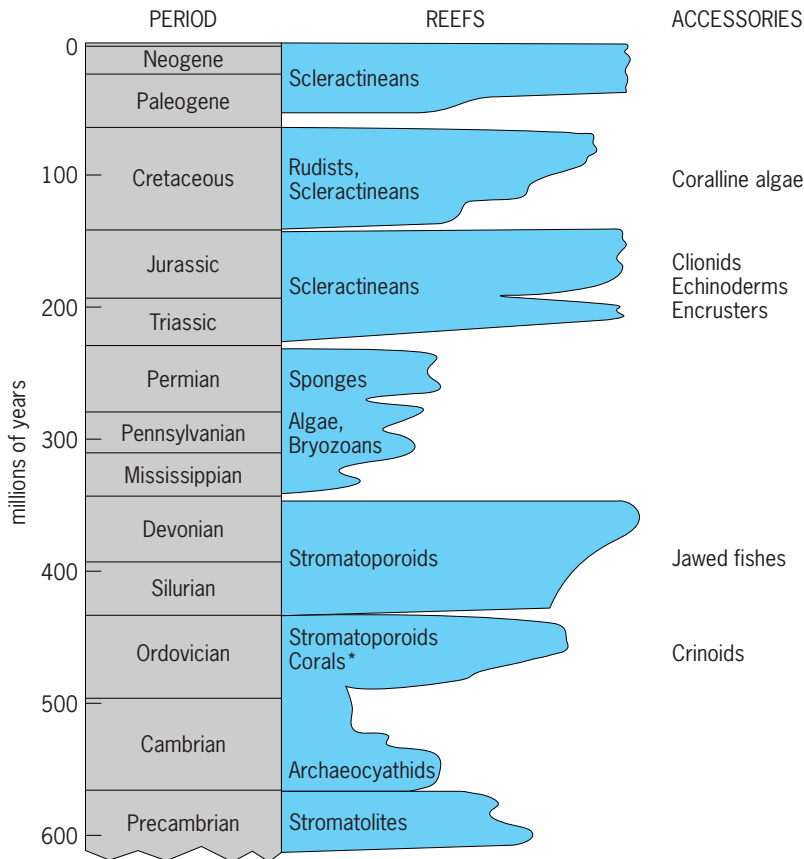
*Mid-Ordovician rugose and tabulate corals

Fig. 4. Major reef-building episodes (Reefs column) through geologic time. The major organisms responsible are listed. The origin of other important organisms that were associated with these reefs are also indicated (ACCESSORIES column).

growing atop one another over the course of hundreds to thousands of years (6000–12000 years for modern reefs). However, recent coring studies have recognized that much of the reef interior consists of toppled and degraded corals, sediment, and open cavities. Most of the sediment is produced by (1) fish and urchins grazing on algae that cover dead coral surfaces and (2) sponges, mollusks and worms that excavate into the carbonate substrate seeking shelter (a process known as bioerosion). This nearly equal mix of intact and broken corals plus loose sediment is subsequently bound together by encrusting organisms (for example, coralline algae) and carbonate cement. Thus, the edifice that is the coral reef owes its final structure as much to bioerosion, cementation, and encrustation as it does to skeletal production by corals. *See* ALGAE; CORAL-LINALES.

Reefs through time. Reefs have not always been built by corals. Microbial stromatolites formed nonskeletal mounds approximately 3 billion years ago (**Fig. 4**). Photosynthesis by these widespread reef-builders provided the oxygen needed for the evolution of complex organisms. Over the ensuing 550 million years, the evolution of reef-building organisms was controlled by changes in ocean chemistry (which affected calcification), changes in sea level (which increased or decreased space for colonization), increases in habitat variability, and increases in competition and predation. *See* PALEOE-COLOGY.

In the mid-Cambrian period (543–490 million years ago), archaeocyathids, followed by early

corals (rugose and tabulate) and stromatoporoids (calcifying sponges), may have evolved due to increased grazing pressure (Fig. 4). Early coral evolution paralleled the rapid evolution of herbivorous fish, particularly those with jaws in the Devonian (417–354 million years ago). The diversification of photosynthesizing dinoflagellates may have coincided with the radiation of both grazing and boring species from Triassic through Cretaceous time (approximately 250–65 million years ago). In this regard, increases in grazing/boring versus calcification resembled an "arms race" between organisms that built reefs and those that bioeroded them for food or shelter. *See* ARCHAEOCYATHA; RUGOSA; STROMATOPOROIDEA; TABULATA.

Reefs under threat. Dramatic changes have occurred in reefs over the past three decades. In the mid-1970s, coral cover commonly exceeded 50–75%. Since then, coral abundance has plummeted on many reefs, especially those closer to population centers. The low nutrient levels characteristic of healthy reefs rose in response to increased septic waste, agricultural fertilizers, and sedimentation. Macroalgae (seaweeds) that were previously limited by low nutrients suddenly took over. Sedimentation from dredging and deforestation not only blocked sunlight needed for photosynthesis, it also eliminated the hard, stable substrate needed for coral recruitment. The ensuing shift from coral-dominated reefs to dense macroalgal communities was exacerbated by overfishing (especially of grazers that kept algae populations balance) in and the loss of other key species, such as the long-spined sea urchin (*Diadema altillarum*) in 1983 from disease. Even the deep reefs described earlier are under threat, mostly by breakage related to trawling for deep-sea fishes.

Recently, global factors have added to the more direct assaults on reef habitat and the organisms that keep the benthos in balance. Rising temperatures have caused corals to expel their zooxanthellae, exposing the bare skeleton through the remaining clear coral tissue (bleaching). If temperatures remain elevated for more than 2–3 weeks, the corals die. Starting with white band disease in the late 1970s, corals have been assaulted by episodes of tissue degeneration. Arguments abound over the specific causes of recent declines, but the scientific community is unanimous in its assessment of likely anthropogenic ties for the global degradation of coral reefs.   Dennis Hubbard

Bibliography. D. K. Hubbard, Dynamic processes of coral-reef development, in C. Birkeland (ed.), *Life and Death of Coral Reefs*, Chapman and Hall, 1997; T. P. Hughes, Catastrophes, phase shifts and large-scale degradation of a Caribbean coral reef, *Science*, 265:1547–1551, 1994; J. Pandolfi et al., Are U.S. coral reefs on the slippery slope to slime?, *Science*, 307:1725–1726, 2005; L. L. Richardson, Coral diseases: What is really known?, *Trends Ecol. Evolut.*, 13:438–443, 1998; G. D. Stanley, *The History and Sedimentology of Ancient Reef Systems*, Klewer Academic/Plenum Press, New York, 2001.

Reengineering

The application of technology and management science to the modification of existing systems, organizations, processes, and products in order to make them more effective, efficient, and responsive. Responsiveness is a critical need for organizations in industry and elsewhere. It involves providing products and services of demonstrable value to customers, and thereby to those individuals who have a stake in the success of the organization. Reengineering can be carried out at the level of the organization, at the level of organizational processes, or at the level of the products and services that support an organization's activities. The entity to be reengineered can be systems management, process, product, or some combination. In each case, reengineering involves a basic three-phase systems-engineering life cycle comprising definition, development, and deployment of the entity to be reengineered.

Systems-management reengineering. At the level of systems management, reengineering is directed at potential change in all business or organizational processes, including the systems acquisition process life cycle itself. Systems-management reengineering may be defined as the examination, study, capture, and modification of the internal mechanisms or functionality of existing system-management processes and practices in an organization in order to reconstitute them in a new form and with new features, often to take advantage of newly emerged organizational competitiveness requirements, but without changing the inherent purpose of the organization itself.

Process reengineering. Reengineering can also be considered at the levels of an organizational process. Process reengineering is the examination, study, capture, and modification of the internal mechanisms or functionality of an existing process or systems-engineering life cycle, in order to reconstitute it in a new form and with new functional and non-functional features, often to take advantage of newly emerged or desired organizational or technological capabilities, but without changing the inherent purpose of the process that is being reengineered.

At the level of processes alone, the reengineering effort would be almost totally internal. It would consist of modifications to whatever standard life-cycle processes are in use in a given organization in order to better accommodate new and emerging technologies or new customer requirements for a system. For example, an explicit risk-management capability might be incorporated at several different phases of a given life cycle and accommodated by a revised configuration of management processes. Alternatively, the work-flow processes that lead to services could be reengineered. This capability could be implemented into the processes for research, development, test, and evaluation (RDT&E); acquisition or procurement; and systems planning and marketing. Basically, reengineering at the level of processes would consist of the determination, or synthesis, of an efficacious process for ultimately fielding a product on the basis of a knowledge of generic

customer requirements, and with appropriate consideration for the objectives and critical capabilities of the systems-engineering organization. *See* RISK ASSESSMENT AND MANAGEMENT.

However, reengineering of processes alone, without attention to reengineering at a higher level, may, in many instances, represent an incomplete and not fully satisfactory way to improve organizational capabilities. Thus, the processes considered as candidates for reengineering should be high-level managerial processes as well as operational processes. Information technology is a major enabling catalyst for process reengineering. The benefits of using a process reengineering include shorter development cycles, fewer engineering change orders, products that fulfill customer expectations, and reduced program and product development costs throughout the life cycle. Thus, the process improvement results ultimately in an increase in effectiveness of product for the same cost, a reduction in cost for the same effectiveness, or some blend of these two.

Product reengineering. The term "reengineering" could mean some sort of reworking or retrofit of an already engineered product, and could be interpreted as maintenance or refurbishment. Reengineering could also be interpreted as reverse engineering, in which the characteristics of an already engineered product are identified, such that the product can perhaps be modified or reused. Inherent in these notions are two major facets of reengineering: it improves the product or system delivered to the user for enhanced reliability or maintainability, or to meet a newly evolving need of the system users; and it increases understanding of the system or product itself. This interpretation of reengineering is almost totally product-focused.

Thus, product reengineering may be redefined as the examination, study, capture, and modification of the internal mechanisms or functionality of an existing system or product in order to reconstitute it in a new form and with new features, often to take advantage of newly emerged technologies, but without major change to the inherent functionality and purpose of the system. This definition indicates that product reengineering is basically structural reengineering with, at most, minor changes in purpose and functionality of the product. This reengineered product could be integrated with other products having rather different functionality than was the case in the initial deployment. Thus, reengineered products could be used, together with this augmentation, to provide new functionality and serve new purposes. There are a number of synonyms for product reengineering, including renewal, refurbishing, rework, repair, maintenance, modernization, reuse, redevelopment, and retrofit.

A specific example of a product reengineering effort is that of taking a legacy system written in Cobol or Fortran, reverse-engineering it to determine the system definition, and then reengineering it in C^{++}. Depending upon whether any modified user requirements are to be incorporated into the reengineered product, the product is reengineered through a forward engineering effort, either just after reverse engineering has determined the initial development (technical) system specifications, or after reverse engineering has been carried out far enough to determine user requirements and user specifications, and these have then been updated. This reverse engineering concept, in which salient aspects of user requirements or technological specifications are recovered from examination of characteristics of the product, predates product reengineering. *See* PROGRAMMING LANGUAGES.

In the late 1990s, much product reengineering was focused on eliminating the potential Y2K difficulty that might ensue when older software was used that assumed that all years could be expressed by $19xy$, where x and y varied between 0 and 9. The year 2000 could not be accommodated in such a representation, and so calculations made using the xy number might be in serious error. Often, this software was written many years ago, and full documentation of the software might be lacking.

Taxonomy and life-cycle phases. Much of product reengineering is very closely associated with reverse engineering to recover either design specifications or user requirements. Then follows refinement of these requirements or specifications and forward engineering to achieve an improved product (see **illus.**). Forward engineering is the original process of defining, developing, and deploying a product, or realizing a system concept as a product; whereas reverse engineering, sometimes called inverse engineering, is the process though which a given system or product is examined in order to identify or specify the definition of the product either at the level of technological design specifications or at system- or user-level requirements. Reverse engineering may be broken down into redocumentation and design recovery.

Redocumentation is a subset of reverse engineering in which a representation of the subject system or product is recreated for the purpose of generating functional explanations of original system behavior
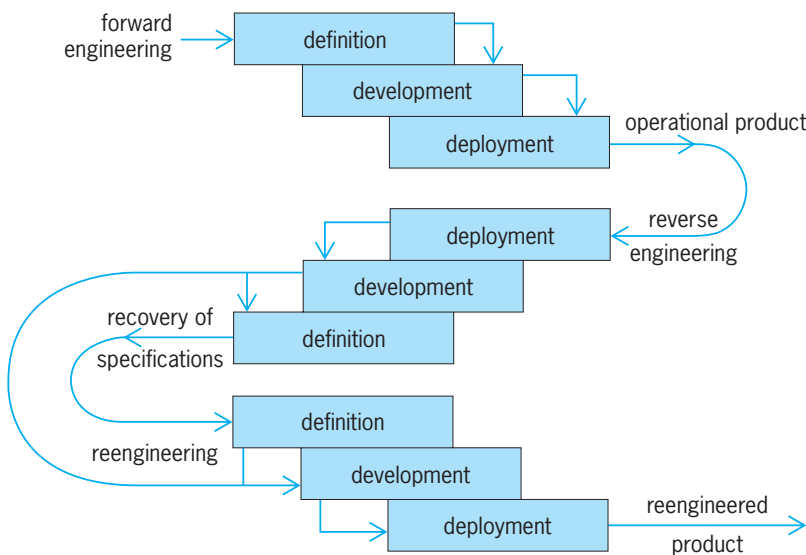
Product reengineering as a sequence of forward, reverse, and forward engineering.

and, perhaps more importantly, to aid the reverse engineering team in better understanding the system at both a functional and a structural level. A number of redocumentation tools are available for software. One major purpose of redocumentation is producing new documentation for an existing product where the existing documentation is faulty, or perhaps virtually absent.

Design recovery is a subset of reverse engineering in which the redocumentation knowledge is combined with other efforts, often involving others' experiences with and knowledge of the system, that lead to functional abstractions and enhanced product or system understanding at the level of function, structure, and even purpose.

Restructuring involves transformation of the reverse engineering information concerning the original system structure into another representation form. Restructuring generally preserves the initial functionality of the original system, or modifies it slightly in a purposeful manner that is in accord with the user requirements for the reengineered system and the way in which they differ from the requirements for the initial system.

Objectives. There are four major objectives for product reengineering.

1. *Improved maintainability*. Often, it is not easy to justify reengineering on the basis of improving maintainability because of the lack of appropriate maintenance data. *See* RELIABILITY, AVAILABILITY, AND MAINTAINABILITY.

2. *Migration to a new platform*. It may be necessary to move the system, especially software, to a better operational environment, often because of improvements in technology, such as the development of a better microprocessor or a new operating system.

3. *Greater reliability*. There may be reliability problems associated with the old system, and a reengineered version may be considerably more reliable. Restructuring reveals most such potential defects.

4. *Functional integration*. Reengineering the system into a collection of subsystems may improve the structure of these systems and thereby make it easier to change or add functionality through the addition of new subsystems without affecting existing ones.

These are not mutually exclusive objectives, and a reengineered product may gain performance in accordance with each of them.

Justification. Often, it is not easy to justify reengineering an existing product. A reengineering judgment is often reached as a last alternative after it turns out not to be possible to engineer a new product from the beginning, to purchase an available commercial product, or to continue to maintain the old one. A program of systematic measurements that can demonstrate improved reliability, maintainability, functionality, and organizational value is needed. Generally, a number of concerns must be addressed in order to demonstrate the cost effectiveness, or lack thereof, of a product reengineering effort. It is

necessary to determine the operating costs, maintenance costs, business value, and estimated future life of the current system; and the costs, estimated time to completion, and associated risks of the reengineering process. For the reengineered system, the operating costs, the business value and maintenance costs after reengineering, and the system life must be estimated.

Planning. Planning for product reengineering is essential. Any planning effort must involve the definition phase, the development phase, and the deployment phase. In term, the definition phase involves formulation of the reengineering issue such as to determine the need for, and requirements to be satisfied by, the reengineered product; identification of potential alternative candidates for reengineering; analysis of the alternatives to enable determination of costs and benefits of the various alternatives; and interpretation and selection of a preferred plan for reengineering. In the development phase, the detailed specifications for implementation of the reengineering plan are determined; and in the deployment phase, operational plans, including contracting, are set forth to enable reengineering of the product in a cost-effective and trustworthy manner.

Needs. Several needs must be considered if a product reengineering process is to yield appropriate and useful results:

1. Long-term organizational and technological issues must be examined in developing a product reengineering strategy.

2. Human, leadership, and cultural issues must be considered, including how they will be affected by the development and deployment of a reengineered product, as a part of the definition of the specifications for the product.

3. It must be possible to demonstrate that the reengineering process and product are, or will be, cost-effective and high-quality, and that they support continued evolution of future capabilities.

4. Reengineered products must be considered within an enhanced framework that also considers the potential need for reengineering at the levels of systems management and organizational processes, since technological fixes alone will generally not be sufficient to resolve organizational difficulties at these levels.

5. In particular, product reengineering for improved postdeployment maintainability must consider maintainability at the level of process rather than at the level of product only, such as would result in the case of software through rewriting source-code statements. Use of model-based management systems or code generators should yield much greater productivity, in this connection, than rewriting source code.

6. Product reengineering must consider the need for reintegration of the reengineered product with existing legacy systems that have not been reengineered.

7. Increased conformance to standards must be a result of the product reengineering process.

8. Product reengineering must consider legal issues associated with reverse engineering.

Relation to product reuse. Product reengineering is closely related to product reuse, especially for software. The reengineering of legacy software and the reuse-based production of new software are closely related concepts. The cost of developing software for one or a few applications is approximately the same as the cost of developing domain reuse components and reengineering approaches to legacy software.

Risks. Reengineering is accompanied by a variety of risks that are associated with processes, people, tools, strategies, and the application area for reengineering. (1) Integration risk is associated with having a reengineered product that cannot be satisfactorily integrated with, or interfaced to, existing legacy systems. (2) Maintenance improvement risk is the possibility that the reengineered product will exacerbate, rather than ameliorate, maintenance difficulties. (3) Systems management risk arises when the reengineered product attempts to impose a technological fix on a situation where the major difficulties call not for greater support but for organizational reengineering at the level of systems management. (4) Process risk is associated with having a reengineered product that might well represent an improvement in a situation where the specific organizational process in which the reengineered process is to be used is defective and in need of reengineering. (5) Cost risk is associated with having major cost overruns in order to obtain a deployed reengineered product that meets specifications. (6) Schedule risk is associated with having schedule delays in order to obtain such a product. (7) Human acceptance risk is associated with obtaining a reengineered product that is not suitable for human interaction, or one that is unacceptable to the user organization for other reasons. (8) Application supportability risk is associated with having a reengineered product that does not really support the application or purpose it was intended to support. (9) Tool and method availability risk is associated with proceeding with reengineering a product based upon promises for a method or tool, needed to complete the effort, which does not become available or which is faulty. (10) Leadership, strategy, and culture risk is associated with imposing a technological fix in the form of a reengineered product in an organizational environment that cannot adapt to this product. These risks are not mutually exclusive, the risk attributes are not independent, and there are additional risks, such as legal risks associated with ownership issues. *See* INFORMATION SYSTEMS ENGINEERING; SOFTWARE ENGINEERING; SYSTEMS ENGINEERING.   Andrew P. Sage

Bibliography. M. Hammer, *Beyond Reengineering: How the Process-Centered Organization Is Changing Our Work and Our Lives*, Harper Business, New York, 1996; S. Hannaford, *Workflow Reengineering*, Hayden Publishing, 1996; M. A. Ould, *Business Processes: Modeling and Analysis for Re-engineering and Improvement*, John Wiley, New York, 1995; A. P. Sage, *Systems Management for Information Technology and Software Engineering*, Wiley, New York, 1995; A. P. Sage and W. B. Rouse (eds.), *Handbook of Systems Engineering and Management*, Wiley, New York, 1999.

Reference electrode

An electrode with an invariant potential. In electrochemical methods, where it is necessary to observe, measure, or control the potential of another electrode (denoted indicator, test, or working electrode), it is necessary to use a reference electrode, which maintains a potential that is practically unchanged during the course of an electrochemical measurement. Potentials of indicator or working electrodes are measured or expressed relative to reference electrodes. *See* POTENTIALS.

Hydrogen electrode. One such electrode, the normal hydrogen electrode, has been chosen as a reference standard, relative to which potentials of other electrodes and those of oxidation-reduction couples are often expressed. A hydrogen electrode consists of a platinum foil, the surface of which is covered with finely divided platinum (platinum black). This is achieved by electroplating spongelike platinum onto the surface of the foil from a solution of chloroplatinic acid. The surface of the electrode is saturated with hydrogen (H_2) gas that is bubbled into the test solution (**Fig. 1***a*). Alternatively, the hydrogen gas (saturated with water vapor) can be bubbled continuously over the surface of the platinum (Fig. 1*b*). Electrical contact to a potentiometer can be made through the mercury in the glass tube to which the platinum foil is attached, and the metal leads are immersed into mercury.
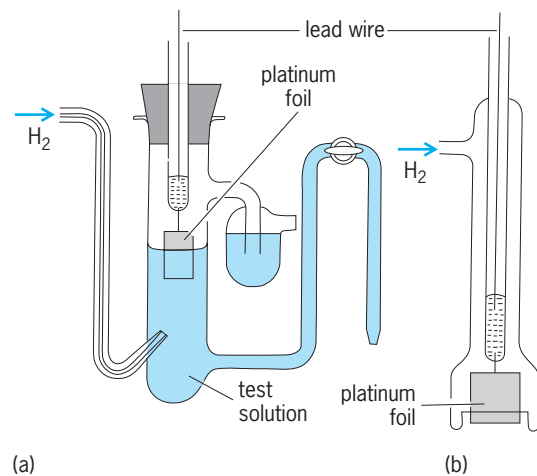

Fig. 1. Hydrogen electrodes. (*a*) Stationary bubbling type. (*b*) Dipping type.

When the hydrogen electrode is immersed in a solution containing hydrogen ions, such ions accept electrons from the electrode and form hydrogen atoms [reaction (1)], which then combine to form hy-

$$H^+ + e \rightleftharpoons H \tag{1}$$

drogen molecules [reaction (2)]. In the presence of

$$2H \rightleftharpoons H_2 \qquad (2)$$

platinum black as a catalyst, both of these equilibria are rapidly established, forming a reversible system. The potential (E) of a hydrogen electrode depends on the pressure of hydrogen gas (p_{H2}) and on activity of hydrogen ions (a_{H^+}) in the solution. The potential is given by Eq. (3), where R is the gas constant, T

$$E = -\frac{RT}{F} \ln a_{H^+} + \frac{RT}{2F} \ln p_{H_2} \qquad (3)$$

the absolute temperature, and F the Faraday. At 25°C (77°F) this equation has the form of Eq. (4).

$$E = -0.0591 \log a_{H^+} + 0.0296 \log p_{H_2} \qquad (4)$$

Hence, at a constant hydrogen pressure, a tenfold change in hydrogen-ion activity yields a change of 0.0591 V in the potential of the hydrogen electrode. Alternatively, at constant hydrogen-ion activity, a tenfold change in the pressure of the hydrogen gas results in a change of 0.0296 V in the potential of the hydrogen electrode. *See* ACTIVITY (THERMODYNAMICS).

Measurement of the potential of the hydrogen electrode can be used for determination of the pressure of hydrogen gas; however, such applications are rare. On the other hand, by maintaining a constant pressure of hydrogen gas the potential of a hydrogen electrode can be used for determination of the activity of hydrogen ions in the tested solution. Achievement of a well-functioning hydrogen electrode is difficult, and some components of the test solution can affect its function. Thus, in practice the determination of the hydrogen-ion activity (pH) is performed by using a glass electrode. The hydrogen electrode itself is used only in fundamental studies and some nonaqueous solutions. *See* PH.

The hydrogen electrode, however, remains important for providing a reference standard. For this purpose, the normal hydrogen electrode standard is achieved with a hydrogen electrode immersed in a test solution in which the hydrogen-ion activity is unity and the hydrogen gas pressure is 1 atm (100 kilopascals). The potential of this electrode, the normal hydrogen electrode, is conventionally assigned a value of 0 V at all temperatures. All electrodes with positive potential relative to the normal hydrogen electrode have greater reducing power than hydrogen; conversely, those of negative potential have lower reducing power. Such comparisons enable arrangement of the chemical elements into a series according to their reducing power (electrochemical series). *See* ELECTROCHEMICAL SERIES; OXIDATION-REDUCTION.

Secondary reference electrodes. In practice, potentials are measured against reference electrodes that are easier to work with than the normal hydrogen electrode. Such electrodes are known as secondary reference electrodes; the most common are the calomel and silver–silver chloride electrodes.

Calomel electrode. This consists of a small pool of mercury covered with a layer of a paste formed by dispersing a small amount of metallic mercury in mercury(I) chloride. The electrode is placed in an aqueous solution of potassium chloride of known concentration. Most frequently it is a saturated solution, and such an electrode is known as a saturated calomel electrode (SCE). Sometimes calomel electrodes contain either a 1.0 M (molar) or a 0.1 M solution of potassium chloride. The potential of the saturated calomel electrode measured against the normal hydrogen electrode is given by Eq. (5), where t is the

$$E = -0.2440 + 0.00076(t - 25°C) \quad \text{volts} \qquad (5)$$

temperature in degrees Celsius.

When the calomel electrode is used in measurement of pH and potentials of other electrodes, in polarography, and in voltammetry, a reservoir for the solution of potassium chloride may be provided on the side arm, which can serve as a salt bridge. The reservoir enables flushing and replacing the solution in the side arm with a fresh potassium chloride solution (**Fig. 2**).

Silver–silver chloride electrode. This consists of silver wire or foil covered with a layer of silver chloride, which is usually produced by electrolysis. The silver wire or foil is made an anode (positive electrode) in a solution containing chloride ions. Silver may also
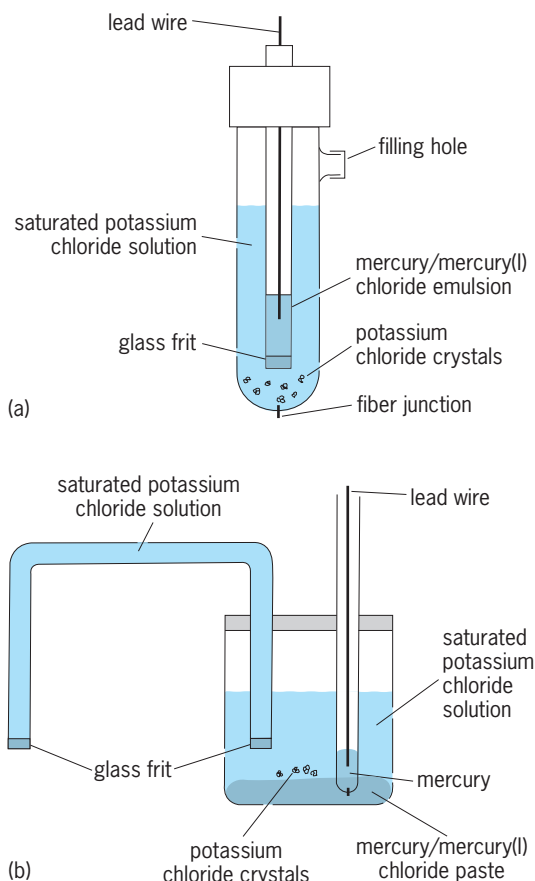

Fig. 2. Calomel electrodes. (*a*) "Upside-down" type, the type most used commercially. (*b*) Two liquid-junction potential types.

be deposited at first on a noble metal by reduction of silver oxide or silver oxalate by heating to 500°C (932°F). This process is then followed by electrodeposition of silver chloride. Alternatively, an intimate mixture of silver oxide and silver chlorate is pasted on platinum and heated to 500°C.

The potential of the silver–silver chloride electrode is defined only in a solution of known concentration of chloride ions. It is a logarithmic function of chloride-ion concentration, and measurement of the potential of such an electrode can be used for determination of chloride-ion concentration. When the silver–silver chloride electrode is used as a reference electrode, the concentration of the chloride ions must be known and stated. In 1 *M* potassium chloride solution at 25°C (77°F), the potential of this electrode is −0.222 V relative to the normal hydrogen electrode. Silver–silver chloride electrodes are used in potentiometric measurements (particularly in connection with ion selective electrodes), polarography, and voltammetry. *See* ION-SELECTIVE MEMBRANES AND ELECTRODES; POLAROGRAPHIC ANALYSIS.

In nonaqueous systems, an electrode consisting of a silver wire or foil in a solution of silver nitrate is often used as a reference. Such an electrode is suitable for comparison of standard, half-wave or peak potentials of various compounds in a given solvent. Nevertheless, the use of such reference electrodes containing one solvent (including water) for comparison of potentials of a given compound in different solvents is difficult. At the interface between the solvent in the reference electrode and the solvent in the studied solution, a liquid junction potential is established. No simple means for measuring this potential are available.

To avoid this complication, a different approach is used, involving adoption of a reference system that is different from the standard hydrogen electrode used in aqueous solutions. These reference systems are oxidation-reduction couples, the potential of which can be assumed to be only slightly sensitive to the composition of the solvent. Ferrocene-ferricinium ion, cobaltocene ion, and bis(diphenyl) chromium are the most frequently used systems. A platinum electrode immersed in the solution containing equal concentrations of the oxidized and reduced forms in the given solvent can serve as a reference electrode.

Alternatively, one of these systems can be added to the investigated solution and the potential of the studied compound measured relative to added system; this is known as an internal standard. Such an approach can be used even in aqueous solutions in voltammetric and polarographic studies. Reduction of thallium(I) ions is independent of composition of the solution in most common electrolytes. Potential of thallium(I) ion is then used as an internal standard, and potentials of studied substances are measured relative to it. *See* ELECTRODE; SOLVENT.   Petr Zuman

Bibliography. V. S. Bagotzky, *Fundamentals of Electrochemistry*, 1993; A. J. Bard, R. Parsons, and J. Jordan (eds.), *Standard Potentials in Aqueous Solutions*, 1985; J. Koryta, J. Dvořák, and L. Kavan, *Principles of Electrochemistry*, 1993; P. H. Rieger, *Electrochemistry*, 2d ed., 1993.

Reflecting microscope

A microscope whose objective is composed of two mirrors, one convex and the other concave. The imaging properties are independent of the wavelength of light, and this freedom from chromatic aberration allows the objective to be used even for infrared and ultraviolet radiation. Although the reflecting microscope is simple in appearance, the construction tolerances are so small and so difficult to achieve that the system is used only when refracting objectives are unsuitable. The distance from the objective to the specimen can be made very large; this large working distance is useful in special applications, such as examining objects situated within metallurgical furnaces. Reflecting microscopes have been mainly used for microspectrometry in the infrared and the ultraviolet, and for ultraviolet microphotography.

Optical system. The concave mirror collects light from the specimen under examination. The convex mirror intercepts this light just above the point where the rays would be focused by the concave mirror, and redirects the light up the microscope tube to the plane where the primary image is formed (see **illus.**). The curvatures and spacing of the mirrors can be so chosen that the optical aberrations
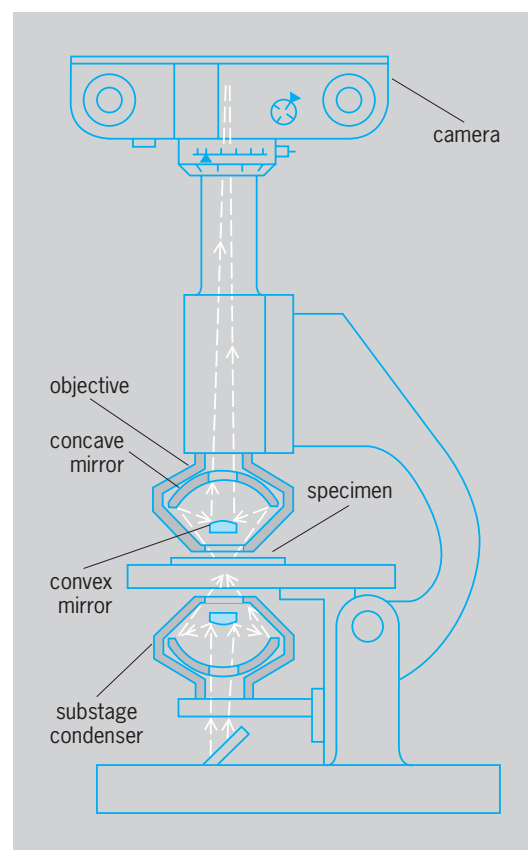
Reflecting microscope arranged for photomicrography.

of the mirrors are mutually compensating, provided that the numerical aperture does not exceed about 0.60 (visible light). The limiting aperture is somewhat smaller than this for ultraviolet radiation and somewhat larger for infrared, and is smaller for large working distances than for small working distances.

The optical designer, in attempting to achieve numerical apertures larger than about 0.60, may resort to mirrors that depart slightly from spherical curvature; alternatively, he may add one or more lens elements. By such means, objectives of numerical aperture greater than unity and free of chromatic aberration can be realized.

The convex mirror unavoidably obscures a small central portion of the aperture. This alters the diffraction of light (upon which the ultimate resolving power of the objective depends) but fortunately this effect is negligible if proper care is used in the design and construction.

Substage illumination of the specimen can sometimes be achieved by a conventional refracting condenser if only visible light is used. Usually, however, the chromatic aberration of the condenser is troublesome and a condenser similar to the objective is used. Chromatic aberration of the ocular is usually not serious; hence the ocular may be of conventional form, although usually constructed of fused silica, fluorite, or other materials transparent to the infrared or ultraviolet. The ocular may be omitted for special applications in which the primary image formed by the objective is examined directly.   David S. Grey

Illumination. The reflecting microscope has several unique characteristics that make it highly suitable for applications which require nonvisible radiation to illuminate the specimen. Reflecting objectives are achromatic to wavelengths ranging from ultraviolet to the infrared. As a result, specimens can be illuminated with single- or multiple-wavelength radiation without the problem of correcting for chromatic aberrations. There is also a fairly large depth of focus at higher magnifications, so that there is little or no need to adjust the focus as the wavelength of the light source is changed, making photomicrography less problematic for the microscopist using nonvisible illumination.

The reflecting objective lens has a long working distance at higher magnifications and, therefore, is more useful than the common refracting objectives for performing micromanipulations on a specimen. Since their introduction in the 1700s, reflecting lenses have been dramatically improved and can now be constructed with high numerical apertures (0.90) and fewer requirements for aspherical corrections.

Uses. The reflecting microscope can be used with a photometer system for the nondestructive spectral investigation of the absorption and transmission of visible and nonvisible radiation from various objects such as specific regions of cells and tissues, crystals, minerals, synthetic or natural dyed fibers, hairs, inks on documents, paint chips, cosmetics, and polymers. An exceedingly small amount of specimen, with little or no sample preparation, can reveal important optical and chemical information. The reflecting microscope in conjunction with ultraviolet illumination and an ultraviolet spectrophotometer is useful in characterizing the position and content of deoxyribonucleic acid (DNA) in various normal and abnormal cells. The image resolution with infrared illumination is somewhat poorer than with ultraviolet illumination. However, when the microscope is interfaced with a Fourier transform infrared (FTIR) spectrometer, it reveals molecular microstructural information on infrared-illuminated, opaque or transparent objects as small as 1 micrometer in diameter. As a result, the Fourier transform infrared reflecting microscope has found a wide range of applications in the biochemical, medical, and forensic sciences. *See* SPECTROSCOPY.   Lawrence Kobilinsky

Bibliography. G. H. Needham, *The Practical Use of the Microscope Including Photomicrography*, 1958; J. A. Reffner, Molecular microspectral mapping with the FT-IR microscope, *Inst. Phys. Conf. Ser. 98*, 1990.

Reflection and transmission coefficients

When an electromagnetic wave passes from a medium of permeability μ_1 and dielectric constant ϵ_1 to one with values μ_2 and ϵ_2, part of the wave is reflected at the boundary and part transmitted. The ratios of the amplitudes in the reflected wave and the transmitted wave to that in the incident wave are called the reflection and transmission coefficients, respectively. For oblique incidence, the reflection and refraction formulas of optics are most convenient, but for normal incidence of plane waves on plane boundaries, such as occur with transmission lines, waveguides, and some free waves, the concept of wave impedance and characteristic impedance is useful.

For a z-directed wave with electric intensity **E** in the x direction and magnetic intensity **H** in the y direction, the total phasor fields on the incident side are given by Eqs. (1) and (2), where primes are used

$$\check{E}_x = E_0 e^{-jkz} + E_0' e^{jkz} \tag{1}$$

$$\check{H}_y = (\eta)^{-1}(E_0 e^{-jkz} - E_0' e^{jkz}) \tag{2}$$

for reflected quantities and η is the wave impedance. The sign difference in Eqs. (1) and (2) is due to the fact that Poynting's vector, $\frac{1}{2}\mathbf{E} \times \mathbf{H}$, is positive for the incident and negative for the reflected wave. For the transmitted wave, Eqs. (3) hold. Since the tangen-

$$\check{E}_x'' = E_0'' e^{-jkz} \qquad \check{H}_y'' = (\eta'')^{-1} E_0'' e^{-jkz} \tag{3}$$

tial components of **E** and **H** are continuous across the boundary at $z = 0$, $\check{E}_x = \check{E}_x''$ and $\check{H}_y = \check{H}_y''$, so that Eqs. (4) hold. The ratios for the reflected and trans-

$$E_0 + E_0' = E_0'' \qquad \eta''(E_0 - E_0') = \eta E_0'' \tag{4}$$

mitted fields obtained by solving these equations are given by Eqs. (5), which are the reflection and

$$\frac{E_0'}{E_0} = \frac{\eta'' - \eta}{\eta'' + \eta} \qquad \frac{E_0''}{E_0} = \frac{2\eta''}{\eta'' + \eta} \tag{5}$$

transmission coefficients, respectively. *See* ELECTRO-MAGNETIC RADIATION; POYNTING'S VECTOR.

Coefficients for optics. Equations (5) hold for normal incidence in optics, if the velocities v and v'' are written for η and η''. For a plane wave whose electric vector is normal to the plane of incidence and whose direction makes an acute angle θ with the normal to the interface, the reflection and transmission coefficients are given by Eqs. (6), where $v'' \sin \theta''$. When

$$\frac{E_0'}{E_0} = -\frac{\sin(\theta - \theta'')}{\sin(\theta + \theta'')}$$

$$\frac{E_0''}{E_0} = \frac{2 \sin \theta'' \cos \theta}{\sin(\theta + \theta'')} \qquad (6)$$

the electric vector lies in the plane of incidence, the coefficients are given by Eqs. (7). The ratio v/v'' is

$$\frac{E_0'}{E_0} = \frac{\tan(\theta - \theta'')}{\tan(\theta + \theta'')}$$

$$\frac{E_0''}{E_0} = \frac{2 \sin \theta'' \cos \theta}{\sin(\theta + \theta'') \sin(\theta - \theta'')} \qquad (7)$$

the index of refraction. *See* REFRACTION OF WAVES.

Waveguides. In waveguides, as in free space, the characteristic impedance is defined as the ratio of the transverse electric field E_t to the transverse magnetic field H_t. For waveguides, this radio depends on the frequency and the dimensions of the waveguide, as well as on the permeabilities and dielectric constants. For a transverse interface, the boundary conditions used for Eqs. (4) on the tangential fields still hold. Thus, Eqs. (5) for the reflection and transmission coefficients are valid if η and η'' are replaced by the characteristic impedances on the incident and emergent sides, respectively. *See* WAVEGUIDE.

Transmission lines and networks. Let $\check{Z}$ and $\check{Z}''$ be the characteristic impedances on the incident and emergent sides of a discontinuity in a transmission line or on the two sides of a junction between two networks. Then the relations between the potentials and currents of the incident, reflected, and transmitted waves are given by Eqs. (8a)–(8c), respectively.

$$\check{V} = \check{Z}\check{I} \qquad (8a)$$

$$\check{V}' = -\check{Z}\check{I}' \qquad (8b)$$

$$\check{V}'' = \check{Z}'' \check{I}'' \qquad (8c)$$

At the discontinuity, potential and current must be continuous so that Eqs. (9) hold. Solution for the ratios gives Eqs. (10), which are the reflection and

$$\check{V} + \check{V}' = \check{V}'' \qquad \check{I} + \check{I}' = \check{I}'' \qquad (9)$$

$$\frac{\check{V}'}{\check{V}} = \frac{\check{Z}'' - \check{Z}}{\check{Z}'' + \check{Z}} \qquad \frac{\check{V}''}{\check{V}} = \frac{2 \check{Z}''}{\check{Z}'' + \check{Z}} \qquad (10)$$

transmission coefficients. *See* TRANSMISSION LINES.

Coefficients for acoustics. Equations (10) hold in acoustics, provided acoustic impedance is substituted for electrical impedance. Acoustic impedance is defined as the product of the density of a medium by the speed of sound in it. There are two types

of waves in solid mediums, longitudinal waves and shear waves, and thus there are two impedances. *See* ACOUSTIC IMPEDANCE.   William R. Smythe

Bibliography. M. F. Iskander, *Electromagnetic Fields and Waves*, 2000; P. Lorrain and D. R. Corson, *Electromagnetic Fields and Waves*, 3d ed., 1987; S. Ramo, J. R. Whinnery, and T. Van Duser, *Fields and Waves in Communication Electronics*, 3d ed., 1994; W. R. Smythe, *Static and Dynamic Electricity*, 3d ed., 1968, reprint 1984; D. H. Staelin, A. W. Morgenthaler, and J. A. Kong, *Electromagnetic Waves*, 1998.

Reflection of electromagnetic radiation

The returning or throwing back of electromagnetic radiation such as light, ultraviolet rays, radio waves, or microwaves by a surface upon which the radiation is incident. In general, a reflecting surface is the boundary between two materials of different electromagnetic properties, such as the boundary between air and glass, air and water, or air and metal. Devices designed to reflect radiation are called reflectors or mirrors.

Reflection angle. The simplest reflection laws are those that govern plane waves of radiation. The law of reflection concerns the incident and reflected rays (as in the case of a beam from a flashlight striking a mirror) or, more precisely, the wave normals of the incident and reflected rays and the normal to the reflecting surface all lie in one plane, called the plane of incidence, and that the reflection angle θ_{refl} equals the angle of incidence θ_{inc} as in Eq. (1) [**Fig. 1**]. The

$$\theta_{refl} = \theta_{inc} \qquad (1)$$

angles θ_{inc} and θ_{refl} are measured between the surface normal and the incident and reflected rays, respectively. The surface (in the above example, that of the mirror) is assumed to be smooth, with surface irregularities small compared to the wavelength of the radiation. This results in so-called specular reflection. In contrast, when the surface is rough, the reflection is diffuse. An example of this is the diffuse scattering
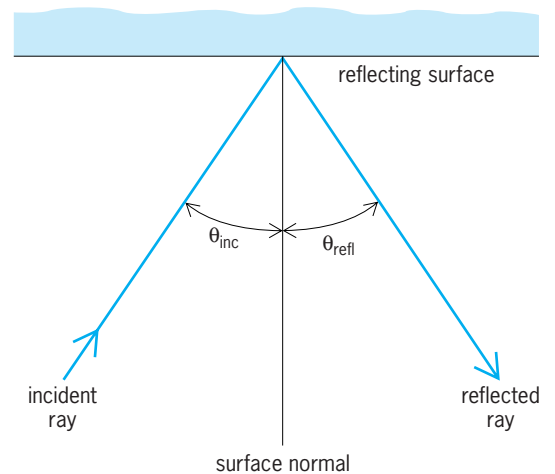

Fig. 1. Reflection of electromagnetic radiation from a smooth surface.

of light from a screen or from a white wall where light is returned through a whole range of different angles.

Reflectivity. The reflectivity of a surface is a measure of the amount of reflected radiation. It is defined as the ratio of the intensities of the reflected and incident radiation. The reflectivity depends on the angle of incidence, the polarization of the radiation, and the electromagnetic properties of the materials forming the boundary surface. These properties usually change with the wavelength of the radiation. Reflecting materials are divided into two groups: transparent materials and opaque conducting materials. Radiation penetrating a transparent material propagates essentially unattenuated, while radiation penetrating a conducting material is heavily attenuated. Transparent materials are also called dielectrics. In the wavelength range of visible light, typical dielectrics are glass, quartz, and water. Conducting materials are usually metals such as gold, silver, or aluminum, which are good reflectors at almost all wavelengths. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; DIELECTRIC MATERIALS.

Reflection from metals. The reflectivity of polished metal surfaces is usually quite high. Silver and aluminum, for example, reflect more than 90% of visible light. The reflectivity of some common metals as a function of wavelength is given in **Fig. 2** for normal incidence ($\theta_{inc} = 0$). The reflectivities vary considerably with wavelength, generally falling off toward the shorter wavelengths, with silver exhibiting a reflection "window" at 320 nanometers. The reflectivity values depend somewhat on the way the metal surface was prepared; for example, whether or not it was polished or was produced by evaporation. The presence of an oxidation layer is also a factor influencing (and usually decreasing) the reflectivity. In ordinary mirrors the reflecting surface is the interface between metal and glass, which is thus protected from oxidation, dirt, and other forms of deterioration. When it is not permissible to use this protection for technical reasons, one uses "front-surface"
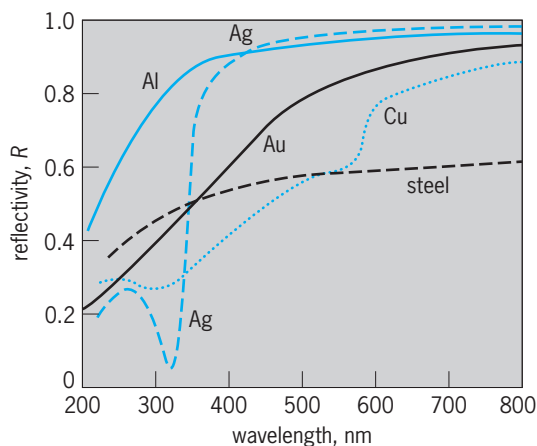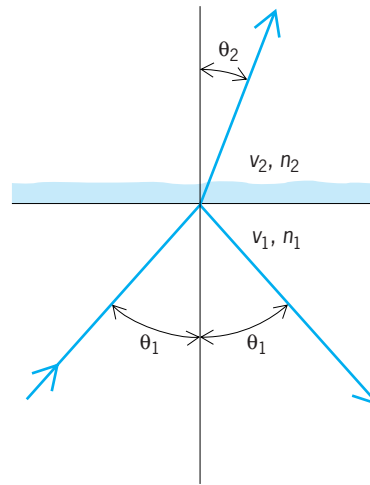
Fig. 3. **Reflection and refraction of electromagnetic radiation at an interface between two dielectric media.**

mirrors, which are usually coated with evaporated aluminum. The aluminum has a high reflectivity and deteriorates relatively little in the atmosphere. Front-surface mirrors are used in scientific applications such as interferometry and in large reflecting telescopes.

Reflection from dielectrics. The material property that determines the amount of radiation reflected from an interface between two dielectric media is the phase velocity v of the electromagnetic radiation in the two materials. In optics one uses as a measure for this velocity the refractive index n of the material, which is defined by Eq. (2) as the ratio

$$n = \frac{c}{v} \tag{2}$$

of the velocity of light c in vacuum and the phase velocity in the material. For visible light, for example, the refractive index of air is about $n = 1$, the index of water is about $n = 1.33$, and the index of glass is about $n = 1.5$. *See* PHASE VELOCITY; REFRACTION OF WAVES.

For normal incidence ($\theta_{inc} = 0$) the reflectivity R of the interface is given by Eq. (3), in which the

$$R = \left(\frac{v_1 - v_2}{v_1 + v_2}\right)^2 = \left(\frac{n_2 - n_1}{n_2 + n_1}\right)^2 \tag{3}$$

material constants are labeled 1 and 2 as shown in **Fig. 3**, where the radiation is incident in material 1. The reflectivity of an air-water interface is about 2% ($R = 0.02$) and that of an air-glass interface about 4% ($R = 0.04$); the other 98% or 96% are transmitted through the water or glass, respectively.

A ray incident upon the interface at an oblique, nonnormal angle θ_1 is deviated as it penetrates material 2 as shown in Fig. 3. This is called refraction, and the refracted angle θ_2 follows from Snell's law of refraction, Eq. (4). Again, there is partial reflec-

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \tag{4}$$

tion, with the reflectivity depending on the angle of incidence θ_1 and the polarization of the radiation.

Fig. 2. **Reflectivity of some common metals for normal incidence as a function of wavelength. (*After F. A. Jenkins and E. White, Fundamentals of Optics, 4th ed., McGraw-Hill, 1976*)**
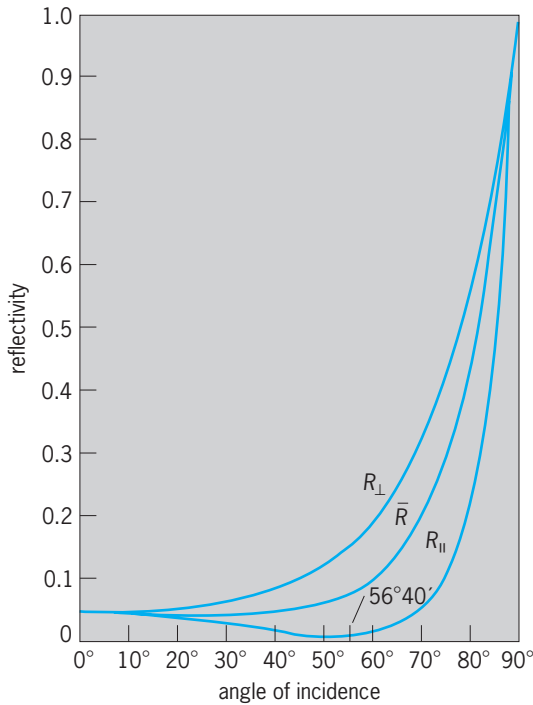
Fig. 4. Reflectivity as a function of angle of incidence for an air-glass interface. (*After M. Born and E. Wolf, Principles of Optics, 5th ed., Pergamon Press, 1975*)

When the electric field is polarized parallel to the plane of incidence, the reflectivity $R_{\parallel}$ is given by Eq. (5), and when the electric field is polarized perpendicular to the plane of incidence, the reflectivity $R_{\perp}$ is given by Eq. (6). These formulas are known as

$$R_{\parallel} = \frac{\tan^2 (\theta_1 - \theta_1)}{\tan^2 (\theta_1 + \theta_2)} \qquad (5)$$

pendicular to the plane of incidence, the reflectivity $R_{\perp}$ is given by Eq. (6). These formulas are known as

$$R_{\perp} = \frac{\sin^2 (\theta_1 - \theta_2)}{\sin^2 (\theta_1 + \theta_2)} \qquad (6)$$

the Fresnel formulas. For unpolarized radiation, in which the electric field varies rapidly in a random, unpredictable manner, the reflectivity $\overline{R}$ is the average of $R_{\parallel}$ and $R_{\perp}$. **Figure 4** shows $R_{\parallel}$, $R_{\perp}$, and $\overline{R}$ for an air-glass interface as a function of angle. The reflectivity approaches 100% at grazing incidence ($\theta_1 \approx 90°$) for both polarizations. *See* REFLECTION AND TRANSMISSION COEFFICIENTS.

At an angle of about $\theta_1 = \theta_B = 56°$ the reflectivity $R_{\parallel}$ assumes zero value. This angle is called Brewster's angle, which is, in general, obtained from formula (7). At this angle, which is also called the

$$\tan \theta_B = \frac{n_2}{n_1} \qquad (7)$$

polarizing angle, only radiation polarized perpendicular to the plane of incidence is reflected.

Total internal reflection. Total reflection, that is, reflection of 100% of the incident radiation, occurs at the interface of two dielectrics when the radiation is incident in the denser medium, that is, when $n_1 > n_2$ and when the angle of incidence θ_1 is larger than

the critical angle θ_0 given by Eq. (8). Total internal

$$\sin \theta_0 = \frac{n_2}{n_1} \qquad (8)$$

reflection can be observed, for example, by a submerged diver looking up at the water-air interface for which the critical angle is about $\theta_0 = 49°$. For a glass-air interface the critical angle is approximately $\theta_0 = 42°$.

Selective reflection from crystals. The discussion of reflection from dielectrics to this point has been concerned with reflectivity of nonabsorbing media far removed from absorption bands. These bands are located in the spectral regions where the frequency of the radiation corresponds to a resonance frequency of the atoms, molecules, or crystal lattice of the medium. Since this radiation is strongly absorbed, it is also strongly reflected. The metallic sheen of dye crystals, which have very strong absorption bands in the visible spectrum, is caused by selective reflection. Crystalline solids such as rocksalt or quartz, the lattices of which are built up of atoms bearing net electric charges, show strong selective reflection in the infrared region at wavelengths near those of the strong absorption bands associated with lattice vibrations in the crystal. By reflecting an infrared beam several times from such a material, highly monochromatic radiation can be obtained at the specific wavelengths. These monochromatic beams are referred to as residual rays or reststrahlen. **Figure 5** shows residual rays for some crystals. *See* IONIC CRYSTALS.

Antireflection coatings. In order to reduce undesired reflections from surfaces of optical components such as photographic lenses, one can coat the surface with a thin film. To minimize reflection, the refractive index n_f of this film should be the geometric mean of the indices n_1 and n_2 of the incident and refracting media, Eq. (9), and the film thickness

$$n_f^2 = n_1 n_2 \qquad (9)$$

d_f should be equal to a quarter of the wavelength of light in the film material, that is, it should satisfy Eq. (10), where λ is the vacuum wavelength of the

$$n_f d_f = \frac{\lambda}{4} \qquad (10)$$

light for which the reflectivity of the surface is minimized. Such a film is called a quarter-wave layer. Materials such as magnesium fluoride with an index of $n = 1.38$ are used for antireflection coatings have been made which reduce the reflectivity to less than 0.1% on certain materials; on glass surfaces less than $^1/_2$% has been achieved.
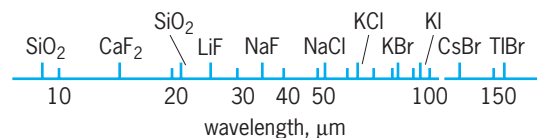

Fig. 5. Residual rays for various crystals.

High-reflectivity multilayers. A series of dielectric quarter-wave layers of alternating high and low refractive index can be used to produce reflectors of high reflectivity at a specified wavelength. This is called a quarter-wave stack. The high reflectivity is due to an interference effect. Stacks with about 20 layers are used to make reflectors with reflectives higher than 99%. Mirrors such as these are used for the construction of laser resonators. Multilayer coatings are also used to enhance the reflectivity of metals such as aluminum. About four quarter-wave layers can enhance the metal reflectivity to values better than 99%. *See* ALBEDO; GEOMETRICAL OPTICS; INTERFERENCE FILTER; MIRROR OPTICS; REFLECTION OF SOUND.   Herwig Kogelnik

Bibliography. *American Institute of Physics Handbook*, 3d ed., 1972; M. Born and E. Wolf, *Principles of Optics*, 7th ed., 1999; B. D. Guenther, *Modern Optics*, 1990; E. Hecht and A. Zajac, *Optics*, 3d ed., 1997; F. A. Jenkins and E. White, *Fundamentals of Optics*, 4th ed., 1976; J. R. Meyer-Arendt, *Introduction to Classical and Modern Optics*, 1995.

Reflection of sound

The return of sound waves from surfaces on which they are incident. Suppose a sound wave strikes a smooth surface that is large compared to the wavelength of the sound and suppose that the path of this sound wave is represented by a ray, that is, by a line perpendicular to the advancing wavefront. By the law of reflection, the angle of reflection *r* for this ray equals the angle of incidence *i*, and the reflected ray lies in the plane of incidence. **Figure 1** shows reflection from a plane surface.

The geometrical laws for reflection of sound waves are the same as those for light waves. The apparent differences involve only questions of scale, because the average wavelength of sound is about 100,000 times that of light. For example, a mirror or lens used to produce a beam of sound waves must be enormously large compared to mirrors and lenses used in optical systems. *See* REFLECTION OF ELECTROMAGNETIC RADIATION.

A concave surface tends to concentrate the reflected sound waves (**Fig. 2**). Such surfaces are sometimes used to advantage as reflectors, but if used indiscriminately they may lead to poor acoustics as a result of undesirable focusing effects. Convex re-
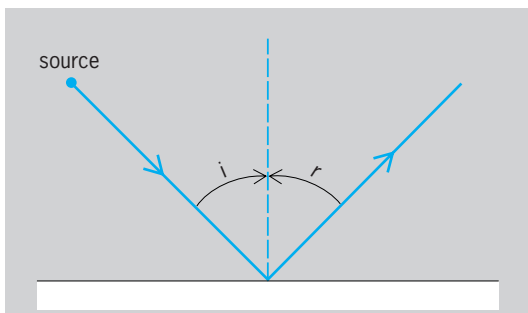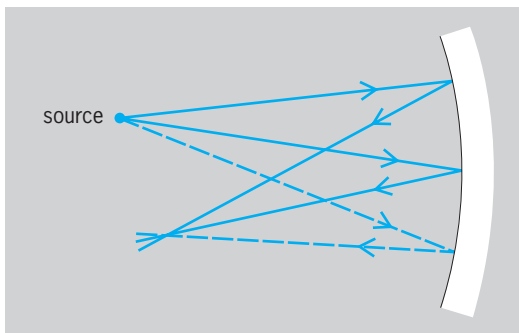

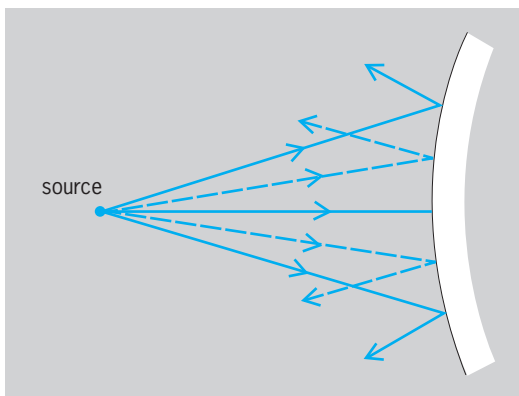Fig. 2. Reflection from a concave spherical surface.

Fig. 3. Reflection from a convex surface.

flectors (**Fig. 3**) tend to spread the reflected waves. Therefore, when placed at the boundaries of a room, they tend to diffuse the sound throughout the room. For this reason, some radio-broadcasting studios employ cylindrical convex panels as part of their wall construction to promote diffusion.

The law of reflection is often used in investigating the effects of various shapes of a proposed room on the distribution of sound in that room. Such studies can lead to the design of interior surfaces that will give beneficial reflections, or to the elimination or modification of surfaces that otherwise would give rise to echoes. However, caution must be exercised in such applications of the law of reflection, because the wave properties of sound are neglected in this simplification of the behavior of sound. For a discussion of sound reflections in water *see* UNDERWATER SOUND; ARCHITECTURAL ACOUSTICS; ECHO; SOUND.   Cyril M. Harris

Reflex

A simple, unlearned, yet specific behavioral response to a specific stimulus. Reflexes are exhibited by virtually all animals from protozoa to primates. Along with other, more complex stimulus-bound responses such as fixed action patterns, they constitute much of the behavioral repertoire of invertebrates. In higher animals, such as primates, where learned behavior dominates, reflexes nevertheless persist as an important component of total behavior.

Fig. 1. Reflection from a plane surface.

Reflexes are characteristic of a species and its relatives; that is, they are genetically determined. In humans, no effort is needed to acquire reflexes; they simply occur automatically, and many take place below the conscious level. For example, although heart rate and ventilation rate are constantly adjusted by reflexes, most of the time people are unaware of these changes. Similarly, pupillary diameter and blood pressure are reflexively regulated without conscious knowledge. Other reflex responses, however, such as perspiring, shivering, blinking, and maintaining posture, are more visible but occur without conscious intervention.

The simplest known reflexes require only one neuron or, in the strictest sense, none. For example, ciliated protozoa, which are single cells and have no neurons, nevertheless exhibit apparently reflexive behaviors. When a paramecium swims into an object, it reverses the power stroke of the cilia, backs away a short distance, turns, and again moves forward. When touched caudally, the animal accelerates in the forward direction. Here, the animal's cell membrane itself serves as the receptor of the stimulus, and the cilia act as effectors for directed movement. In higher organisms, there are also very simple reflexes. For example, when human skin is injured sufficiently to stimulate a single pain neuron, an unknown substance is released which causes small local blood vessels to dilate. In contrast to these simple responses, however, most reflexes require activity in a large circuit of neurons. *See* NEURON.

Anatomy and physiology. The neurons involved in most reflexes are connected by specific synapses to form functional units in the nervous system. Such a circuit begins with sensory neurons and ends with effector cells such as skeletal muscles, smooth muscles, and glands, which are controlled by motor neurons. The central neurons which are often interposed between the sensory and motor neurons are called interneurons. When first described, this sequence of neurons was called a reflex arc. The sensory side of the reflex arc conveys specificity as to which reflex will be activated. That is, the sensory cells themselves determine which environmental change is sensed, either within the body or without. The remainder of the reflex response is governed by the specific synaptic connections that lead to the effector neurons.

A familiar reflex is the knee-jerk or stretch reflex. It involves the patellar (kneecap) tendon and a group of upper leg muscles. Other muscle groups show similar reflexes, but this one is described here because of its widespread clinical use. It is a relatively simple reflex arc in which the terminals of the sensory cells synapse directly on the effector neurons. It is thus an example of a monosynaptic reflex (**Fig. 1**). When the patellar tendon is tapped, a brief stretch is imposed on the attached muscles (quadriceps femoris), including the muscle spindles which are embedded within the quadriceps. This stretch causes the sensory neurons of the spindles to fire impulses which return to the spinal cord. Within the spinal cord, excitatory synapses are activated on the specific motor neurons supplying the same muscles that

were stretched, that is, the quadriceps. This results in a volley of impulses in the motor neurons which activate a brief contraction of the quadriceps, which in turn extends the leg. A possible function of stretch reflexes is to oppose sudden changes in length of skeletal muscle, and these reflexes are likely operating even during normal movements.

The stretch reflex as well as other more complex reflexes involve the inhibition of antagonistic muscle groups. This requires the activation of inhibitory interneurons in the spinal cord which synapse upon the motor neurons of antagonistic muscles. For example, although not shown in Fig. 1, the sensory neurons of the quadriceps also produce inhibition of motor neurons that contact antagonistic (flexor) muscles on the opposite side of the leg. In still other complex reflexes, there is a spread of activity to higher and lower levels of the spinal cord (multisegmental) and to the opposite side of the body (contralateral). Examples of reflexes with these characteristics include the flexor reflex and accompanying crossed extensor reflex (**Fig. 2**). These reflexes are initiated by a painful stimulus. The returning sensory signals excite the motor neurons of the leg flexor muscles while inhibiting those of the leg extensor muscles on the same side as the stimulus (ipsilateral side). This assures that the foot is quickly lifted away from the harmful stimulus. Balance, however, must also be maintained, and this is accomplished by exciting the leg extensors of the opposite (contralateral) limb; thus the term crossed extensor reflex. The flexor reflex is a prime example of a protective response (see below). It occurs so rapidly that the pain is felt only after the withdrawal response is complete. Speed of response is particularly important when the severity of the injury is time-dependent, such as in response to a chemical exposure or burn.

Although only a few neurons were drawn into Figs. 1 and 2 for clarity, many thousands are usually
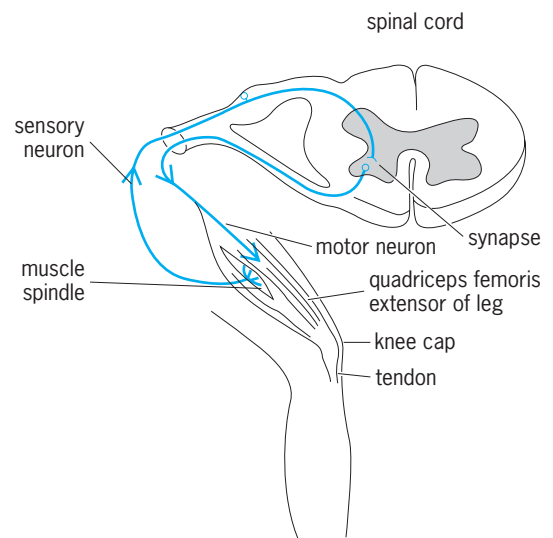

Fig. 1. Simplified diagram of the excitatory component of a stretch reflex. This reflex is said to be monosynaptic because there is a single synapse between the sensory neuron and motor neuron.
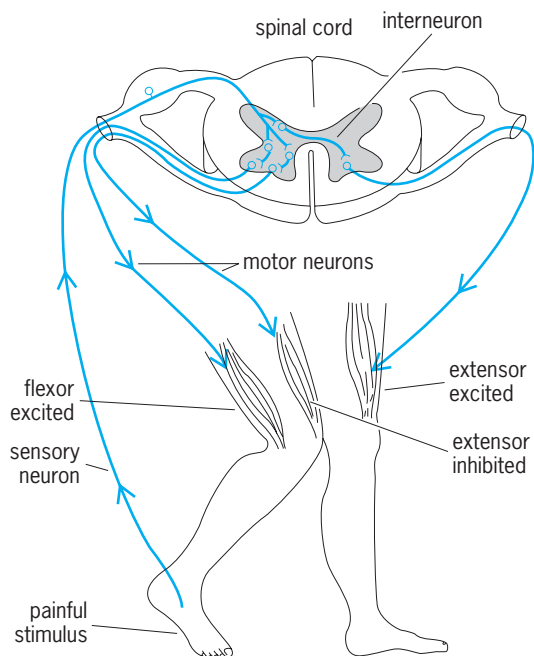
Fig. 2. Diagram of the flexor and crossed extensor reflexes. This reflex is said to be monosynaptic because there is a single synapse between the sensory and motor neuron.

involved. A few to many sensory neurons may activate several hundred to several thousand motor neurons, depending upon the reflex involved. Each motor neuron in turn branches to synapse on as many as several hundred muscle cells. In addition, in multisegmental reflexes and in those reflexes with bilateral effects, another thousand or more interneurons may become involved.

Functions. Reflexes adjust many important biological functions rapidly and efficiently without conscious effort, while other reflexes are largely protective. For example, the most delicate and sensitive sensory systems, the eye and the ear, can be damaged or destroyed by overstimulation or by accident. The amount of light admitted to the retina is controlled by the pupillary reflex, which in humans can effect a change in pupillary diameter from about 8 mm (0.30 in.) in darkness to 2 mm (0.08 in.) in bright light. A sudden flash of intense light can evoke the reflex closure of the eyelids, further protecting the delicate retina. The eyeball itself is protected from drying by the blink reflex and from mechanical injury by the eyelid closure reflex. The latter reflex is triggered when an object approaches the eye or when the lashes or the cornea are touched. The ear is similarly protected from potentially damaging sounds through the reflex contraction of middle-ear muscles in response to loud noise. This reflex serves to lower the efficiency with which sound vibrations are transmitted through the bones of the middle ear, and thus reduces the possibility of damage to the delicate hair cell receptors of the inner ear (cochlea). *See* EAR (VERTEBRATE); EYE (VERTEBRATE); HEARING (HUMAN).

Stereotyped escape responses constitute another category of protective reflexes exhibited by many animals. For example, a startled squid takes evasive action by contracting its mantle muscles, forcing a jet of water through its siphon. Fishes respond to waterborne vibrations by contracting muscle on one side of their bodies. The reflex contraction occurs on the side opposite the source of vibration, and the result is a sudden move away from possible harm.

Equally important for survival are the various feeding reflexes exhibited by animals. For example, flies and many other insects have chemoreceptors on their feet, mouthparts, and antennae. If a hungry fly walks on a surface moistened with nutrients, a set of reflexes is initiated. When the receptors on the feet are stimulated, there is a reflex extension of the proboscis. If the proboscis receptors are favorably stimulated, the animal begins to drink. Drinking continues until the crop is sufficiently distended to stimulate its stretch receptors. Finally, this stimulus initiates the reflex termination of feeding.

Instinctive or innate behaviors such as courtship rituals, nestbuilding, aggressive, and territorial behaviors show considerable similarities to reflexes. Although generally more complex, they are, like reflexes, unlearned, species-specific, genetically determined, and stereotypic in nature. Importantly, fixed action patterns such as these are like reflexes in being initiated by a specific stimulus, called a sign stimulus or a releaser. In the broad sense, both forms of behavior are also similar in that they are thought to be controlled by specific sets of neurons that underlie each behavior. Like the still more complex learned behaviors, the neural basis of the instinctive behaviors remains largely known. *See* INSTINCTIVE BEHAVIOR; REPRODUCTIVE BEHAVIOR; TERRITORIALITY.

Modification (plasticity). In general, the strength of a reflexive response depends upon the magnitude of the stimulus. However, many reflexes tend to gradually weaken if the stimulus is applied repeatedly. This phenomenon, called habituation, allows an animal to ignore a familiar or repeated stimulus. If a strong, novel stimulus is presented along with a reflex-evoking stimulus, however, the strength of the reflex is often enhanced. This process, called sensitization, causes the animal to attend to a potentially threatening situation by responding strongly to its environmental cues. Recent analysis of the changes in reflex circuits during habituation and sensitization have shown changes in synaptic transmission between neurons in the reflex. These changes are now understood in some detail at the cellular and molecular levels.

Similarly, reflexes may be modified by classical (pavlovian) conditioning, a form of associative learning. For example, in dogs, the sight and smell of food cause the reflexive secretion of saliva. A light or the sound of a bell does not. Ivan Pavlov showed that pairing the sight and smell of food (unconditioned stimuli) with a light or bell caused the animal to associate the light or bell with food. Thus, after several training sessions, presentation of the light or bell alone caused salivation. Whether classical conditioning provides a significant portion of learned behavior is still controversial. *See* CONDITIONED REFLEX.

Finally, reflexes are often modified during voluntary movement or locomotion. Motion itself

stimulates many sensory receptors and could therefore elicit reflexes which oppose the intended movement. Such reflexes are thought to be overridden or suppressed by commands from the brain in order for the desired movements to be executed.

Clinical uses. Since reflex responses are essentially invariant in healthy individuals, their examination can often provide valuable information about neurological disorders. Relatively simple tests, such as pupillary responses or auditory perceptions, can give information about the normal operation of brain functions. Scratching the skin of the foot, which leads normally to flexion of the toes, can show the status of the spinal circuits that control them. For example, if the toes extend and spread in response to scratching the sole of the foot (Babinski response), this is indicative of a spinal lesion in the pyramidal tracts descending from the motor centers of the cerebral cortex. When observations on these and many other reflexes are combined with data on paralyses, spasms, or losses of sensation, it is often possible to localize the site of the lesion. *See* NERVOUS SYSTEM (INVERTEBRATE); NERVOUS SYSTEM (VERTEBRATE); NEUROBIOLOGY.   J. L. Larimer

Bibliography. J. M. Camhi, *Neuroethology*, 1984; R. Gilles et al. (eds.), *Advances in Comparative and Environmental Physiology*, vol. 2, 1988; J. Nolte, *Principles of Neurobiology*, int. ed., 1991; R. A. Rhodes and G. A. Tanner (eds.), *Medical Physiology*, 1995; G. M. Shepherd, *Neurobiology*, 3d ed., 1997; R. N. Singh and N. J. Strausfeld (eds.), *Neurobiology of Sensory Systems*, 1990.

Reforestation

The reestablishment of forest cover either naturally or artificially. Given enough time, natural regeneration will usually occur in areas where temperatures and rainfall are adequate and when grazing and wildfires are not too frequent. For example, Canada was basically treeless 16,000 years ago (during the last glacial period). During that time, spruce trees probably dominated much of present Illinois and Missouri. Due to global warming, spruce forests gradually spread north and now cover much of Canada. Within 16,000 years, more than 1 billion acres (400 million hectares) of forests in Canada were regenerated naturally.

Human impact. Rapid reforestation was not important when human populations were low in number and food was obtained by gathering wild plants and following game. Mature natural forests usually produced wood at a rate of 14–28 ft³/(acre)(yr) [1–2 m³/(ha)(yr)], and this often exceeded the rate of harvesting firewood. However, with the adoption of agriculture, humans settled down and their populations increased, placing pressures on forests. Not only did the demand for fuelwood increase, but in many places forests were cut down to make room for agriculture. For example, in 1600 about 49% of the continental United States was covered with forests, but this had been reduced to 33% by 1900. In just

300 years, population pressures reduced forestlands by about 1 million acres/yr (400,000 hectares/yr). Since natural regeneration was not keeping up with the rate of harvests, the need for rapid reforestation became more apparent.

During the late 1800s, a few foresters and government officials expressed concern over future wood supplies in the United States and the floods caused by deforestation. Two solutions to these problems were proposed. One resulted in the creation of national forests where wood and clean water could be produced in perpetuity. The other promoted artificial regeneration (the establishment of trees by planting or direct seeding). Due, in part, to a combination of improved technology in both agricultural and artificial regeneration, the United States now has about the same amount of forestland as it did 80 years ago. This is not true in many countries where human populations are increasing but artificial regeneration has been limited.

Planting and seeding. Reforestation occurs on land where trees have been recently removed due to harvesting or to natural disasters such as a fire, landslide, flooding, or volcanic eruption. When abandoned cropland, pastureland, or grasslands are converted to tree cover, the practice is termed afforestation (where no forest has existed in recent memory). Afforestation is common in countries such as Australia, South Africa, Brazil, India, and New Zealand. Although natural regeneration can occur on abandoned cropland, planting trees will decrease the length of time required until the first harvest of wood. Planting also has an advantage in that both tree spacing and tree species can be prescribed. The selection of tree species can be very important since it affects both wood quality and growth rates. Direct seeding is also used for both afforestation and reforestation, although it often is less successful and requires more seed than tree planting. Unprotected seed are often eaten by birds and rodents, and weeds can suppress growth of newly germinated seed. For these reasons, direct seeding accounts for only about 5% and 1% of artificial reforestation in Canada and the United States, respectively.

Growth rate improvements. In just a 5-year period (1990–1995), the world population increased by 0.4 billion and forests declined by 139 million acres (56 million hectares) due mostly to an increase in pastureland and cropland. Since the world population is expected to increase by more than 3 billion by the year 2050, some predict wood consumption will increase by 50%. How much this additional demand will deplete natural stands will depend, in part, on how much wood can be produced from tree plantations. The amount of plantation wood harvested in 2050 is predicted to meet half the demand. However, the actual amount will depend on improvements in growth rates and on the amount of afforestation (**Fig. 1**). Growth rates of plantations can be increased by selecting fast-growing species, by genetic selection, and by various cultural treatments.

Species selection. The objectives of the landowner determine which tree species are planted. Some
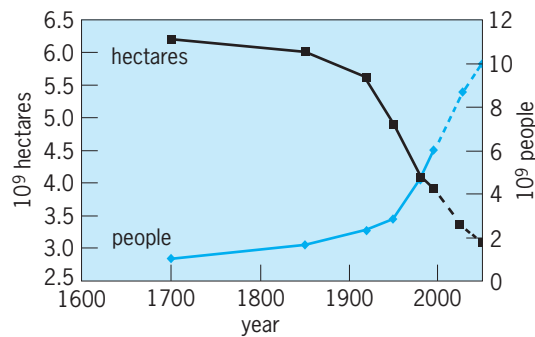
Fig. 1. Forests and people in the world.

species grow fast and straight, while others grow slow and crooked. If a landowner's objective is to produce fuelwood quickly, then an exotic species might be selected. In many cases, exotic species grow faster than native trees and sometimes can produce the same weight of wood from one-fifth the land. Some believe the faster growth rates of exotics is due to a lack of belowground pests.

Genetic improvement. Most developed countries have improvement programs for their most commonly planted tree species. Trees with rapid growth rates and desirable wood quality traits are bred in special areas called seed orchards. The progeny of various crosses are tested, and the best offspring are selected to be included in future seed orchards. In some cases when exceptional crosses are made, the genotypes are replicated using various vegetative propagation techniques. The most common method is to produce rooted cuttings. For example, about 30% of the pines planted in New Zealand are rooted cuttings. A more expensive method of vegetative propagation involves tissue culture. Although tissue culture techniques are used mainly by forest researchers, some companies in Brazil, Canada, and New Zealand are producing a limited amount of trees using either organogenesis or somatic embryogenesis. If the cost of these methods can be greatly reduced, it will be much easier to mass-produce genotypes with high growth rates. *See* FOREST GENETICS.

Seedling size. The best genetic improvements can be wasted if the nursery produces inferior seedlings. The size and condition of the seedling is related to its survival and growth after transplanting. Seedlings with small roots usually exhibit a limited amount of new root growth, undergo transplanting shock, and may die when soil moisture is limited. Small seedlings usually do not compete well with weeds and may require more intensive site preparation than seedlings with large root systems. In some regions of Canada, companies have moved away from applying herbicides to increase the growth of small seedlings. Instead, they are now planting large stock that can outgrow the weeds. It was once believed that money could be saved by purchasing small nursery stock. But it is now generally recognized that in many cases total establishment costs can be reduced by planting large stock instead of applying overly intensive mechanical site preparation to improve the growth of small seedlings.

Mechanical site preparation. On some sites, environmental conditions limit early growth of trees, and large machines are used to ameliorate the soil. The soil is sometimes too compacted, or too wet or too cold to allow rapid root growth. When the water table is close to the soil surface, ditches may need to be made to allow the soil to breathe. On some poorly drained sites, mounds or ridges are used to prevent young seedlings from being inundated during wet weather. Rippers are sometimes used to break up compacted soils. Although mechanical site preparation is sometimes used to control woody weeds, a reduction in site productivity can occur when part of the topsoil is removed during root-raking operations. In some regions the use of herbicides has reduced soil degradation caused when inappropriate mechanical methods are used.

Weed control. Foresters tend to group weeds into either herbaceous or woody. Herbaceous weeds include grasses and annuals, while woody weeds include shrubs and noncrop trees. Herbaceous weeds are usually found on afforestation sites and compete with trees mainly during the early years of the stand. Controlling herbaceous weeds at the time of planting can increase early growth and can shorten rotation lengths by 1 or 2 years. Controlling herbaceous weeds after the plantation is 10 years old will likely have no effect on wood production. In contrast, woody weeds can compete with crop trees throughout the rotation. Therefore, reducing the growth of woody weeds at either time of establishment or at midrotation can increase the growth of crop trees. In many cases, complete failures have occurred when woody sprouts were not controlled. Although it was once believed that a mixture of tree species would be the best economic approach, this is not the case since all woody species do not have the same value at harvest.

Fertilization. In the Southern Hemisphere, afforestation sites are usually fertile and growth can exceed 300 ft^3/(acre)(yr) [21 m^3/(ha)(yr)]. In contrast, in developed countries in the Northern Hemisphere, naturally fertile sites are usually occupied by agricultural crops and forest trees are planted on eroded and nutrient-poor sites. Some plantations on nutrient-poor sites produce only 140 ft^3/(acre)(yr) [10 m^3/(ha)(yr)]. Not only can growth be reduced, but wood quality can be reduced when certain micronutrient deficiencies exist. On phosphorus-deficient sites, fertilization can greatly increase yields of pine plantations. In the southern United States, past agricultural practices have greatly reduced topsoil depth, and although rainfall and temperatures are favorable, a limited amount of available nutrients restricts tree growth. Therefore, fertilization in pine plantations is conducted by forest industry. In total, forest industry uses only about 0.2% of all the nitrogen fertilizer used in the United States. Currently more than 1.2 million acres (485,000 ha) of plantations in this region were fertilized in 1998. It is estimated that more than half of all forest fertilization in the world is used on nutrient-deficient sites in the United States.

Fertilization can be made either at time of planting or at some point in midrotation. In general, phosphorus applications are made at the time of planting. Nitrogen applications are more beneficial after the tree roots have expanded their size and are better able to capture the added nitrogen. In either case, it is often wise to control weeds so they do not benefit from the fertilization more than the trees. *See* SOIL ECOLOGY.

Tree planting rates. Globally, the rate of tree planting gradually increased throughout the twentieth century. For example, only about 139,000 acres (56,000 ha) of land were planted in 1930, but now more than 2.5 million acres (1 million hectares) are planted annually in the United States. Some of this is due to afforestation since farmers are finding it more profitable in the long run to plant trees on cropland than to grow certain annual crops. Although tree planting is currently increasing on land owned by farmers and companies, the acreage of tree planting on federal land has been declining since 1990, when it peaked at 388,000 acres (157,000 ha). The planting rate by the federal government is now about half that amount, due mainly to a reduction in harvesting from National Forests. Governments in several other developed countries have also reduced their planting rate. Currently, a majority of industrial plantations are located in developed countries in Europe, North America, and the former Soviet Union. However, this may change in the future as governments in countries with expanding populations are continuing to encourage the planting of trees.

Fiber farms. Tree plantations can be classified into three groups according to management intensity. Most tree farms are managed with a low level of management intensity, where almost no inputs are made after the trees are 2 years old. These plantations might never receive any fertilization, weed control, insect control, or irrigation. This level of management is often used by a large number of private landowners who usually have limited amounts of capital. Medium-level management may involve herbicides and fertilization at establishment, and may also receive fertilizers one or two times during midrotation. Insect populations may be monitored, and preventative treatments may be applied when economical. Forest companies usually have the capital for this level of management. The increased growth rates allow the company to produce the same amount of wood on fewer acres (**Fig. 2**). This allows some land to be devoted to other objectives such as wildlife and streamside management. A high level of input on fiber farms increases the yields. Fiber farms typically grow hardwoods, have a 10-year or less rotation, and are irrigated and fertilized annually, and each farm is less than 25,000 acres (10,000 ha) in size. The management of fiber farms is closer to that of intensive horticulture than it is to traditional forestry. Currently less than 25 fiber farms exist in North America. But over time, as pressures by a growing population increase, the higher yields from fiber farms [400–800 ft^3/(acre)(yr) or 28–56 m^3/(ha)(yr)] will help to reduce the demand for wood that peo-
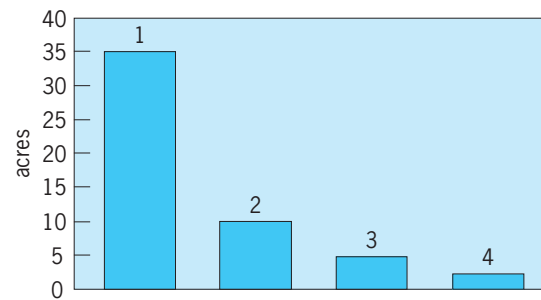

Fig. 2. Relative amounts of land needed to produce 1000 ft^3 of wood per year. (1) Mature, upland, natural hardwood stand. (2) Pine plantation owned by a farmer and managed with minimum inputs. (3) Pine plantation managed by a forest company and managed with medium inputs. (4) Hybrid-poplar fiber farm using trickle irrigation, fertilization, and weed control.

ple place on natural forests [at 14–28 ft^3/(acre)(yr) or 1–2 m^3/(ha)(yr)]. *See* BREEDING (PLANT); FOREST AND FORESTRY; FOREST MANAGEMENT.   David B. South

Bibliography. J. R. Boyle et al., *Planted Forests: Contributions to the Quest for Sustainable Societies*, Kluwer Academic, Dordrecht, 1999; J. Evans, *Plantation Forestry in the Tropics*, Oxford University Press, Oxford, 1992; Food and Agriculture Organization of the United Nations, *State of the World's Forests*, FAO, Rome, 1999; T. Gardner-Outlaw and R. Engelman, *Forest Futures*, Population Action International, Washington, DC, 1999; R. W. Haynes, *An Analysis of the Timber Situation in the United States: 1989–2040*, USDA Forest Service, Washington, DC, 1990; J. Pöyry, *Global Outlook for Plantations*, Australian Bureau of Agricultural and Resource Economics, Canberra, 1999; M. Williams, *Americans and Their Forests: A Historical Geography*, Cambridge University Press, Cambridge, 1989.
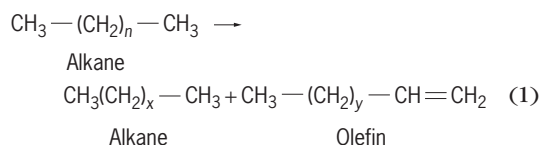
Reforming processes

Those processes used to convert, with limited cracking (thermal decomposition), petroleum liquids into higher-octane gasoline. When the demand for higher-octane gasoline developed during the early 1930s, attention was directed to ways of improving the octane number of fractions within the boiling range of gasoline. Straight-run (distilled without change) gasoline frequently had low octane numbers, and any process that would improve its octane numbers would aid in meeting the demand for higher-octane gasoline. Such a process, called thermal reforming, was developed from thermal cracking processes and is used widely. *See* CRACKING; GASOLINE; OCTANE NUMBER.

The production of liquid product streams by distillation or by cracking processes is the first of a series of steps leading to the production of marketable liquid petroleum products. Several other unit processes are involved in the production of a final product. Such processes are generally called secondary processes or product improvement processes since they

are not used directly on the crude petroleum (oil) but are used on primary product streams that have been produced from the crude petroleum. Product improvement processes, such as reforming processes, are those processes in which the molecular structure of the feedstock is reorganized. Examples are the conversion (reforming, molecular rearrangement) of *n*-hexane to cyclohexane, and cyclohexane to benzene. These processes reform or rearrange one particular molecular type to another, thereby changing the properties of the product relative to the feedstock. *See* PETROLEUM.

Chemistry. Upgrading by reforming may be accomplished, in part, by an increase in volatility (reduction of molecular size) or by the conversion of *n*-paraffins to isoparaffins, olefins, and aromatics, and of naphthenes (cycloalkanes) to aromatics. The nature of the final product is influenced by the structure and composition of the straight-run (virgin) naphtha (hydrocarbon mixture) feedstock. In thermal reforming, the reactions resemble those in the cracking of gas oils. The molecular size is reduced, while olefins and some aromatics are synthesized. For example, an alkane can be converted to another alkane and an olefin [reaction (1), where $n > x + y$], or an alkane can be converted to a cycloalkane, which is in turn converted to an aromatic compound [reaction (2)]. Hydrocracking of high-molecular-weight paraffins to lower-molecular-weight paraffins and an olefin:
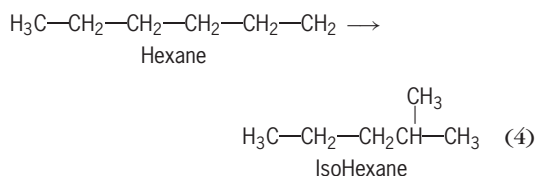
$$CH_3 - (CH_2)_n - CH_3 \longrightarrow$$

Alkane

$$CH_3(CH_2)_x - CH_3 + CH_3 - (CH_2)_y - CH=CH_2 \quad (1)$$

Alkane   Olefin

Dehydrocyclization of paraffin compounds to aromatic compounds:
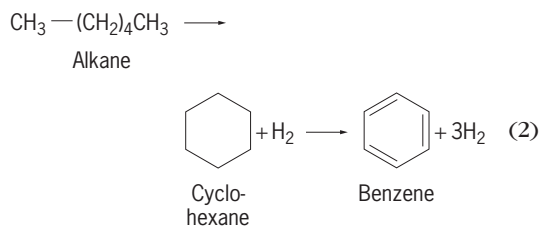
$$CH_3 - (CH_2)_4CH_3 \longrightarrow$$

Alkane

$$(2)$$

Cyclo-   Benzene
hexane

Hydrocracking of high-molecular-weight paraffins to lower-molecular-weight paraffins:

$$CH_3(CH_2)_{10}CH_3 \xrightarrow{H_2} 2CH_3(CH_2)_4CH_3$$

Dodecane   Hexane

$$\xrightarrow{H_2}$$

$$CH_3(CH_2)_6CH_3 + CH_3(CH_2)_2CH_3 \quad (3)$$

Octane   Butane
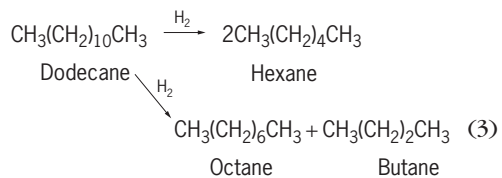
Isomerization of *n*-paraffins to isoparaffins:

$$H_3C-CH_2-CH_2-CH_2-CH_2-CH_2 \longrightarrow$$

Hexane

$$H_3C-CH_2-CH_2\overset{\overset{\displaystyle CH_3}{|}}{C}H-CH_3 \quad (4)$$

IsoHexane

Isomerization of methylcyclopentane to cyclohexane:

$$(5)$$

Methylcyclo-   Cyclo-
pentane   hexane
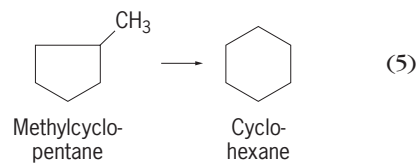
The liquid product (reformate) has the potential for further reaction by virtue of the presence of the olefinic constituents, but still contains appreciable quantities of these unsaturated materials and, as such, is classed as being unstable. Further hydrotreatment is required for stabilization. *See* PARAFFIN.

In the presence of catalysts and in the presence of the hydrogen available from dehydrogenation reactions, hydrocracking of paraffins to yield two lower-molecular-weight paraffins takes place, and olefins that do not undergo dehydrocyclization (dehydrogenation accompanied by formation of cyclic compounds) are dehydrogenated (with or without isomerization) so that the end product contains only traces of olefins. Examples of typical reactions are dehydrocyclization and dehydrogenation [reaction (2)], hydrocracking [reaction (3)], and isomerization [reactions (4) and (5)].

Other reactions such as the demethylation of branched paraffins to *n*-paraffins and the dealkylation of aromatics also occur. *See* DEHYDROGENATION; HYDROCRACKING; ISOMERIZATION.

Thermal reforming. Thermal reforming was a natural development from thermal cracking, since reforming is also a thermal decomposition reaction. Cracking converts heavier oils into gasoline constituents, whereas reforming converts (reforms) these gasoline constituents into higher-octane molecules. The equipment for thermal reforming is essentially the same as for thermal cracking, but higher temperatures are used.

In carrying out thermal reforming, a feedstock, such as a 205°C (400°F) end-point naphtha or a straight-run gasoline, is heated to 510–595°C (950–1100°F) in a furnace much the same as a cracking furnace, with pressures of 400–1000 lb/in.^{2} (2.8–6.9 megapascals). As the heated naphtha leaves the furnace, it is cooled or quenched by the addition of cold naphtha. The quenched, reformed material then enters a fractional distillation tower where any heavy products are separated. The remainder of the reformed material leaves the top of the tower and is separated into gases and reformate. The higher octane number of the product (reformate) is due primarily to the cracking of longer-chain paraffins into higher-octane olefins. *See* DISTILLATION.

The products of thermal reforming are gases, gasoline, and residual oil, the last formed in very small amounts (about 1%). The amount and quality of the reformate are very dependent on the temperature. As a rule, the higher the reforming temperature, the higher the octane number of the product but the lower the reformate yield. For example, a gasoline with an octane number of 35 when reformed at

515°C (960°F) yields 92.4% of 56-octane reformate; when reformed at 555°C (1030°F) the yield is 68.7% of 83-octane reformate. By using catalysts, as in the catalytic reforming processes described later, higher yields of much higher-octane gasoline can be obtained for a given temperature.

Thermal reforming is less effective and less economical than catalytic processes and has been largely supplanted. As it was practiced, a single-pass operation was used at 540–760°C (1000–1140°F) and 500–1000 lb/in.^{2} (3.4–6.9 MPa). The degree of octane-number improvement depended on the extent of conversion but was not directly proportional to the extent of cracking. The octane number was changed by the severity of the cracking, and the product had increased volatility, compared to the volatility of the feedstock. At high conversion, the production of coke and gas increased. The gases produced, although high in methane, generally contained olefins, and the process was generally accompanied by either a separate gas polymerization operation or one in which the C_3-C_4 gases, autogenous as well as extraneous, were added back to the reforming system.

Modifications of the thermal reforming process due to the inclusion of hydrocarbon gases with the feedstock are known as gas reversion and polyforming. Thus, olefinic gases produced by cracking and reforming can be converted into liquids boiling in the gasoline range by heating them under high pressure. Since the resulting liquids (polymers) have high octane numbers, they increase the overall quantity and quality of gasolines produced in a refinery.

The gases most susceptible to conversion to liquid products are olefins with three and four carbon atoms. These are propylene ($CH_3CH\!=\!CH_2$), which is associated with propane in the C_3 fraction, and butylene ($CH_3CH_2CH\!=\!CH_2$ or $CH_3CH\!=\!CHCH_3$) and isobutylene [$(CH_3)_2C\!=\!CH_2$], which are associated with butane ($CH_3CH_2CH_2CH_3$) and isobutane [$(CH_3)_2CHCH_3$] in the C_4 fraction. When the C_3 and C_4 fractions are subjected to the temperature and pressure conditions used in thermal reforming, they undergo chemical reactions that result in a small yield of gasoline. When the C_3 and C_4 fractions are passed through a thermal reformer as a mixture with naphtha, the process is called naphtha-gas reversion or naphtha polyforming.

Reversion and polyforming are essentially the same but differ in the manner in which the gases and naphtha are passed through the heating furnace. In gas reversion, the naphtha and gases flow through separate lines in the furnace and are heated independently of one another. Before leaving the furnace, both lines join to form a common soaking section where the reforming, polymerization, and other reactions take place. In naphtha reforming, the C_3 and C_4 gases are premixed with the naphtha and pass together through the furnace. Except for the gaseous components in the feedstock, both processes operate in much the same manner as thermal reforming and produce similar products.

Catalytic reforming. Like thermal reforming, catalytic reforming converts low-octane gasoline into high-octane gasoline (reformate). Although thermal reforming can produce reformate with a research octane number of 65–80 depending on the yield, catalytic reforming produces reformate with octane numbers on the order of 90–95. Catalytic reforming is conducted in the presence of hydrogen over hydrogenation-dehydrogenation catalysts, which may be supported on alumina or silica-alumina. Depending on the catalyst, a definite sequence of reactions takes place, involving structural changes in the feedstock. Furthermore, this process has rendered thermal reforming obsolete. *See* CATALYSIS.

Although subsequent olefin reactions occur in thermal reforming, the product contains appreciable amounts of unstable unsaturated compounds. In the presence of catalysts and hydrogen (available from dehydrogenation reactions), hydrocracking of paraffins to yield two lower paraffins occurs. Olefins that do not undergo dehydrocyclization are also produced. The olefins are hydrogenated with or without isomerization, so that the end product contains only traces of olefins.

The addition of a hydrogenation-dehydrogenation catalyst to the system yields a dual-function catalyst complex. Hydrogen reactions—hydrogenation, dehydrogenation, dehydrocyclization, and hydrocracking—take place on one catalyst, and cracking, isomerization, and olefin polymerization take place on the other (acid) catalyst. *See* HYDROCRACKING.

Under the high-hydrogen partial pressure conditions used in catalytic reforming, sulfur compounds are readily converted into hydrogen sulfide which, unless removed, builds up to a high concentration in the recycle gas. Hydrogen sulfide is a reversible poison for platinum catalysts and decreases the catalysts' dehydrogenation and dehydrocyclization activity. In the first catalytic reformers, hydrogen sulfide was removed from the gas cycle stream by absorption in diethanolamine. Sulfur is generally removed from the feedstock by use of a conventional desulfurization over cobalt-molybdenum catalyst. An additional benefit of desulfurization of the feed to a level of <5 ppm sulfur is the elimination of hydrogen sulfide (H_2S) corrosion problems in the heaters and reactors.

Organic nitrogen compounds are converted into ammonia under reforming conditions, and this neutralizes acid sites on the catalyst, thus repressing its activity for isomerization, hydrocracking, and dehydrocyclization reactions. Straight-run materials do not usually present serious problems with regard to nitrogen, but feeds such as coker naphtha may contain around 50 ppm nitrogen and removal of this quantity may require high-pressure hydrogenation (800–1000 lb/in.^{2}; 5.5–6.9 MPa) over a nickel-cobalt-molybdenum on alumina catalyst.

The commercial catalytic processes can be broadly classified as the moving-bed, fluid-bed, and fixedbed types. The fluid-bed and moving-bed processes use mixed nonprecious metal oxide catalysts in units equipped with separate regeneration facilities. Fixed-bed processes use predominantly
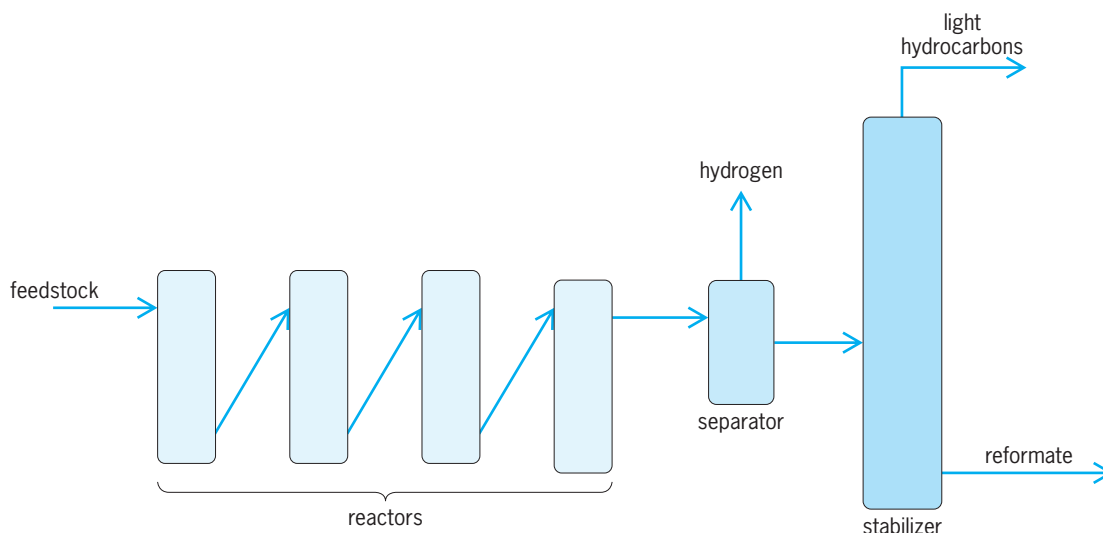
Catalytic reforming process.

platinum-containing catalysts in units equipped for cycle, occasional, or no regeneration.

Catalytic reformer feeds are saturated (not olefinic) materials. In most cases, the feed is a straight-run naphtha; however, other by-product low-octane naphtha (such as coker naphtha) can be processed after treatment to remove olefins and other contaminants. Hydrocarbon naphtha, which contains substantial quantities of naphthenes, is also a suitable feed.

The yield of gasoline of a given octane number and at given operating conditions depends on the hydrocarbon types in the feed. For example, high-naphthene stocks, which readily give aromatic gasoline, are the easiest to reform and give the highest gasoline yields. Paraffinic stocks which depend on the more difficult isomerization, dehydrocyclization, and hydrocracking reactions, require more severe conditions and give lower gasoline yields than the naphthenic stocks. The end points of these feeds are usually limited to about 190°C (375°F), partially because of increased coke deposition on the catalyst at the end point during processing at about 15°C (27°F). Limiting the feed end point avoids redistillation of the product to meet the gasoline end-point specification of 205°C (400°F), maximum.

Dehydrogenation is a main chemical reaction in catalytic reforming, and hydrogen gas is consequently produced in large quantities. The hydrogen is recycled through the reactors where the reforming takes place to provide the atmosphere necessary for the chemical reactions; and the hydrogen also prevents carbon from being deposited on the catalyst, thus extending its operating life. An excess of hydrogen above whatever is consumed in the process is produced, and as a result, catalytic reforming processes are unique in that they are the only petroleum refinery processes to produce hydrogen as a by-product. *See* DEHYDROGENATION; HYDRO-GENATION.

Catalytic reforming is usually carried out by feeding a naphtha (after pretreating with hydrogen if necessary) and hydrogen mixture to a furnace where the mixture is heated to 450–520°C (840–965°F) and then passed through fixed-bed catalytic reactors at hydrogen pressures of 100–1000 lb/in.^{2} (0.69–6.9 MPa). Normally two (or more) reactors are used in series, and reheaters are located between adjoining reactors to compensate for the endothermic reactions taking place. Sometimes as many as four or five are kept on-stream in series while one or more is being regenerated. The on-stream cycle of any one reactor may vary from several hours to many days, depending on the feedstock and reaction conditions (see **illus.**).

The product from the last catalytic reactor is cooled and sent to a high-pressure separator where the hydrogen-rich gas is split into two streams: one stream goes to recycle, and the other stream is excess hydrogen available for other uses. The excess hydrogen is vented from the unit and used in hydrotreating, as a fuel, or for the manufacture of chemicals (such as ammonia). The liquid product (reformate) is stabilized (by removal of light ends) and used directly in gasoline or extracted for aromatic blending stocks for aviation fuel. *See* AIRCRAFT FUEL.

Fixed-bed processes. Reforming processes that use catalysts arranged in beds are known as fixed-bed processes.

Hydroforming. The hydroforming process uses molybdena-alumina (MoO_2-Al_2O_3) catalyst pellets arranged in mixed beds. The hydroformer has four reaction vessels or catalyst cases, two of which are on the process cycle, while the other two are being regenerated. Naphtha feed is preheated to 400–540°C (900–1000°F) and passed in series through the two catalyst cases under a pressure of 150–300 lb/in.^{2} (1–2 MPa). Gas containing 70% hydrogen produced by the process is passed through the catalyst cases with the naphtha. The material leaving the final catalyst case enters a four-tower system where it is fractionally distilled as hydrogen-rich gas, a product (reformate) suitable for gasoline, and an aromatic polymer boiling above 205°C (400°F).

After 4–16 h on process cycle, the catalyst is regenerated. This is done by burning carbon deposits from the catalyst at 565°C (1050°F) by blowing air diluted with flue gas through the catalyst. The air also reoxidizes the reduced catalyst (9% molybdenum oxide on activated alumina pellets) and removes sulfur from the catalyst.

Platforming. The first step in the platforming process is preparation of the naphtha feed. For gasoline manufacture, the naphtha feed is distilled to separate a fraction boiling in the 120–205°C (250–400°F) range. Since sulfur adversely affects the platinum catalyst, the naphtha fraction may be treated to remove sulfur compounds. Otherwise, the hydrogen-rich gas produced by the process, which is cycled through the catalyst cases, must be scrubbed free of its hydrogen sulfide content.

The prepared naphtha feed is heated to 455–540°C (850–1000°F) and passed into a series of three catalyst cases under a pressure of 200–1000 lb/in.^{2} (1.4–6.9 MPa). Further heat is added to the naphtha between each of the catalyst cases in the series. The material from the final case is fractionated into hydrogen-rich gas and reformate. The catalyst is composed of 1/8-in. (0.3-cm) pellets of alumina containing chlorine and about 0.5% platinum. Each pound of catalyst reforms up to 100 barrels (16 m^3) of naphtha before losing its activity. It is possible to regenerate the catalyst, but it is more usual to replace the spent catalyst with new catalyst.

Other fixed-bed processes include catforming, in which the catalyst is a platinum, alumina, silica-alumina composition, which results in very high hydrogen purity. Regeneration to prolong catalyst life is practiced on a blockout basis (one of several reactions is taken off stream) with a dilute air in-stream mixture. In addition, Houdriforming is a process in which the catalyst may be regenerated, if necessary, on a blockout basis. A guard case catalytic hydrogenation pretreating stage using the same Houdry catalyst as the Houdriformer reactors is available for high-sulfur feedstocks. Lead and copper salts are also removed under the mild conditions of the guard case operation.

Powerforming. The cyclic Powerforming process is based on frequent catalyst regeneration (carbon burn-off) and permits continuous operation. Reforming takes place in four or five reactors, and regeneration is carried out in the last, or swing, reactor. Thus the plant need not be shut down to regenerate a catalyst reactor. The cyclic process assures a continuous supply of hydrogen gas for hydrorefining operations and tends to produce a greater yield of higher-octane reformate. The choice between the semiregenerative process and the cyclic process depends on the size of plant required, the type of feedstocks available, and the octane number needed in the product.

Iso-Plus Houdriforming. This is a combination process using a conventional Houdriformer operated at moderate severity, in conjunction with one of three alternatives: (1) conventional catalytic reforming plus aromatic extraction and separate catalytic reforming of the aromatic product; (2) conventional catalytic reforming plus aromatic extraction and recycling of the aromatic product aligned to the reforming state; and (3) conventional catalytic reforming followed by thermal reforming of the Houdriformer product and catalytic polymerization of the C$_3$ and C$_4$ olefins from thermal reforming. A typical feedstock for this type of unit is naphtha, and the use of a Houdry guard case permits charging stocks of relatively high sulfur content.

Rexforming. Rexforming is a combination process using platforming and aromatic extraction processes in which low-octane raffinate is recycled to the platformer. Operating temperatures may be as much as 27°C (50°F) lower than conventional platforming, and higher space velocities (rates of feedstock introduction into reactor) are used. A balance is struck between hydrocyclization and hydrocracking, excessive coke and gas formation thus being avoided. The glycol solvent in the aromatic extraction section is designed to extract low-boiling high-octane isoparaffins as well as aromatics.

Fixed-bed swing reactor. Fixed-bed, continuous catalytic reforming can be classified by catalyst type: (1) cyclical regenerative with nonprecious metal oxide catalysts and (2) cyclic regenerative with platinum-alumina catalysts. Both types use swing reactors to regenerate a portion of the catalyst while the remainder stays on-stream.

The cyclic regenerative fixed-bed, operation, using a platinum catalyst, is basically a low-pressure process (250–350 lb/in.^{2}; 1.7–2.4 MPa) which gives higher gasoline yields because of fewer hydrocracking reactions, as well as higher-octane products from a given naphtha charge and better hydrogen yields because of more dehydrogenation and fewer hydrocracking reactions. The coke yield, with attendant catalyst deactivation, increases rapidly at low pressures.

Selectoforming. The Selectoforming process uses a fixed-bed reactor operating under a hydrogen partial pressure. Typical operating conditions depend on the process configuration but are in the ranges 200–600 lb/in.^{2} (1.4–4.1 MPa) and 315–450°C (600–900°F). The catalyst used in the Selectoforming process is nonnoble metal with a low potassium content. As with the large-pore hydrocracking catalysts, the cracking activity increases with decreasing alkali metal content.

There are two configurations of the Selectoforming process that are being used commercially. The first Selectoformer was designed as a separate system and integrated with the reformer only to the extent of having a common hydrogen system. The reformer naphtha is mixed with hydrogen and passed into the reactor containing the shape-selective catalyst. The reactor effluent is cooled and separated into hydrogen, liquid petroleum, gas, and high-octane gasoline. The removal of *n*-paraffins reduces the vapor pressure of the reformate since these paraffins are in higher concentration in the front end of the feed. The separate Selectoforming system has the additional flexibility of being able to process other refinery streams.

The second process modification is the terminal reactor system. In this system, the shape-selective catalysts replace all or part of the reforming catalyst in the last reforming reactor. Although this configuration is more flexible, the high reforming operating temperature causes butane and propane cracking and consequently decreases the liquid petroleum gas yield and generates higher ethane and methane production. The life of a Selectoforming catalyst used in a terminal system is between 2 and 3 years, and regeneration only partially restores its catalytic activity.

Moving-bed processes. In moving-bed processes, the catalyst moves downward through the reactor by gravity flow and returns to the top by means of a solids-conveying technique (hyperflow)—this moves the catalyst at a low velocity and with minimum loss.

Hyperforming. Hyperforming is a moving-bed reforming process that uses catalyst pellets of cobalt molybdate with a silica-stabilized alumina base. Feedstock (naphtha vapor) and recycle gas flow upward, countercurrent to the catalyst, and regeneration of catalyst is accomplished in either an external vertical lift line or a separate vessel.

Hyperforming naphtha (65–230°C; 150–450°F) can result in improvement of the gasoline component; in addition, sulfur and nitrogen removal is accomplished. Light gas oil stocks can also be charged to remove sulfur and nitrogen under mild hydrogenation conditions for the production of premium diesel fuels and middle distillates. Operating conditions in the reactor are 400 lb/in.^{2} (2.8 MPa) and 425–480°C (800–900°F), the higher temperature being used for a straight-run naphtha feedstock; catalyst regeneration takes place at 510°C (950°F) and 415 lb/in.^{2} (2.9 MPa).

Thermofor catalytic reforming (TCR). This is a moving-bed process that uses a synthetic bead, co-precipitated chromia (CrO$_2$) and alumina (Al$_2$O$_3$) catalyst. Catalyst-to-naphtha ratios have little effect on product yield or quality when varied over a wide range. The catalyst flows downward through the reactor, and the naphtha-recycle gas feed enters the center of the reactor. The catalyst is transported from the base of the reactor to the top of the regenerator by bucket-type elevators.

Fluid-bed processes. In catalytic reforming processes using a fluidized bed, the catalyst is continuously regenerated with a separate or integrated reactor to maintain catalyst activity by coke and sulfur removal. Cracked or virgin naphtha is charged with hydrogen-rich recycle gas to the reactor. A molybdena (Mo$_2$O$_3$, 10.0%) on alumina catalyst, not materially affected by normal amounts of arsenic, iron, nitrogen, or sulfur, is used. Operating conditions in the reactor are about 200–300 lb/in.^{2} (1.4–2.0 MPa) and 480–950°C (900–950°F). *See* FLUIDIZATION.

Fluidized-bed operation with its attendant excellent temperature control prevents over- and under-reforming operations, resulting in more selectivity in the conditions needed for optimum yield of the desired product.

Catalysts. Reforming reactions typically proceed through a number of elementary steps. For example, a straight-chain paraffin is converted into an isoparaffin by first being converted into an olefin, which is isomerized to an isoolefin, and then converted into an isoparaffin. Correspondingly, the catalyst has two functions: a hydrogenation-dehydrogenation function for the paraffin-olefin conversions and an isomerization function, which is associated with the catalyst acidity. The catalysts used until the early 1950s were chromium oxide or molybdenum oxide supported on alumina, which incorporated both the catalytic functions on the surface of the metal oxide. More recently developed reforming catalysts have crystallites of a metal such as platinum on an acidic support such as alumina, and the two functions are present in separate phases. The metal (palladium, Pd; platinum, Pt; or rhenium, Re; or a noble-metal-containing trimetallic alloy) provides the hydrogenation-dehydrogenation activity, and the promoted acidic alumina provides the isomerization activity. The hydrogenation-dehydrogenation activity of the supported metal and the isomerization activity of the alumina are much greater than the respective activities of the early-generation metal oxides.

The composition of a reforming catalyst is dictated by the composition of the feedstock and the desired reformate. The catalysts used are principally molybdena-alumina (MoO$_2$-Al$_2$O$_3$), chromia-alumina (Cr$_2$O$_3$-Al$_2$O$_3$), or platinum on a silica-alumina (SiO$_2$-Al$_2$O$_3$) or alumina (Al$_2$O$_3$) base. The nonplatinum catalysts are widely used in regenerative process for feeds containing sulfur, which poisons platinum catalysts, although pretreatment processes (such as hydrodesulfurization) may permit platinum catalysts to be used.

The purpose of platinum on the catalyst is to promote dehydrogenation and hydrogenation reactions, that is, the production of aromatics, participation in hydrocracking, and rapid hydrogenation of carbon-forming precursors. For the catalyst to have an activity for isomerization of both paraffins and naphthenes—the initial cracking step of hydrocracking—and to participate in paraffin dehydrocyclization, it must have an acid activity. The balance between these two activities is most important in a reforming catalyst.

In the production of aromatics from cyclic saturated materials (naphthenes), it is important that hydrocracking be minimized to avoid loss of the desired product. Thus, the catalytic activity must be moderated relative to the case of gasoline production from a paraffinic feed, where dehydrocyclization and hydrocracking play an important part.

The acid activity can be obtained by means of halogens (usually fluorine or chlorine up to about 1% by weight in catalyst) or silica incorporated in the alumina base. The platinum content of the catalyst is normally in the range 0.3–0.8 wt %. At higher levels there is some tendency to effect demethylation and naphthene ring opening, which is undesirable; at lower levels the catalysts tend to be less resistant to poisons.

Most processes have a means of regenerating the catalyst as needed. The time between regeneration, which varies with the process, the severity of the reforming reactions, and the impurities of the feedstock, ranges from every few hours to several months. Several processes use a nonregenerative catalyst that can be used for a year or more, after which it is returned to the catalyst manufacturer for reprocessing. In processes that have moving beds, the catalysts are continuously regenerated in separate reactors.

The processes using bauxite (cycloversion) and clay (isoforming) differ from other catalytic reforming processes in that hydrogen is not formed and hence none is recycled through the reactors. Since hydrogen is not concerned in the reforming reactions, there is no limit to the amount of olefin that may be present in the feedstock. The cycloversion process is also used as a catalytic cracking process and as a desulfurization process. The isoforming process causes only a moderate increase in octane number.

The modern reforming catalysts in use today contain platinum supported on a silica or silica-aluminum base. In most cases, rhenium is combined with platinum to form a more stable catalyst that permits operation at lower pressures. Platinum is thought to serve as the catalytic site for hydrogenation and dehydrogenation reactions, and chlorinated alumina provides the acidic site for isomerization, cyclization, and hydrocracking reactions. Reforming catalyst activity is a function of surface area, pore volume, and active platinum and chlorine content. Catalyst activity is reduced during operation by coke deposition and chloride loss. In a high-pressure process, up to 20 barrels (3.2 m^3) of charge can be processed per pound of catalyst before regeneration is needed. The activity of the catalyst can be restored by high-temperature oxidation of the carbon followed by chlorination. This type of process is referred to as semiregenerative and is able to operate for 6- to 24-month periods between regenerations. The activity of the catalyst decreases during the on-stream period and the reaction temperature is increased as the catalyst ages to maintain the desired operating severity. Normally the catalyst can be regenerated at least three times before it has to be replaced and returned to the manufacturer for reclamation. *See* HETEROGENEOUS CATALYSIS; PETROLEUM PROCESSING AND REFINING. James G. Speight

Bibliography. G. J. Antos, A. M. Aitani, and J. M. Parera (eds.), *Catalytic Naphtha Reforming: Science and Technology*, Marcel Dekker, New York, 1995; J. G. Speight, *The Chemistry and Technology of Petroleum*, 3d ed., Marcel Dekker, New York, 1999.

Refraction of waves

The change of direction of propagation of any wave phenomenon which occurs when the wave velocity changes. The term is most frequently applied to visible light, but it also applies to all other electromagnetic waves, as well as to sound and water waves.
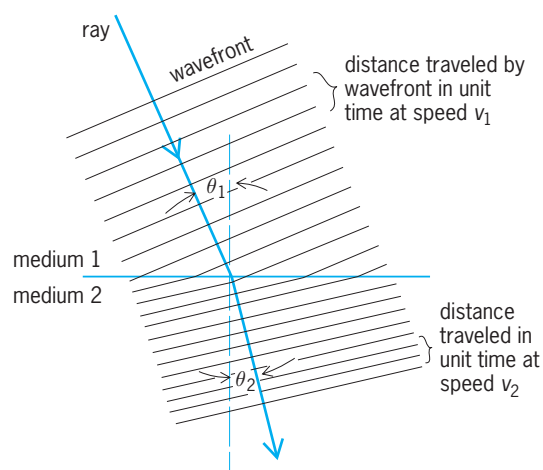

Fig. 1. Physical basis for Snell's law.

The physical basis for refraction can be readily understood with the aid of **Fig. 1**. Consider a succession of equally spaced wavefronts approaching a boundary surface obliquely. The direction of propagation is in ordinary cases perpendicular to the wavefronts. In the case shown, the velocity of propagation is less in medium 2 than in medium 1, so that the waves are slowed down as they enter the second medium. Thus the direction of travel is bent toward the perpendicular to the boundary surface (that is, $\theta_2 < \theta_1$). If the waves enter a medium in which the velocity of propagation is faster than in their original medium, they are refracted away from the normal.

Snell's law. The simple mathematical relation governing refraction is known as Snell's law. If waves traveling through a medium at speed v_1 are incident on a boundary surface at angle θ_1 (with respect to the normal), and after refraction enter the second medium at angle θ_2 (with the normal) while traveling at speed v_2, then Eq. (1) holds. The index of

$$\frac{v_1}{v_2} = \frac{\sin\theta_1}{\sin\theta_2} \qquad (1)$$

refraction n of a medium is defined as the ratio of the speed of waves in vacuum c to their speed in the medium. Thus $c = n_1 v_1 = n_2 v_2$, and therefore Eq. (2)

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \qquad (2)$$

holds. The refracted ray, the normal to the surface, and the incident ray always lie in the same plane.

The relative index of refraction of medium 2 with respect to that of medium 1 may be defined as $n = n_1/n_2$. Snell's law then becomes Eq. (3). For

$$\sin\theta_1 = n \sin\theta_2 \qquad (3)$$

sound and other elastic waves requiring a medium in which to propagate, only this last form has meaning. Equation (3) is frequently used for light when one medium is air, whose index of refraction is very nearly unity.

When the wave travels from a region of low velocity (high index) to one of high velocity (low index), refraction occurs only if $(n_1/n_2) \sin \theta_1 \leqq 1$. If θ_1 is too large for this relation to hold, then $\sin \theta_2 > 1$, which is meaningless. In this case the waves are totally reflected from the surface back into the first medium. The largest value that θ_1 can have without total internal reflection taking place is known as the critical angle θ_c. Thus $\sin \theta_c = n_2/n_1$. When the angle of incidence $\theta_1 < \theta_c$, refraction occurs, as in **Fig. 2a**. When $\theta_1 = \theta_c$, the emergent ray just grazes the surface (Fig. 2b). Total internal reflection (Fig. 2c) represents the only practical case for which 100% of the incident energy is reflected and none is absorbed. When it is desired to change the direction of a beam of light without loss of energy, totally reflecting prisms are often used, as in prism binoculars.

If waves travel through a medium having a continuously varying index of refraction, the rays follow smooth curves with no abrupt changes of direction. Suppose (**Fig. 3**) that $n = n(y)$, and that the incident ray lies in the xy plane. If θ is the angle between the direction of the ray and the y axis, then Snell's law can be written in the differential form given by Eq. (4). In a particular case Eq. (4) can be integrated

$$\frac{d\theta}{dn} = -\frac{1}{n} \tan \theta \qquad (4)$$

to give the path of the ray.

Visible light. Many interesting cases of refraction occur for visible light. The refraction of light by a prism in air affords a particularly simple and useful example. *See* LIGHT.

As a ray passes through the prism of **Fig. 4**, its total deflection or deviation $D = \theta_1 + \theta_4 - A$, where A is the vertex angle of the prism. Also, by Snell's law, Eq. (5) holds. It is found that the deviation is a minimum when the ray passes through the prism symmetrically (that is, when $\theta_1 = \theta_4$). For minimum deviation, Eq. (6) holds. For a given prism, the dispersion,

$$n = \frac{\sin \theta_1}{\sin \theta_2} = \frac{\sin \theta_4}{\sin \theta_3} \qquad (5)$$

imum when the ray passes through the prism symmetrically (that is, when $\theta_1 = \theta_4$). For minimum deviation, Eq. (6) holds. For a given prism, the dispersion,

$$n = \frac{\sin \frac{1}{2}(A + D)}{\sin \frac{1}{2}A} \qquad (6)$$

or lateral spread of the spectrum formed is maximum for that wavelength of light which passes through the prism at minimum deviation. *See* OPTICAL PRISM.

For most optical materials, the dispersion $dn/d\lambda$ (λ is the wavelength) is negative; therefore red light is bent less than blue light. Typical values of n for optical materials range from 1.5 for ordinary crown glass, 1.7 or 1.8 for dense flint glass, and up to 2.42 for diamond. For water, n is 1.33. Some special substances have even higher values. Many substances show anisotropy in the refraction of light, with different indices of refraction in different directions. *See* OPTICAL MATERIALS.

For a lens (**Fig. 5**), refraction occurs at both surfaces. If the lens is thin and the rays all make small angles with the axis of the system, application of
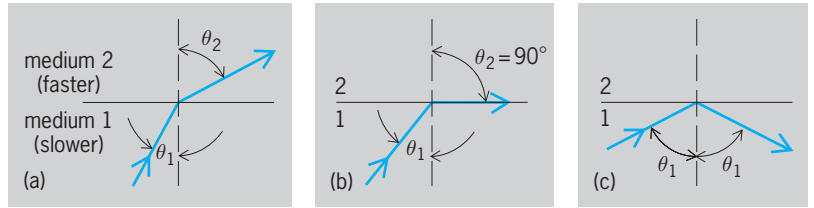

Fig. 2. Behavior of ray traveling from medium of high refractive index to medium of low refractive index. (*a*) When $\theta_1 < \theta_c$, ray is refracted. (*b*) When $\theta_1 = \theta_c$, ray grazes surface. (*c*) When $\theta_1 > \theta_c$, ray is reflected.

Snell's law to the two spherical surfaces yields the well-known lens formula, Eq. (7), where s is the ob-

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \qquad (7)$$

ject distance from the lens, s' the image distance, and f the focal length of the lens. Magnifying instruments such as binoculars, telescopes, microscopes, and projectors make use of refraction by lenses or prisms in their operation. *See* LENS (OPTICS).

Double refraction. Some anisotropic single crystals such as those of calcite and quartz are birefringent, or doubly refracting. If one looks through such a crystal at a dot on a piece of paper, two images are seen. As the crystal is rotated in the plane of the paper, one image remains stationary while the other appears to rotate about it. *See* BIREFRINGENCE.

Two separate rays propagate through the crystal; they are called the ordinary ray and the extraordinary ray. These rays are linearly polarized at right angles to each other. The ordinary ray obeys Snell's
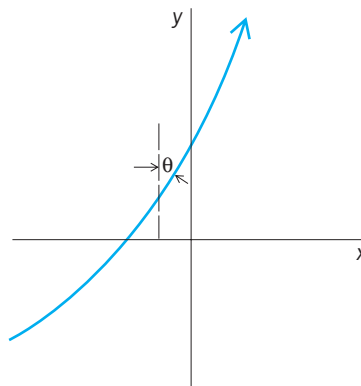

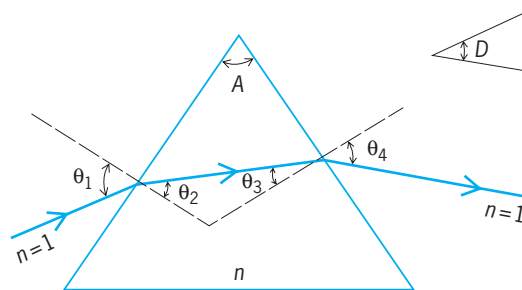**Fig. 3.** Path of light in medium having continuously varying index of refraction $n = n(y)$.
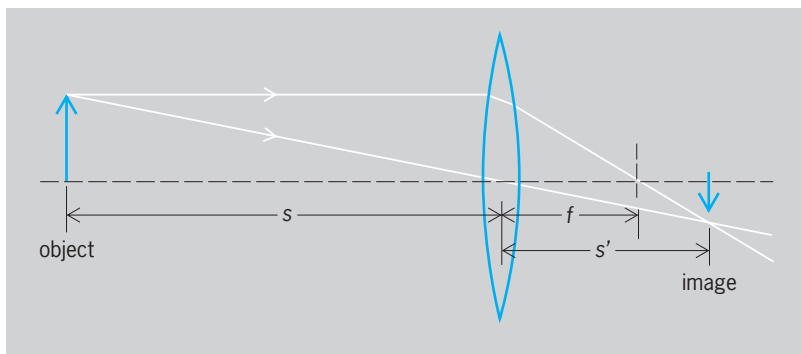

Fig. 4. Refraction of light by a prism.

Fig. 5. Refraction of light by a lens.

law; the extraordinary ray in general does not. The extraordinary ray does not propagate perpendicularly to its wavefronts. The separation between the two rays depends upon the direction in which the light travels through the crystal relative to that of the optic axis of the crystal. Light traveling parallel to the optic axis is only singly refracted.

Birefringent crystals are either uniaxial or biaxial, depending upon whether they have one optic axis or two. They are said to be positive or negative, depending upon whether the velocity of propagation (within the crystal) of the extraordinary wave is greater or less than that of the ordinary wave. Calcite is a uniaxial negative crystal, quartz a uniaxial positive crystal. The most commonly used biaxial crystal is mica. Doubly refracting crystals are frequently employed as polarizers, such as the nicol prism. *See* CRYSTAL OPTICS; POLARIZED LIGHT.

Refractometry. The measurement of indices of refraction, called refractometry, can be made in several ways. A very accurate technique is to determine in a prism spectrometer the minimum deviation D for a prism made from the material in question. The value of n is then calculated from Eq. (6). Hollow prisms can be used in this manner to determine the values of indices of refraction of various liquids. Alternatively, the critical angle for total internal reflection may be measured. Another method is to observe visually the apparent thickness of a slab of material by looking straight through it, and to compare the apparent thickness with the real thickness as measured with a micrometer. Then Eq. (8) holds.

$$n = \frac{\text{real thickness}}{\text{apparent thickness}} \qquad (8)$$

Interferometric methods are particularly convenient for gases. In the Jamin refractometer, for example, a simple count of fringes as the gas is slowly admitted to an initially evacuated tube in the optical path yields n. These techniques can also be used for solids, particularly when the material is available in the form of thin films. *See* INTERFEROMETRY.

Refractometry is an important tool in analytical chemistry. For example, information about the composition of an unknown solution can frequently be obtained by measurement of its index of refraction. *See* REFRACTOMETRIC ANALYSIS.

Atmospheric refraction. Gases have indices of refraction only slightly greater than unity. In general, $n - 1$ is proportional to the density of the gas, or to the ratio of pressure to absolute temperature. The index of refraction of the Earth's atmosphere increases continuously from 1.000000 at the edge of space to 1.000293 (yellow light) at $32°$F ($0°$C) and 760 mmHg (101.325 kilopascals) pressure. Thus celestial bodies as seen in the sky are actually nearer to the horizon than they appear to be. The effect decreases from a maximum of about 35 minutes of arc for an object on the horizon to zero at the zenith, where the light enters the atmosphere at perpendicular incidence. Thus the Sun (and all other bodies) appears to rise 2 or more minutes earlier (depending upon latitude) and to set 2 or more minutes later than would be the case without refraction. This must be taken into account when the altitude of a celestial body is observed for navigational purposes.

Other manifestations of atmospheric refraction are the mirages and "looming" of distant objects which occur over oceans or deserts, where the vertical density gradient of the air is quite uniform over a large area. The twinkling of stars is caused by the rapid small fluctuations in density along the light path in the atmosphere. Rainbows are produced by the multiple reflections, refraction, and dispersion of sunlight by spherical raindrops. *See* METEOROLOGICAL OPTICS; MIRAGE; RAINBOW; TWINKLING STARS.

Other electromagnetic waves. Although refraction is most frequently encountered for the visible portion of the spectrum, it is of importance for other electromagnetic radiation. For very long-wavelength radiation, the index of refraction of many materials is equal to the square root of the relative permittivity, ϵ_r. In general, $dn/d\lambda$ is negative except in the regions of so-called anomalous dispersion near absorption bands. On the short-wavelength side of an absorption band, n can be less than 1.00. Since it is the phase velocity of the wave rather than the group velocity which is involved in the definition of the index of refraction, this does not represent a violation of the principle of relativity, that is, that energy cannot be propagated at a velocity faster than the velocity of light in vacuum. At very high frequencies, the index of refraction of all materials is also slightly less than unity. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; GROUP VELOCITY; PERMITTIVITY; PHASE VELOCITY; RELATIVITY.

Refraction plays a role in the propagation beyond the line of sight of radio waves in the Earth's atmosphere. *See* RADIO-WAVE PROPAGATION.

The interaction of electromagnetic radiation with more or less opaque substances is often described in terms of a complex index of refraction. The real part of this quantity has the usual meaning for the small amount of light which penetrates into the material before it is absorbed. The imaginary part is a measure of the absorption.

Sound waves. The velocity of sound in a gas is proportional to the square root of the absolute temperature. Because of the vertical temperature gradients in the atmosphere, refraction of sound can be quite

pronounced. As in mirage formation, to allow large-scale refraction the temperature at a given height must be uniform over a rather large horizontal area. If the temperature decreases with altitude (the usual situation), sound waves initially traveling at a small angle with the horizontal are refracted upward. A sound out of doors is thus not normally audible at a great distance. However, if there is a temperature inversion (as over a body of water on a calm sunny day), the waves would be refracted downward. This is the main reason that sound carries long distances across water on a calm day. On a windy day the horizontal temperature strata are broken up and the sound is dissipated.

Refraction accompanied by reflection accounts for the fact that large explosions are sometimes heard in several distinct regions at surprisingly large distances, with zones of silence in between. A temperature inversion at high levels can refract the waves downward into a zone of audibility. The sound is then reflected from the ground, and must again be refracted downward to give the next zone of audibility. *See* SOUND; WAVE MOTION IN FLUIDS.

Seismic waves. The velocity of elastic waves in a solid depends upon the modulus of elasticity and upon the density of the material. Waves propagating through solid earth are refracted by changes of material or changes of density. Worldwide observations of earthquake waves enable scientists to draw conclusions on the distribution of density within the Earth. These waves may be totally internally reflected at the boundary of the core. It was through such observations that the existence of the much denser core of the Earth was first postulated.

Refraction of compressional waves from explosions set off on the ground is (combined with reflection) used in prospecting for oil, natural gas, and minerals which have large differences in density and elastic constants from the surrounding rocks. *See* SEISMOLOGY.

Water waves. The speed of water waves in shallow water is proportional to the square root of the depth. As the waves enter shallower water they travel more slowly. As a train of waves approaches a coastline obliquely, its direction of travel becomes more nearly perpendicular to the shore because of refraction. *See* GEOMETRICAL OPTICS; NEARSHORE PROCESSES; WAVE MOTION IN LIQUIDS.   John W. Stewart

Bibliography. M. Born and E. Wolf, *Principles of Optics*, 7th ed., 1999; B. D. Guenther, *Modern Optics*, 1990; D. Halliday and R. Resnick, *Physics*, 4th ed., 1992; F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed., 1976; J. Meyer-Arendt, *Introduction to Classical and Modern Optics*, 4th ed., 1995.

Refractometric analysis

A method of chemical analysis based on the measurement of the index of refraction of a substance. As shown in the **illustration**, when light impinges on the surface of a material at an angle i to the normal to the surface, its direction is changed on passing into the material so that it then travels at an angle r to the normal. The index of refraction is defined in the equation below. It varies as a function

$$\text{Index of refraction} = \frac{\sin i}{\sin r} = n_D^{20}$$

of temperature and wavelength of light, and also of pressure in gases. Refractive indices are usually measured at 20°C (68°F), using the yellow D line of the sodium spectrum. The indices of refraction of a few substances are water, 1.333; benzene, 1.5014; chloroform, 1.4464; and acetone, 1.3589.

The most common type of refractometer is the Abbe refractometer. It is simple to use, requiring but a drop or two of sample and allowing a measurement of refractive index to be made in 1–2 min, with a precision of 0.0001. More precise measurements of refractive indices, within 0.00003, may be made by using a dipping or immersion refractometer, the prism of which is completely immersed in the sample. This requires about 15 ml of sample and is widely used to detect trace impurities or to control the quality of a product. The most precise measurements of the refractive indices of gases or solutions containing small traces of impurities are made with an interferometer, based on the interference of light. Its precision is 0.000001. For measurements in flowing systems, differential refractometers, capable of detecting a difference of 10^{-7} refractive index unit, are used.

The measurement of refractive index is used to identify compounds whose other physical constants are quite similar. Because minute amounts of impurities often cause a measurable change in the refractive index of a pure material, refractive index is often used as a criterion for purity. A measurement
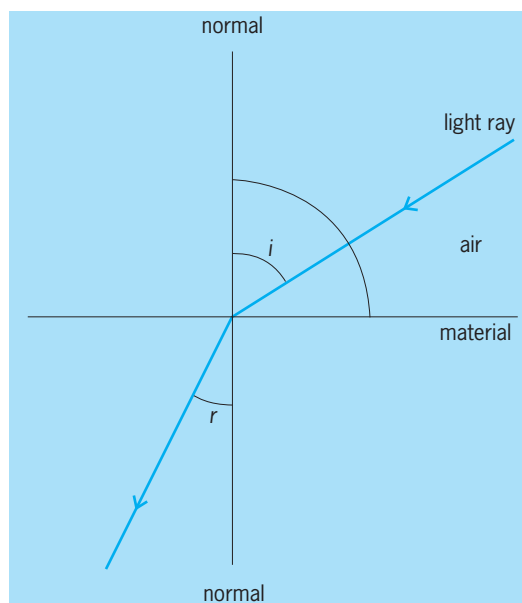
Refraction of light. The light ray changes direction as it passes into the material from the air.

of refractive index gives information as to the gross amount of impurity; it does not serve to identify the impurity. To give qualitative information, measurements would have to be made at different wavelengths, a rare procedure. In systems containing only two components, such as water and alcohol mixtures or aqueous salt solutions, refractive index is a sensitive and rapid method of determining the composition. Refractometric analysis has also been used for the determination of the water content in milk, the amount of sulfur in rubber, and the isotopic analysis of D_2O–H_2O mixtures. *See* REFRACTION OF WAVES.

Robert F. Goddu; James N. Little

Bibliography. J. D. Winefordner, *Treatise on Analytical Chemistry*, 2d ed., 1989.

Refractory

One of a number of ceramic materials for use in high-temperature structures or equipment. The term high temperatures is somewhat indefinite but usually means above about 1800°F (1000°C), or temperatures at which, because of melting or oxidation, the common metals cannot be used. In some special high-temperature applications, the so-called refractory metals such as tungsten, molybdenum, and tantalum are used.

The greatest use of refractories is in the steel industry, where they are used for construction of linings of equipment such as blast furnaces, hot stoves, and open-hearth furnaces. Other important uses of refractories are for cement kilns, glass tanks, nonferrous metallurgical furnaces, ceramic kilns, steam boilers, and paper plants. Special types of refractories are used in rockets, jets, and nuclear power plants. Many refractory materials, such as aluminum oxide and silicon carbide, are also very hard and are used as abrasives; some applications, for example, aircraft brake linings, use both characteristics. *See* ABRASIVE.

Refractory materials are commonly grouped into (1) those containing mainly aluminosilicates; (2) those made predominately of silica; (3) those made of magnesite, dolomite, or chrome ore, termed basic refractories (because of their chemical behavior); and (4) a miscellaneous category usually referred to as special refractories. *See* METAL COATINGS.

Aluminosilicate refractories. Fireclay is the raw material from which the bulk (about 70%) of refractories is manufactured. Different grades are distinguished according to the softening temperature or the pyrometric cone equivalent (PCE), the number of the standard pyrometric cone which deforms under heat treatment in the same manner as the fireclay. Thus, the minimum PCEs for low, intermediate, high, and superduty fireclays are 19, 29, 31/32, and 33, respectively. Fireclays are also classified by their working properties into two other classes: plastic (those which form a moldable mass when mixed with water), and flint (a hard, rocklike clay that does not become plastic when mixed with water). In general, flint clays have higher PCEs than plastic clays and are mixed with them to form higher-grade

fireclay brick. *See* CLAY, COMMERCIAL.

High alumina refractories are made from clays which contain, in addition to the alumina (Al_2O_3) in the clay minerals, hydrates of aluminum oxide. These raise the total Al_2O_3 content and make the material more refractory. Different grades are distinguished on the basis of the total Al_2O_3 content (50, 60, 70% alumina refractories).

Sillimanite and kyanite are anhydrous aluminosilicate minerals used to make special refractory objects, such as crucibles, tubes, and muffles, or as an addition to fireclay to increase refractoriness and to control its shrinkage during firing.

Silica refractories. These account for about 15% of total production. They are made from crushed and ground quartzite (ganister) to which about 2% lime has been added to assist in bonding, both before and after firing. The quality of silica refractories is to a great extent determined by the amount of Al_2O_3 impurity, even small amounts having a deleterious effect on refractoriness. This is just opposite to the case of alumina in fireclays, where a higher alumina content means greater refractoriness. High-grade silica brick contains less than 0.6% Al_2O_3, and even the standard grade contains less than 1%. During firing, the mineral quartz transforms to cristobalite and tridymite, the high-temperature forms of silica. Since these are less dense than quartz, the brick expands on firing and the true density of the solid is often taken as a test of adequate firing. An example would be the case in which the density of acceptably fired material must be below 1.36 oz/in³ (2.35 g/cm³). The outstanding characteristic of silica is its ability to withstand high loads at elevated temperatures, for example, as a sprung-arch roof 30 to 40 ft (9–12 m) wide over an open hearth. The hearth may be operated within 120°F (50°C) of the melting point of silica. *See* SILICON.

Semisilica refractories are made from clay with a high silica (sand) content (over 70% total silica); their main advantage is their dimensional stability when heated, or fired. Apparently the expansion of the silica, as sand, offsets the contraction of the clay.

Basic refractories. Magnesite refractories are so named because magnesium carbonate mineral was for many years the sole raw material. Since World War II seawater has become a significant source of magnesium oxide refractory, and such material is often called seawater magnesite. In any case, the raw material is calcined to form a material largely magnesium oxide, MgO; about 5% iron oxide is usually added before calcining. *See* MAGNESITE.

Chrome refractories are made from chrome ore, a complex mineral containing oxides of chromium, iron, magnesium, aluminum, and other oxides crystallized in the spinel structure. These crystals are usually embedded in a less refractory matrix called gangue. *See* CHROMITE.

In an attempt to combine the best properties of each, magnesite and chrome are often mixed to form chrome-magnesite or magnesite-chrome refractories (the first-named is the dominant constituent).

Dolomite is a mixed calcium-magnesium carbonate, $CaMg(CO_3)_2$, which, when calcined to a

mixture of MgO and CaO, is used in granular form to patch the bottoms of open hearths and also to make bricks.

Miscellaneous materials. Special refractories are made of a great many materials, and it is possible to mention here only a few of the more important.

Silicon carbide, SiC, is used for many refractory shapes, its outstanding properties being good thermal and electrical conductivity (it is used to make electric heating elements for furnaces), good heat-shock resistance, strength at high temperatures, and abrasion resistance. The first silicon carbide refractories were bonded with clay, so that the refractory properties of the bond placed the ultimate limit on the material. A method of making self-bonded silicon carbide has been developed to remove this limitation. Although silicon carbide tends to oxidize to form SiO_2 and either CO or CO_2, the silica-oxidation product forms a glassy coating on the remaining material and to a certain extent protects it from further oxidation.

Insulating firebrick is made from refractory clays to which a combustible material (sawdust, cork, coal) has been added; when this burns during the firing operation, it leaves a brick of high porosity. The low thermal conductivity of insulating brick reduces heat losses from furnaces, and the low bulk density and consequent low heat capacity reduce the amount of heat needed to bring the furnace itself up to temperature. The main disadvantage of such bricks is their low strength, but even this is useful in that they can be cut or ground to shape quite readily.

Pure oxides, of which alumina, Al_2O_3, is the prime example, are used for many special refractories. Zircon, $ZrSiO_4$, and zirconia, ZrO_2, are finding increased and significant uses as refractory materials. Some, such as beryllia, BeO, thoria, ThO_2, and uranium oxide, UO_2, are of particular interest for nuclear applications.

Carbides, nitrides, borides, silicides, and sulfides of various sorts have been considered as refractory materials, and some study made of them; aside from a few carbides and nitrides, however, none have found much use.

Cermets are an intimate mixture of a metal and a nonmetal, for example, Al_2O_3 and chromium. Although the nonmetal may be an oxide, it is more commonly a carbide or nitride (as in cemented tungsten carbide). *See* CERMET.

Carbon, generally in the form of graphite, is used for such equipment as crucibles and as stopper nozzles in ladles for steel casting. A potentially very large use of carbon is in blocks for construction of blast-furnace hearths. Graphite has very good thermal-shock resistance and moderate electrical conductivity, does not melt but rather sublimes at a significant rate only at temperatures well above 5400°F (3000°C), is quite inert chemically, and is wet by very few molten materials. The main disadvantage of graphite, common to all nonoxide materials at high temperatures, is that it oxidizes; since the products are all gaseous, they offer no protection against further oxidation. *See* GRAPHITE.

Manufacture. Standard ceramic techniques are used. Hand molding, once widely used, is used only for special shapes and small orders. The extrusion or stiff mud process is used for plastic fireclays; very often the extruded blanks are repressed or hydraulically rammed to form special shapes, for example, T-sections of refractory pipe. Power pressing of simple shapes is the most widely used forming method. Hot pressing and hydrostatic pressing are used for some special refractories. Slip casting is used for special refractory shapes. Fusion casting is commonly used for glass tank blocks; these are mainly either Al_2O_3 or Al_2O_3, with significant amounts of SiO_2, ZrO_2, or both. *See* CERAMICS.

Refractories are generally fired in tunnel kilns, but some periodic kilns are still used, particularly for special shapes. *See* KILN.

Some types of basic refractories, known as chemically bonded, are pressed with a chemical binder, such as magnesium oxychloride, and installed without firing. Some of these, the steel-clad refractories, are encased in a metal sheath at the time of pressing. When the refractory is heated after installation, the iron oxidizes and reacts with the refractory, forming a tight bond between the individual bricks.

In all refractory products and in unfired brick in particular, the maximum possible formed density is desired. To this end, careful crushing and sizing of raw materials are carried out so that, as far as possible, the gaps between large pieces are filled with smaller particles, and the space between these with still smaller, and so on. In the case of clay refractories, it is customary to use prefired (calcined) clay or crushed, fired rejects (both are known as grog) to increase the density and to reduce the firing shrinkage.

Properties. A high melting point is of course necessary in a refractory, but many other properties must be considered in choosing a refractory for a specific application.

A definite melting point is characteristic of pure materials; actual minerals from which refractories are made, for example, clay, are far from pure and hence do not melt at a specific temperature. Rather, they form increasing amounts of liquid as the temperature is increased above a certain minimum temperature at which liquid first appears. This characteristic of gradual softening is indicated by the PCE of the material and the underload test.

High-temperature strength is important for refractories, but most materials become plastic and flow at elevated temperatures. Therefore, the rate of flow (creep rate) at a given temperature under a given load is a more important design criterion.

A knowledge of the thermal expansion of high-temperature materials is important, first, so that allowance can be made in furnace construction [long tunnel kilns must be built with expansion joints of several inches every 10 ft (3 m) or so], and second, because of its relation to thermal-shock resistance.

Thermal conductivity determines the amount of heat that will flow through a furnace wall under

given conditions, and a knowledge of this property is essential to furnace design.

Thermal-shock resistance is the ability of a specimen to withstand, without cracking, a difference in temperature between one part and another. For example, if a red-hot brick is dropped into cold water, it is likely to shatter since the outside cools and contracts while the center is still hot. This cracking is often referred to as thermal spalling, the term spalling meaning any cracking off of large pieces of brick. Other causes of spalling are mechanical (hitting the brick and knocking off a piece) and structural (a reaction in the brick which changes the mineral structure and causes cracking). Thermal-shock resistance is enhanced by high strength, low Young's modulus, low thermal expansion, and sometimes, depending on conditions, high thermal conductivity. Whether or not a given specimen cracks under heat shock depends not only on the material of which it is made, but also on its size and shape and on the test conditions, for example, whether it is dropped into water or into still air at the same temperature.

Various chemical properties are important in refractories. For example, the tendency of the magnesium oxide in basic brick to hydrate, that is, to react with water to form $Mg(OH)_2$, should be as low as possible. Turning to high-temperature chemistry, the rate of corrosion of refractories by molten slags and iron oxide fumes is vital to the length of service. Reference to the appropriate phase equilibrium diagrams may give some indication of which combinations of slag and refractory will react; but usually, actual tests are needed to make any precise predictions. The rate of corrosion depends to a great extent on such physical factors as the porosity of the refractory and whether or not the refractory is wet by the slag.

Carbon deposition is another chemical reaction which affects the life of refractories. The reaction is not with the refractory, but is catalyzed by substances in it. When carbon monoxide, perhaps in the top of a blast furnace, comes in contact with certain iron compounds which can occur in fireclays, its reduction to carbon is catalyzed. This carbon deposits at the site of the catalyst in the brick, and causes the brick to shatter. The effect is most pronounced around 950°F (500°C); much below this temperature, the rate of reaction is too slow, and much above it, the equilibrium oxygen pressure necessary for the reduction is lower than is found in practice. Although the reaction is not completely understood, it has been found that high-temperature firing of the fireclay refractories converts the iron to a form which does not catalyze the carbon deposition.

The bursting of spinel (chrome) refractories in contact with iron oxide is another high-temperature chemical reaction; it is not thoroughly understood, but appears to be related to oxidation and reduction reactions in the refractory.   John F. McMahon

Bibliography. J. I. Duffy (ed.), *Refractory Materials Since 1977*, 1981; *Iron and Steel Society Refractories for Modern Steelmaking Systems*, 1987; C. Krause, *Refractories: The Hidden Industry*, 1987; P. Kumar and R. L. Ammon (eds.), *Refractory Metals: State of the Art*, 1989.

Refrigeration

The cooling of a space or substance below the environmental temperature. The art was known to the ancient Egyptians and people of India, who used evaporation to cool liquids in porous earthen jars exposed to dry night air; and to the early Chinese, Greeks, and Romans, who used natural ice or snow stored in underground pits for cooling wine and other delicacies. In the eighteenth and early nineteenth centuries natural ice cut from lakes and ponds in winter was stored underground for use in summer. The technique of mechanical refrigeration began with the invention of machines for making artificial ice. Great strides have been made in the twentieth century in the application of mechanical refrigeration to fields other than ice making, including the direct cooling and freezing of perishable foods and air conditioning for industry and human comfort.

Mechanical refrigeration is primarily an application of thermodynamics wherein the cooling medium, or refrigerant, goes through a cycle so that it can be recovered for reuse. The commonly used basic cycles, in order of importance, are vapor-compression, absorption, steam-jet or steam-ejector, and air. Each cycle operates between two pressure levels, and all except the air cycle use a two-phase working medium which alternates cyclically between the liquid and vapor phases.

The term "refrigeration" is used to signify cooling below the environmental temperature to lower than about 150 K (−190°F; −123°C). The term "cryogenics" is used to signify cooling to temperatures lower than 150 K. *See* CRYOGENICS.

Vapor-compression cycle. The vapor-compression cycle (**Fig. 1**) consists of an evaporator in which the liquid refrigerant boils at low temperature to produce cooling, a compressor to raise the pressure and temperature of the gaseous refrigerant, a condenser in which the refrigerant discharges its heat to the environment, usually a receiver for storing the liquid condensed in the condenser, and an expansion valve through which the liquid expands from the high-pressure level in the condenser to the low-pressure level in the evaporator. This cycle may also be used for heating if the useful energy is taken off at the condenser level instead of at the evaporator level. *See* HEAT PUMP.

The theoretical vapor-compression cycle can best be analyzed on the pressure-enthalpy or temperature-entropy coordinates for a two-phase fluid (**Fig. 2**). Enthalpy is a parameter that replaces heat content. The specific enthalpy, that is, the enthalpy per unit mass (Btu/lbm; kJ/kg), is equal to the specific internal energy (Btu/lbm; kJ/kg) plus the pressure (lbf · in.$^{-2}$; kPa)–volume (ft^3/lbm; m^3/kg) product divided by the mechanical equivalent of heat (778 ft · lbf/Btu; 1 J).
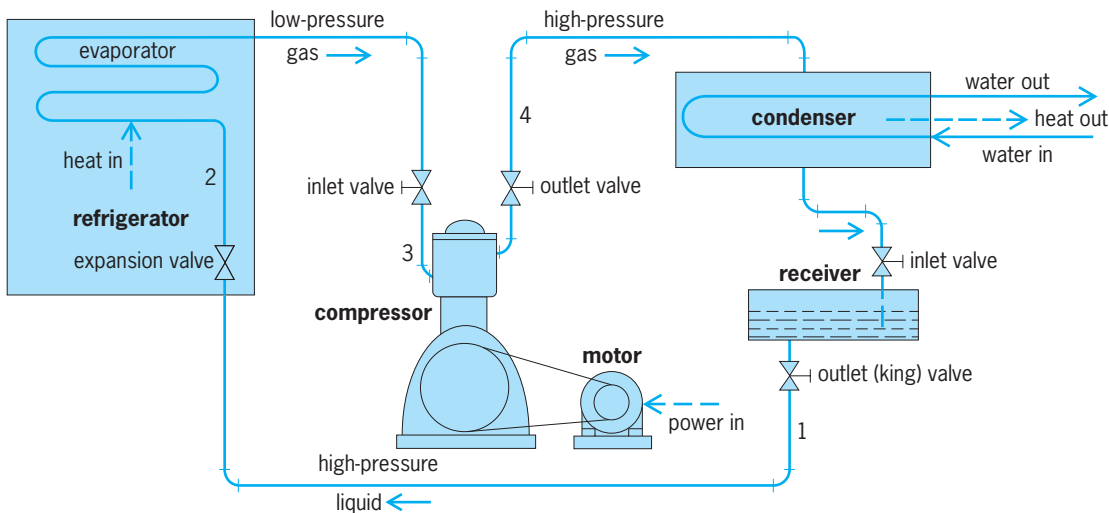
Fig. 1. Diagram of the vapor-compression cycle.

Entropy is a parameter that is obtained by dividing the heat flow by the average absolute temperature during the change. It is expressed in units of Btu per pound mass per degree Rankine (°R = °F + 460) or kilojoule per kilogram degree Kelvin [kJ/(kg · K)]. In Fig. 2, process 1–2 represents adiabatic (constant enthalpy) expansion; 2–3′, constant temperature (and pressure) evaporation; 3′–3, suction superheating at constant pressure; 3–4, ideal frictionless adiabatic (constant entropy) compression; 4–4′, removal of discharge superheat at constant pressure; 4′–1′, condensation at constant pressure (and temperature); and 1′–1, liquid subcooling at constant pressure.

The efficiency of a heat-power cycle is defined as the ratio of useful output to energy input. For a heat engine, efficiency is less than unity. Efficiency is not very meaningful for the refrigeration and heat-pump cycles, for which the term "coefficient of performance" (CP) is used instead. Referring to the theoretical cycle (Fig. 2), the refrigeration CP is the ratio of cooling effect in evaporator 2–3 to compressor energy input 3–4, and the heat-pump CP is the ratio of heating effect in condenser 4–1 to compressor energy input 3–4. The coefficient of performance may be considerably greater than unity, and the theoretical heat-pump CP is 1 plus refrigeration CP. The cooling effect in evaporator 2–3, q_L, is $h_3 - h_2$, and the compressor energy input 3–4, w_C, is $h_4 - h_3$, so that the coefficient of performance of the system considered as a refrigerator, CP_R, is equal to $q_L/W_C = (h_3 - h_2)/(h_4 - h_3)$. The cycle CP is thus determinable from a knowledge of the four enthalpies h_1, h_2, h_3, and h_4.

Compound cycles using two or more compressors in series with a common refrigerant can be used to extend the temperature range with a single compressor stage. However, practical difficulties can occur, and a cascade cycle is used instead, consisting of two or more separate refrigerants each working in its own closed cycle. Such cycles reduce the power consumption and hence increase the thermodynamic efficiency.

Absorption cycle. The absorption cycle accomplishes compression by using a secondary fluid to absorb the refrigerant gas, which leaves the evaporator at low temperature and pressure. Heat is applied, by means such as steam or gas flame, to distill the refrigerant at high temperature and pressure. The most-used refrigerant in the basic cycle (**Fig. 3**) is
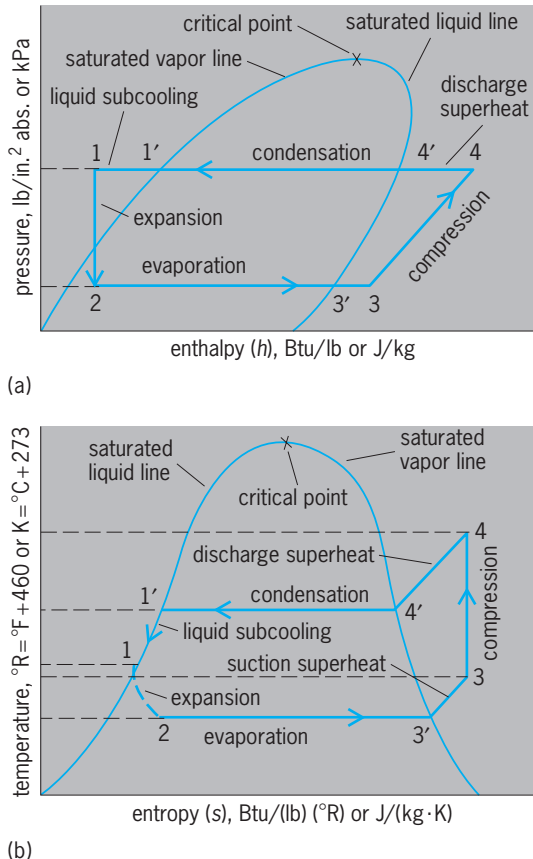

(a)

(b)

Fig. 2. Diagrams showing theoretical analysis of the vapor-compression cycle. (*a*) Pressure-enthalpy diagram. (*b*) Temperature-entropy diagram.
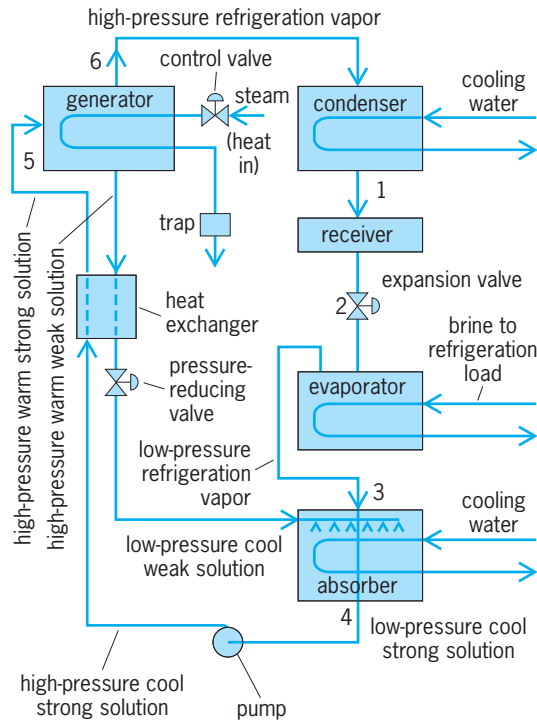
Fig. 3. Basic absorption cycle.

ammonia; the secondary fluid is then water. This system is used for the lower temperatures. Another system is lithium bromide–water, where the water is used as the refrigerant. This is used for higher temperatures. Due to corrosion, special inhibitors must be used in the lithium bromide–water system. The condenser, receiver, expansion valve, and evaporator are essentially the same as in any vapor-compression cycle. The compressor is replaced by an absorber, generator, pump, heat exchanger, and controlling-pressure reducing valve.

The operation of the cycle is based on the principle that the vapor pressure of a refrigerant is lowered by the addition of an absorbent having a lower vapor pressure; and the greater the quantity of absorbent used, the more the depression of the vapor pressure of the refrigerant. *See* DALTON'S LAW.

By maintaining the solution in the absorber at the proper temperature and concentration, the vapor pressure of the solution can be kept lower than that of the refrigerant in the evaporator. Spraying the weak solution in the absorber then causes the refrigerant vapor to flow from the evaporator to the absorber. The strong solution thus formed in the absorber is then pumped through a heat exchanger to the generator, where heat is applied to release the refrigerant vapor. Then follows condensation, expansion, and evaporation, as in the standard vapor-compression cycle. Except for small units, an indirect system is used wherein brine is cooled and circulated to the actual refrigeration load.

For air conditioning, water is the refrigerant and lithium bromide is the absorbent. In terms of the basic absorption cycle (Fig. 3), from 1 to 2 the high-pressure liquid refrigerant is expanded into the evap-

orator where brine is usually cooled, from 2 to 3 the low-pressure refrigerant vapor is drawn into the absorber, from 3 to 4 the low-pressure refrigerant vapor is absorbed in the weak solution, from 4 to 5 the low-pressure strong solution is pumped through the heat exchanger to the high-pressure generator, and from 5 to 6 heat is applied to drive off the refrigerant vapor and force it into the condenser. The hot weak solution drains back to the absorber through the heat exchanger and pressure-reducing valve.

Steam-jet cycle. The steam-jet cycle uses water as the refrigerant. High-velocity steam jets provide a high vacuum in the evaporator, causing the water to boil at low temperature and at the same time compressing the flashed vapor up to the condenser pressure level. Its use is limited to air conditioning and other applications for temperatures above $32°F$ ($0°C$).

The basic steam-jet or ejector cycle (**Fig. 4**) is usually analyzed on temperature-entropy coordinates (**Fig. 5**). High-pressure motive steam (Fig. 4) at 1 is expanded to a low absolute pressure at 2 through a converging-diverging nozzle. Path 1–2 (Fig. 5) is the ideal expansion and 1–2′ the actual expansion allowing for nozzle friction. Water vapor in the evaporator at 3 is entrained by the motive steam at 4, the steam having lost some of its energy from 2′ to 4 because
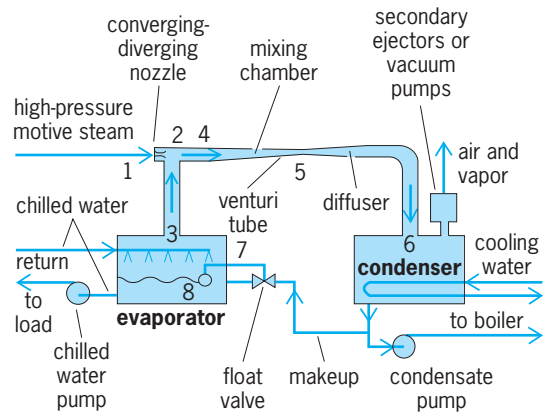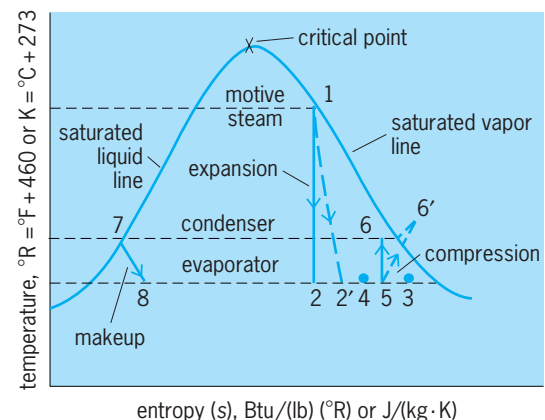

Fig. 4. Steam-jet water-vapor cycle.

Fig. 5. Temperature-entropy coordinates for analyzing basic steam-jet or ejector cycle.

the entrainment efficiency is rather low. The motive steam at 4 plus the entrained moisture at 3 are forced through the venturi tube, in which the velocity of the incoming mixture is reduced and converted into pressure head in the condenser at 6. Path 5–6 is the ideal compression and 5–6′ the actual compression, taking into account compression efficiency. Typical efficiencies are nozzle 88%, entrainment 65%, and compression 80%.

In the evaporator or flash chamber, part of the water is evaporated to cool the rest of the water, which is circulated to the cooling load. Makeup water must be added from 7 to 8. Typical operating conditions are 100°F or 38°C (2 in. Hg or 6.8 kilopascals absolute pressure) in the condenser and 40°F or 4°C (0.25 in. Hg or 850 pascals absolute pressure) chilled water in the evaporator. The condensate from the condenser is pumped back to the boiler; secondary ejectors, or vacuum pumps, are required to remove the air and maintain the high vacuum. Considerably more water is required for condensing than for a vapor-compression system of the same capacity.

The steam jet refrigerator is most widely used for processes where direct vaporization is used for drying of heat-sensitive materials, or in processes using simultaneous desorption and chilling where hard water or seawater can be used. Examples include freeze drying of foods, the concentration of fruit juices, and chilling vegetables.

Air cycle. The air cycle, used primarily in airplane air conditioning, differs from the other cycles in that the working fluid, air, remains as a gas throughout the cycle. Air coolers replace the condenser, and the useful cooling effect is obtained by a refrigerator instead of by an evaporator. A compressor is used, but the expansion valve is replaced by an expansion engine or turbine which recovers the work of expansion. Systems may be open or closed. In the closed system, the refrigerant air is completely contained within the piping and components, and is continuously reused. In the open system, the refrigerator is replaced by the space to be cooled, the refrigerant air being expanded directly into the space rather than through a cooling coil.

One of the typical open air-cycle systems used on airplanes is called the "bootstrap" system (**Fig. 6**). It may be analyzed theoretically on the temperature-entropy coordinates (**Fig. 7**). From 1 to 2 ambient air is compressed ideally in the engine or supercharger of the airplane. Part of this high-pressure air is bled through a primary heat exchanger, where it is cooled from 2 to 3 by ram air, that is, ambient air compressed by the forward motion of the airplane. From 3 to 4 this air is further increased in pressure by the compressor of the refrigeration machine; from 4 to 5 the air is cooled by ram air in the cooler or secondary heat exchanger; and from 5 to 6 the air is further cooled in the expansion turbine, ideally without moisture. However, there is entrained moisture, about 70% of which is removed by the water separator, resulting in an approximate path 5–6′ as the air is warmed by the heat given up by the condensation and removal of this moisture. The balance of the
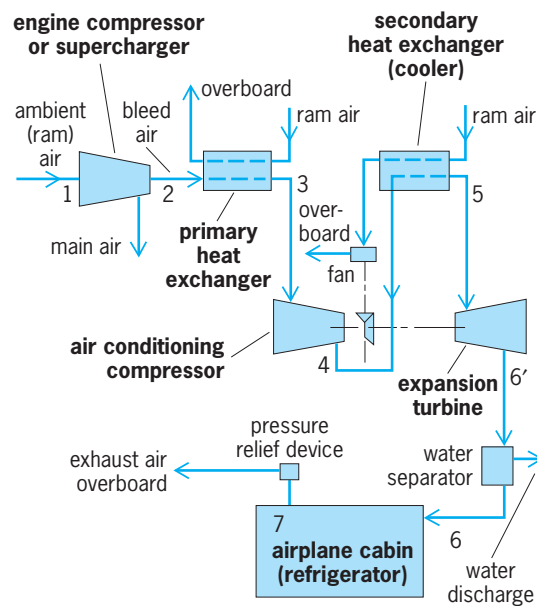

Fig. 6. Open air-cycle bootstrap system for airplanes.

moisture evaporates in the cabin and contributes to the cooling effect 6′–7, where 6′ is the dry-air rated temperature of the air as it enters the cabin. Dry-air rated temperature is the temperature which would be attained by the expansion of the refrigerant air in the absence of any moisture condensation. Refrigerant air at 7, having absorbed the heat load in the cabin, is released overboard through a pressure-relief device. The equipment is proportioned so that the work of expansion is recovered and is sufficient to drive the compressor and the cooler fan to approximate constant entropy expansion and compression. However, compressor and turbine efficiencies and heat-exchanger pressure drops, neglected in this analysis, reduce the ideal performance.

Thermoelectric refrigeration. The effect of heat application to one junction of a couple composed of two dissimilar conductors is to produce a thermoelectric voltage on the circuit (Seebeck effect).
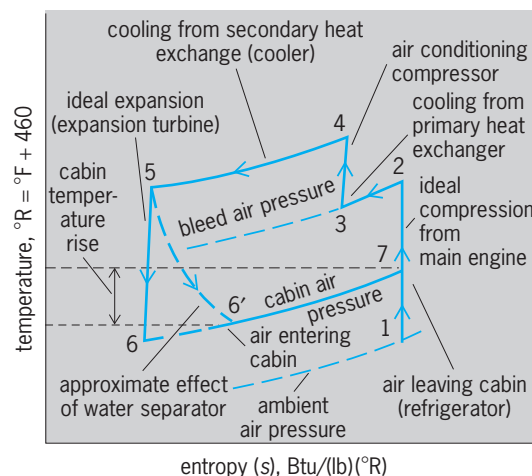

Fig. 7. Temperature-entropy coordinates for analyzing theoretically the air-cycle bootstrap system.

Fluids in REFPROP* Database

Name or R number	Chemical formula	Melting point, °C	Boiling point, °C	Critical temp., °C	Critical press., bar	Molecular weight
Ammonia	NH_3	−77.9	−33.3	133.0	114.2	17.03
Butane	C_4H_{10}	−159.5	−0.4	152.0	38.0	58.13
Carbon dioxide	CO_2	−56.7	−78.4	31.1	73.7	44.01
Ethane, R 170	C_2H_6	−174.2	−88.8	32.2	48.9	30.07
Isobutane, R 600a	C_4H_{10}	−136.2	−11.7	135.0	36.5	58.13
Propane, R 290	C_3H_8	−187.7	−42.1	96.8	42.5	44.10
Propylene, R 1270	C_3H_6	−185	−47.7	91.8	50.5	42.08
R 11	CCl_3F	−111	23.8	198.0	44.1	137.37
R 12	CCl_2F_2	−157.8	−29.8	112.0	41.1	120.91
R 13	$CClF_3$	−181.2	−81.4	28.8	38.7	104.46
R 14	CF_4	−183.7	−127.9	−45.7	37.5	88.00
R 22	$CHClF_2$	−160	−40.8	96.2	49.9	86.47
R 23	CHF_3	−155.2	−82.1	26.3	48.7	70.01
R 32	CH_2F_2	−136	−51.7	−78.2	58.0	52.02
R 41	CH_3F	−114.8	−78.4	44.3	58.8	34.03
R 113	$C_2Cl_3F_3$	−35.0	47.6	214.1	34.4	187.38
R 114	$C_2Cl_2F_4$	−93.9	3.8	145.7	32.5	170.92
R 115	C_2ClF_5	−102	−39.1	79.9	31.5	154.47
R 116	C_2F_6	−100.6	−78.2	19.9	30.4	138.01
R 123	$C_2HCl_3F_3$	−107	27.9	183.7	36.7	152.93
R 124	C_2HClF_4	−100	−12.0	122.5	36.4	136.48
R 125	C_2HF_5	−103	−48.1	66.3	36.3	120.02
R 134a	$C_2H_2F_4$	−101	−26.1	101.1	40.6	102.03
R 141b	$C_2H_3Cl_2F$	−103.5	32.2	204.4	42.5	116.95
R 142b	$C_2H_3ClF_2$	−130.8	−9.8	137.2	41.2	100.50
R 143a	$C_2H_3F_3$	−111	−47.4	73.6	38.3	84.04
R 152a	$C_2H_4F_2$	−117	−24.0	113.3	45.2	66.05
R 227ea	C_3HF_7		−18.3	103.5	29.5	170.03
R 236ea	$C_3H_2F_6$		6.6	141.2	35.3	152.04
R 236fa	$C_3H_2F_6$		−1.1	130.7	31.8	152.04
R 245ca	$C_3H_3F_5$		25.5	178.5	38.6	134.05
R 245fa	$C_3H_3F_5$		15.3	157.6	36.4	134.05
R C318	C_4F_8		−5.8	115.4	27.8	200.04

*Fluids listed in M. O. McLinden et al., NIST Standard Reference Database 23, 1998. Many of these fluids are being evaluated for use as binary mixtures—see M. O. McLinden et al., *Int. J. Refrig.*, 21(4):322–338, 1998. For fluoroethers see D. B. Bivens and B. H. Minor, *Int. J. Refrig.*, 21(7):567–576, 1998. Tables and charts for many of these fluids are contained in the *Handbook of Fundamentals*, ASHRAE, Atlanta, 1997, and in *Perry's Chemical Engineers Handbook*, McGraw-Hill, New York, 1997.

Inversely, this is matched by a cooling effect at one junction of an equivalent couple when a direct-current electrical flow is imposed on the circuit. This is known as the Peltier effect, discovered in 1834. The amount of the refrigerating effect is multiplied by the number of couples in parallel; the heat absorption at the cold junction is rejected at the second junction of each couple (that is, heat-pump effect).

With ordinary materials generally available prior to World War II, the Peltier effect, though recognized, was not considered particularly practical. Since the war and with the advent of the transistor and other semiconductor devices leading to more practical and efficient equipment, there has been considerable worldwide interest in the possibilities of this system. Its application is to special-purpose processes, since its low efficiency has to be compensated by its special features, such as elimination of moving parts and noiseless operation. *See* PELTIER EFFECT; SEEBECK EFFECT; THERMOELECTRICITY.

Refrigerants and equipment. The working fluid in a two-phase refrigeration cycle is called a refrigerant. A useful way to classify refrigerants is to divide them into primary and secondary. Primary refrigerants are those fluids (pure substances, azeotropic mixtures which behave physically as a single pure compound, and zeotropes which have temperature glides in the condenser and evaporator) used to directly achieve the cooling effect in cycles where they alternately absorb and reject heat. Secondary refrigerants are heat transfer or heat carrier fluids.

A number of commonly used refrigerantsm are listed in the **table**. Ammonia and Freon-22 are most important for industrial refrigeration, Freon-11 and Freon-12 being used most often for commercial and air conditioning work in which nontoxic refrigerants are necessary. A secondary cooling liquid that does not change from the liquid phase is called a brine. Solutions of sodium chloride or calcium chloride in water are frequently used as circulating brines in refrigeration systems.

Compressors. Refrigeration compressors may be positive-displacement of the reciprocating, rotary, or gear type for high- and medium-pressure differentials, or of the centrifugal type for low-pressure differentials. Early ammonia compressors were horizontal, double-acting, slow-speed units built like steam engines. Modern reciprocating compressors are vertical, single-acting, multicylinder, high-speed units built like automobile engines. Ammonia and other large compressors require water jacketing, whereas most Freon compressor cylinders are air-cooled.

Condensers. Refrigeration condensers may be air-cooled for small and medium capacities; or

water-cooled, as in the case of the shell-and-tube, shell-and-coil, or double-pipe types. Because of the large quantities of condensing water required, cooling towers or spray ponds are commonly used to re-cool the water for reuse. An evaporative condenser is a device combining a condensing coil and a forced-draft cooling tower in a single unit.

Evaporators. Refrigerant evaporators are the cooling units placed in the room or fluid to be cooled. Plain pipe coils or finned coils, with or without forced circulation of the fluid being cooled, are commonly used. Shell-and-tube coolers, or tanks with wetted or submerged cooling coils, are frequently used where water or brine is circulated as a secondary cooling medium.

Expansion valve. The main flow control in a vapor-compression system is the expansion valve. It permits the liquid refrigerant to expand from the high pressure in the condenser to the lower pressure in the evaporator. The expansion causes part of the liquid to evaporate and thereby to cool the remainder to the evaporator temperature. A float valve with a throttling orifice is often used instead of an expansion valve to provide flooded control and maintain a fixed liquid level in the evaporator. In domestic refrigerators, a capillary tube restricts flow from condenser to evaporator. For completely automatic operation, additional controls are required to maintain the desired temperature in the evaporator, to regulate the compressor operation and the flow of the condensing medium, and to provide safety protection. *See* AIR CONDITIONING; AIR COOLING; AUTOMOTIVE CLIMATE CONTROL; COLD STORAGE; COOLING TOWER; DRY ICE; ICE MANUFACTURE; MARINE REFRIGERATION; REFRIGERATOR.   Peter E. Liley; Carl F. Kayan

Bibliography. Air Conditioning and Refrigeration Institute Staff, *Refrigeration and Air Conditioning*, 2d ed., 1987; R. J. Dossat, *Principles of Refrigeration*, 1996; B. C. Langley, *Fundamentals of Refrigeration*, 1995; B. Parsons (ed.), *1998 ASHRAE Handbook: Refrigeration* (SI or Inch-Pound editions), 1998.

Refrigeration cycle

A sequence of thermodynamic processes whereby heat is withdrawn from a cold body and expelled to a hot body. Theoretical thermodynamic cycles consist of nondissipative and frictionless processes. For this reason, a thermodynamic cycle can be operated in the forward direction to produce mechanical power from heat energy, or it can be operated in the reverse direction to produce heat energy from mechanical power. The reversed cycle is used primarily for the cooling effect that it produces during a portion of the cycle and so is called a refrigeration cycle. It may also be used for the heating effect, as in the comfort warming of space during the cold season of the year. *See* HEAT PUMP; THERMODYNAMIC PROCESSES.

In the refrigeration cycle a substance, called the refrigerant, is compressed, cooled, and then expanded. In expanding, the refrigerant absorbs heat from its surroundings to provide refrigeration. After the refrigerant absorbs heat from such a source, the cycle is repeated. Compression raises the temperature of the refrigerant above that of its natural surroundings so that it can give up its heat in a heat exchanger to a heat sink such as air or water. Expansion lowers the refrigerant temperature below the temperature that is to be produced inside the cold compartment or refrigerator. The sequence of processes performed by the refrigerant constitutes the refrigeration cycle. When the refrigerant is compressed mechanically, the refrigerative action is called mechanical refrigeration.

There are many methods by which cooling can be produced. The methods include the noncyclic melting of ice, or the evaporation of volatile liquids, as in local anesthetics; the Joule-Thomson effect, which is used to liquefy gases; the reverse Peltier effect, which produces heat flow from the cold to the hot junction of a bimetallic thermocouple when an external emf is imposed; and the paramagnetic effect, which is used to reach extremely low temperatures. However, large-scale refrigeration or cooling, in general, calls for mechanical refrigeration acting in a closed system. *See* PARAMAGNETISM; REFRIGERATION.

Reverse Carnot cycle. The purpose of a refrigerator is to extract as much heat from the cold body as possible with the expenditure of as little work as possible. The yardstick in measuring the performance of a refrigeration cycle is the coefficient of performance, defined as the ratio of the heat removed to the work expended. The coefficient of performance of the reverse Carnot cycle is the maximum obtainable for stated temperatures of source and sink. **Figure 1** depicts the reverse Carnot cycle on the *T-s* plane. *See* CARNOT CYCLE.

The appearance of the cycle in Fig. 1 is the same as that of the power cycle, but the order of the cyclic processes is reversed. Starting from state 1 of the figure, with the fluid at the temperature T_H of the hot body, the order of cyclic events is as follows:

1. Isentropic expansion, 1–2, of the refrigerant fluid to the temperature T_c of the cold body.
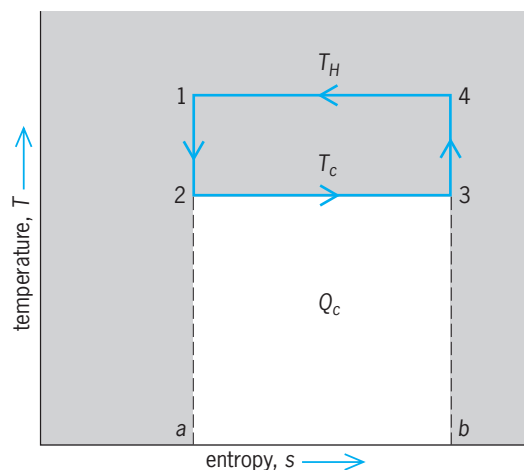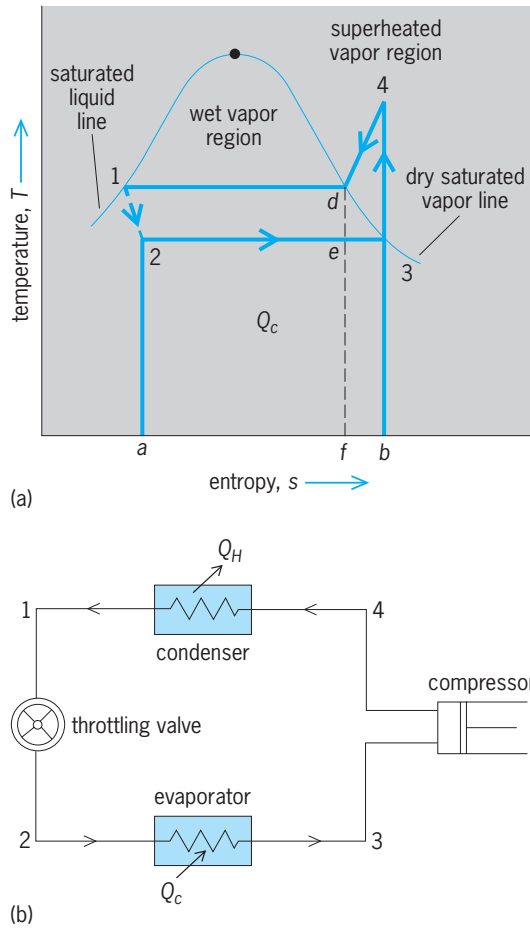

Fig. 1. **Reverse Carnot cycle.**

(a)

(b)

Fig. 2. Vapor-compression refrigeration cycle.
(*a*) Temperature-entropy diagram. (*b*) Diagram of the
components of a vapor-compression system.

2. Isothermal expansion, 2–3, at the temperature
T_c of the cold body during which the cold body gives
up heat to the refrigerant fluid in the amont Q_c, rep-
resented by the area 2–3–*b*–*a*.

3. Isentropic compression, 3–4, of the fluid to the
temperature T_H of the hot body.

4. Isothermal compression, 4–1, at the temper-
ature T_H of the hot body. During this process, the
hot body receives heat from the refrigerant fluid in
the amount Q_H represented by the area 1–4–*b*–*a*.
The difference Q_H–Q_c represented by area 1–2–3–4
is the net work which must be supplied to the cycle
by the external system.

Figure 1 indicates that Q_c and the net work rect-
angles each have areas in proportion to their ver-
tical heights. Thus the coefficient of performance,
defined as the ratio of Q_c to net work, is $T_c/(T_H - T_c)$.

The reverse Carnot cycle does not lend itself to
practical adaptation because it requires both an ex-
panding engine and a compressor. Nevertheless, its
performance is a limiting ideal to which actual refrig-
eration equipment can be compared.

Modifications to reverse Carnot cycle. One change
from the Carnot cycle which is always made in real
vapor-compression plants is the substitution of an
expansion valve for the expansion engine. Even if
isentropic expansion were possible, the work deliv-

ered by the expansion engine would be very small
and the irreversibilities present in any real operations
would further reduce the work delivered by the ex-
panding engine. The substitution of an expansion
valve (throttling orifice), with constant enthalpy ex-
pansion, changes the theoretical performance but
little, and greatly simplifies the apparatus. A typical
vapor-compression refrigeration cycle is shown in
Fig. 2.

Another practical change from the ideal Carnot
cycle substitutes dry compression 3–4 for wet com-
pression *e-d* in Fig. 2, placing state 4 in the super-
heat region above ambient temperature; the process
is called dry compression in contrast to the wet com-
pression of the Carnot cycle. Dry compression intro-
duces a second irreversibility by exceeding the am-
bient temperature, thus reducing the coefficient of
performance. Dry compression is usually preferred,
however, because it simplifies the operation and con-
trol of a real machine. Vapor gives no readily observ-
able signal as it approaches and passes point *e* in
the course of its evaporation, but it would undergo a
temperature rise if it accepted heat beyond point 3.
This cycle, using dry compression, is the one which
has won overwhelming acceptance for refrigeration
work.

Reverse Brayton cycle. The reverse Brayton cycle
constitutes another possible refrigeration cycle; it
was one of the first cycles used for mechanical
refrigeration. Before Freon and other condensable
fluids were developed for the vapor-compression
cycle, refrigerators operated on the Brayton cycle,
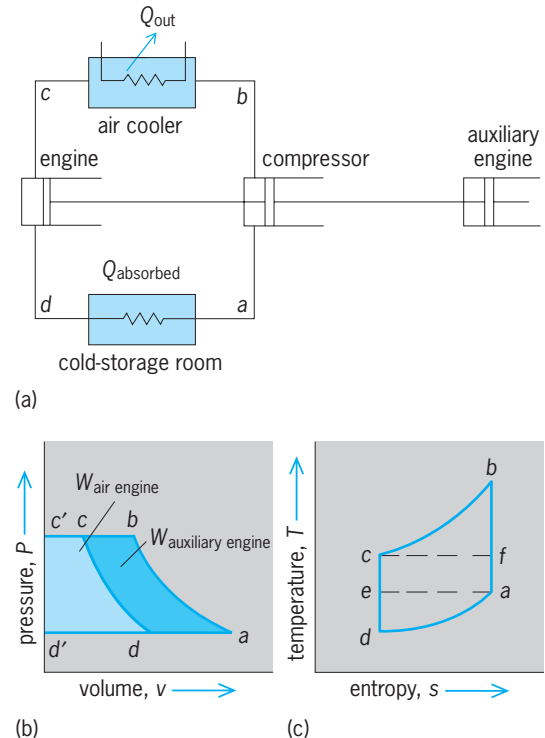using air as their working substance. **Figure 3**

(a)

(b)

(c)

Fig. 3. Reverse Brayton cycle. (*a*) Schematic diagram
of components. (*b*) Pressure-volume diagram.
(*c*) Temperature-entropy diagram.

Ideal refrigeration cycle performance*

Refrigerant	Saturation pressure, lb/in.^{2} absolute		Refrigerating effect, Btu/lb	Refrigerant required per ton refrigeration		Compressor horsepower per ton refrigeration	Coefficient of performance
	Evaporator	Condenser		lb/min	cfm		
Carnot-any fluid							6.6
Freon, F12	47	151	49	4.1	3.6	0.89	5.31
Ammonia	66	247	445	0.44	1.9	0.85	5.55
Sulfur dioxide	24	99	135	1.5	4.8	0.83	5.67
Free air	15	75	32	6.2	7	2.7	1.75
Dense air	50	250	32	6.2	23	2.7	1.75
Steam	0.1	1.275	1000	0.2	590	0.91	5.18

*Performance is based on evaporator temperature = 35°F (2°C), condenser temperature = 110°F (44°C). In vapor-compression cycles, the pressures and temperatures are for saturation conditions; vapor enters the compressor dry and saturated; no subcooling in the condenser.
1 lb/in.^{2} absolute = 6.895 kPa.
1 Btu/lb = 2326 J/kg.
1 (lb/min)/ton refrigeration = 2.15×10^{-6} (kg/s)/W.
1 cfm/ton refrigeration = 1.34×10^{-7} (m^3/s)/W.
1 compressor horsepower per ton of refrigeration = 0.212 watt of compressor power per watt of refrigeration.

presents the schematic arrangement of this cycle. Air undergoes isentropic compression, followed by reversible constant-pressure cooling. The high-pressure air next expands reversibly in the engine and exhausts at low temperature. The cooled air passes through the cold storage chamber, picks up heat at constant pressure, and finally returns to the suction side of the compressor. *See* BRAYTON CYCLE.

The temperature-entropy diagram (Fig. 3c) points up the disadvantage of the dense-air cycle. If the temperature at *c* represents the ambient, then the only way that air can reject a significant quantity of heat along the line *b-c* is for *b* to be considerably higher than *c*. Correspondingly, if the cold body service temperature is *a*, the air must be at a much lower temperature in order to accept heat along path *d-a*. If a reverse Carnot cycle were used with a working substance undergoing changes in state, the fluid would traverse path *a-f-c-e* instead of path *a-b-c-d*. The reverse Carnot cycle would accept more heat along path *e-a* than the reverse Brayton cycle removes from the cold body along path *d-a*. Also, since the work area required by the reverse Carnot cycle is much smaller than the corresponding area for the reverse Brayton cycle, the vapor-compression cycle is preferred in refrigeration practice. *See* THERMODYNAMIC CYCLE.

Comparative performance of refrigerants. The **table** gives the significant theoretic performance data for a selected group of refrigerants when used in the ideal cycles as outlined above. These data include not only the requisite pressures in the evaporator and condenser for saturation temperatures of 35°F (2°C) and 110°F (44°C), respectively, but also the refrigerating effect, the weight and volume of refrigerant to be circulated, the horsepower required, and the coefficient of performance. The vapor compression cycles presuppose that the refrigerant leaves the evaporator dry and saturated (point 3 in Fig. 2) and leaves the condenser without subcooling (point 1 in Fig. 2). The heat absorbed, Q_c, in the evaporator is called the refrigerating effect and is measured in Btu/lb or joules/kg of refrigerant. The removal of heat in the

evaporator at the rate of 200 Btu/min (3.517 kW) is defined as a ton refrigeration capacity. The weight and volume of refrigerant, measured at the compressor suction, and the theoretic horsepower required to drive the compressor per ton refrigeration capacity are also given in the table. These data show a wide diversity of numerical value. Each refrigerant has its advantages and disadvantages. The selection of the most acceptable refrigerant for a specific application is consequently a practical compromise among such divergent data.

Recent developements. The pulse tube refrigerator is a simple device containing a tube with one closed end and a movable piston at the other end. The closed end contains a heat exchanger, while the open end contains a heat exchanger through which a piston can move. With an orifice inserted in the tube, the device is called an orifice pulse tube refrigerator.   Theodore Baumeister; Peter E. Liley

Bibliography. American Society of Heating, Refrigerating and Air-Conditioning Engineers, *Heating, Ventilating, and Air Conditioning Guide*, revised periodically; E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996.

Refrigerator

An insulated, cooled compartment. If it is large enough for the entry of a person, it is termed a walk-in box; otherwise it is called a reach-in refrigerator. Cooling may be by mechanical or gas refrigeration, by water or dry ice, or by brine circulation. Temperatures maintained depend upon the requirements of the product stored, generally varying from 55°F (13°C) down to 0°F (−18°C), and sometimes lower.

A household or domestic refrigerator is a factory-built, self-contained cabinet. The range of storage capacities is wide and varies among manufacturers. Modern designs have a main compartment for holding food above freezing, a second compartment for storage below freezing, and trays for the freezing of

ice cubes. The cabinets are usually all metal with $2\text{-}3\frac{1}{2}$ in. (5–9 cm) of insulation. The refrigeration unit is usually driven by electric motor, but gas refrigerators motivated by the thermal energy of fuel gas have been used extensively where cheap natural gas is available. Low-temperature household refrigerators, or home freezers, for the storage of frozen foods are manufactured in both the chest and the upright, or vertical, types.

A commercial refrigerator is any factory-built refrigerated fixture, cabinet, or room that can be assembled and disassembled readily, in contrast to a built-in refrigerator. Commercial or built-in refrigerators are used in restaurants, markets, hospitals, hotels, and schools for the storage of food and other perishables. A meat cooler is a refrigerator held at about $33°F$ ($1°C$) for the storage of fresh meats. Refrigerators for lower temperatures down to $0°F$ ($-18°C$) and below are called freezer boxes. Insulation thicknesses vary from 2 to 8 in. (5 to 20 cm), depending upon the service. In markets and stores, commercial display refrigerators may be of the self-service type from which customers help themselves. Both vertical types with glass doors, which the customer opens, and chest types with open tops are used. Electric refrigeration units may be built into each fixture, or they may be remotely located. *See* COLD STORAGE; REFRIGERATION.   Carl F. Kayan

Regeneration (engineering)

Any of the various processes for restoring a system to its original state by restoring some property of a system to its original value, or for using the properties of a system at one point in a cycle to modify the properties of the same system at another point in that cycle. The term has a wide variety of uses in different fields of engineering.

In mechanical engineering, regeneration is the process of transferring heat from the working fluid at one part of a cycle to the working fluid at a different part of the same cycle to increase cycle efficiency. An example is the use of extraction steam to preheat incoming boiler water. *See* BOILER.

In analog electronics, regeneration is the process of sampling the output variable, usually a voltage, and adding it to the input variable, again usually a voltage, so that an amplifier senses an increased signal level. Regeneration, in this sense, is called positive feedback. Low-level, controlled regeneration can result in increased amplification at the expense of narrowed bandwidth and decreased signal-to-noise ratio. Higher-level regeneration can result in oscillation, producing an output with no input. *See* FEEDBACK CIRCUIT; OSCILLATOR.

In digital electronics, regeneration is the process of restoring a set of data in a data storage device to a working level or restoring a degraded pulse-code-modulation (PCM) waveform to its original shape. In the data storage device, decay of the value of the variable representing a binary number, 0 or 1, can result in loss of data if the values are not refreshed or regenerated. In a pulse-code-modulation transmission system, a regenerative repeater can restore a degraded pulse waveform to its original shape, thus eliminating the effects of acquired noise or introduced distortion. *See* COMPUTER STORAGE TECHNOLOGY; PULSE MODULATION.

In higher-power electric motor applications, regeneration refers to the recapture of kinetic energy of the system by converting motor action to generator action when the motor is required to reduce speed. An example of this process is when the drive motors of an electrically propelled train are used as generators for dynamic braking, returning energy to the system during braking operations. *See* DYNAMIC BRAKING.

In nuclear engineering, regeneration refers to the process of increasing the fissile content of used nuclear fuel, a process that is achievable but seldom used. *See* NUCLEAR FUELS REPROCESSING.

William L. Beasley

Regenerative biology

The field within developmental biology that focuses on regeneration, the process by which an animal restores a lost part of its body. Broadly defined, the term regeneration can include many kinds of restorative activities shown by a wide variety of organisms. However, regeneration in this context connotes reproduction of a more perfect or complete replacement of a missing tissue or organ than usually results from processes of tissue repair or wound healing. Cells of the epidermis, blood, and digestive and reproductive tracts are regenerated continuously or periodically in all vertebrates. Within the field of developmental biology, however, most research in regeneration involves systems in which removing a complex structure or major part of an organism initiates a chain of events that produces a structure that duplicates the missing part both functionally and anatomically. *See* DEVELOPMENTAL BIOLOGY.

The best-known and most widely studied examples of regeneration are those involving epimorphosis, in which the lost structure is reproduced directly by a combination of cell proliferation and redifferentiation of new tissue. Examples can be found throughout the animal kingdom, from the extensive regenerative capacities of *Hydra* and *Planaria* to the regrowth of antlers in deer. Research on regenerating systems such as in annelid worms, hydra, and amphibian tails and limbs has yielded information regarding basic mechanisms of animal development. Regeneration can involve growth of an existing population of unspecialized cells or cells arising by dedifferentiation of tissues at the site of injury. Both processes are currently very active areas of investigation within the field of regenerative biology, holding promise for new approaches to various medical problems. *See* STEM CELLS; CELL DIFFERENTIATION.

Tissue and organ regeneration in mammals. Mammals, birds, and reptiles have a much more poorly

developed ability than amphibians and fish to regenerate complete organs, but nevertheless can reform missing tissue and restore function after partial removal of certain organs.

Liver. If part of the liver is cut away, the remaining portion increases in size to compensate for the missing tissue and to restore the normal functional capacity of the organ. The process of liver regeneration involves the triggering of active growth in the remaining liver cells, in cells of bile ductules, and in unspecialized stem cells, all of which are usually quiescent in the normal liver. Proliferation of these cells and their subsequent differentiation are key events in the process by which the missing liver mass is replaced and adequate hepatic function restored. *See* LIVER.

Musculoskeletal system. In the musculoskeletal system, different populations of quiescent stem cells allow efficient replacement of damaged or partially removed bones and muscles. Such undifferentiated progenitor cells are present within the inner and outer connective tissue coverings of bones, and are the major source of new proliferating cells from which new bone mass is formed during bone repair or restoration after surgery. Growth and patterning in these cells during bone regeneration are influenced by an important class of growth factors called bone morphogenetic proteins and by patterns of stress and mechanical stimulation placed on the existing bone. In skeletal muscle, undifferentiated cells called satellite cells are normally present along with the highly specialized muscle cells or fibers. When muscle fibers are killed by injury, the accompanying satellite cells are activated to proliferate and give rise to new muscle cells. The lack of satellite cells in heart muscle is one reason for the relatively poor potential for regeneration or repair exhibited by that tissue compared to other types of muscle. *See* BONE; MUSCLE.

While connective tissue cells of the bone marrow have long been recognized as a potential source of stem cells during bone repair, it has recently been shown that bone marrow–derived cells can also participate in regeneration of injured skeletal muscle and a variety of other tissues. Migrating from the marrow cavity to damaged muscle apparently via the circulation, these stem cells proliferate and contribute to the newly formed muscle fibers, supplementing the contribution of local satellite cells. Because these bone marrow cells can be isolated easily and grown in the laboratory, regenerative biologists are investigating whether such stem cells can be developed as grafts for restoring healthy cells in degenerative muscle or brain disorders.

Central nervous system. Brain and spinal cord research is of prime importance within regenerative medicine because of the high frequency and devastating effects of central nervous system (CNS) injuries. Progress remains slow but has been sufficient to dispel the long-held belief that mammalian CNS tissues are inherently unable to regenerate after injury. Limited reorganization and compensation do occur within the injured adult spinal cord, and such tissues now appear to be amenable to therapeutic manipulation. Moreover, new populations of stem cells have been discovered in the CNS. In peripheral nerves regenerative growth of cut or injured nerve fibers (axons) usually occurs readily with only partial interference from scar tissue. In the injured CNS, however, axonal regeneration appears to be blocked physically by a scar produced by several types of local cells and inhibited chemically by exposure to factors produced by these cells. Current attempts to circumvent these blocks to neural regeneration in the CNS include (1) local treatment with factors designed to reduce scar formation, inactivate inhibitors of axon growth, and stimulate nerve fiber regeneration and (2) grafting of stem cells, peripheral nerve tissue, or cells capable of ensheathing and facilitating regrowth of axons. *See* NERVE.

Appendages. Mammals have little capacity to regenerate missing appendages. One exception is the annual regrowth of shed antlers in deer. Regeneration involves a rapidly growing bud of skin-covered tissue that forms cartilage, then bone, until fully grown, at which point the skin is lost. In humans, especially children, clinical experience has shown that fingertips can regenerate fully if the wound is not closed with sutures but is allowed to heal normally with skin. Such observations of regenerative ability hold the promise that regrowth of missing appendages may be inducible in mammals once the various required factors are known.

Role of the immune system. Another model of mammalian appendage regeneration, the reconstitution of tissue to close large holes produced in the cartilaginous portion of the ear, which was thought to occur only in rabbits, was recently found in a strain of mice with specific defects in the immune system. This regenerative capacity in mice with immunological deficits is consistent with the hypothesis that the decline in regenerative ability during the evolution of mammals may correlate with the rise of more complex and efficient adaptive immune systems which allow mammals to mount rapid inflammatory responses in wounds to avoid infection. Such inflammation usually culminates in scar formation, which has been shown in many systems to be incompatible with a normal regenerative response. Animals that lack highly developed adaptive immunity, including invertebrates and primitive vertebrates, generally have greater regenerative potential than mammals. Regeneration of ear tissue in genetically well-characterized immune-deficient mice is important since it may allow molecular dissection of the genes required for the regenerative response in mammals. *See* CELLULAR IMMUNOLOGY; IMMUNITY.

Regeneration of damaged skin, without scarring and with good restoration of hair follicles, occurs if surgical injury is performed at an embryonic stage of skin development, and this has become an experimental model for regeneration in mammals. If skin injury is delayed until later stages of development, regeneration becomes progressively more imperfect as evidenced by scar formation. Research in scar-free wound healing using mice with various genes

"knocked out" again points to roles for cells of the immune system in the response to injury that leads to scarring. The immune system forms during the same embryonic period in which the potential for scarring appears. Greater understanding of the molecular events that lead to scar formation is yielding new treatments to mitigate postsurgical scarring and may yield additional clues regarding the loss of regenerative ability in skin and other organs.

Limb regeneration in amphibians. Of all vertebrates, amphibians have the most highly developed capacity for regeneration. Certain species have the ability to regenerate not only limbs and tails but also parts of the eye, lower jaw, intestine, and heart.

Impact of metamorphosis. Complete regeneration of amputated limbs can occur throughout the lifetime of most urodeles (salamanders and newts) but usually occurs much more rapidly in larval animals. In anurans (frogs and toads) the ability to regenerate limbs is lost during metamorphosis to the adult form. This transition from a regenerative to nonregenerative capacity in frogs has been investigated for insights into the more general inability of other vertebrates to regenerate missing appendages. As with the ontogeny of skin scarring in the mammalian embryo, the evidence suggests a primary role for the developing immune system in the loss of regenerative ability. Comparative immunologists have shown that many aspects of immunity in larval anurans are primitive and that during metamorphosis the adaptive immune system undergoes changes that result in adult frogs having immunity similar to that of mammals. Urodeles, which retain well-developed regenerative potential throughout life, display the more primitive type of adaptive immunity found in frog larvae. Evolution of the adaptive immune system, many aspects of which are involved in inflammation and defense against pathogens during wound healing, may have produced mechanisms that lead to scarring and failure of epimorphic regeneration, a possibility currently under active investigation. *See* METAMORPHOSIS.

Mechanism. All amphibian limbs that are capable of regeneration show essentially the same sequence of events following amputation (**Fig. 1**). The end of the limb stump is rapidly covered with epidermis that migrates out from the cut edge of the skin. Injured tissues immediately beneath this newly formed wound epithelium begin to dissociate, while much of the extracellular material is broken down and most specialized characteristics of the cells in these tissues are lost. This process, termed dedifferentiation, results in the appearance of cells similar to those in the embryonic limb bud. As they dedifferentiate, the cells begin to proliferate and soon form a mass of cells, the blastema, below the epidermis at the end of the limb stump. Continued proliferation of the blastema cells causes a budlike outgrowth on the stump. The regeneration bud gradually elongates into a conical structure that flattens and becomes paddle-shaped. During elongation of the blastema, proximal cells adjacent to tissues in the stump begin to differentiate. Capillaries grow into the blastema from the stump,

and new cartilage begins to form in the area contiguous with cut skeletal elements of the stump. By the paddle stage, cartilage formation is under way in central areas throughout most of the outgrowth. Muscle and connective tissue also redifferentiate from blastemal cells in a proximal-to-distal direction. Reformation of the amputated parts of the limb is completed as digits develop along the distal end of the flattened structure and articulate with skeletal elements differentiating in that region. The typical regenerated limb is morphologically and functionally identical to the amputated appendage. *See* EMBRYONIC DIFFERENTIATION.

Role of electric currents. As in simple wound healing and tissue repair, the initial local stimuli for limb regeneration are generated as a result of injury and include enzymes and growth factors released from damaged tissue and from immigrating white blood cells. In addition to these biochemical factors, electrical properties may be important during the onset of regeneration. Currents driven by the uptake of sodium ions across the skin can be measured leaving the ends of amputated limbs and are much greater in regenerating species, such as newts and salamanders, than in adult frogs. Moreover, reducing the flow of current by various methods inhibits limb regeneration, while increasing the current associated with frog limb stumps by means of implanted batteries produces a regenerative response in this animal. The cellular basis for the electrical effects on growth is
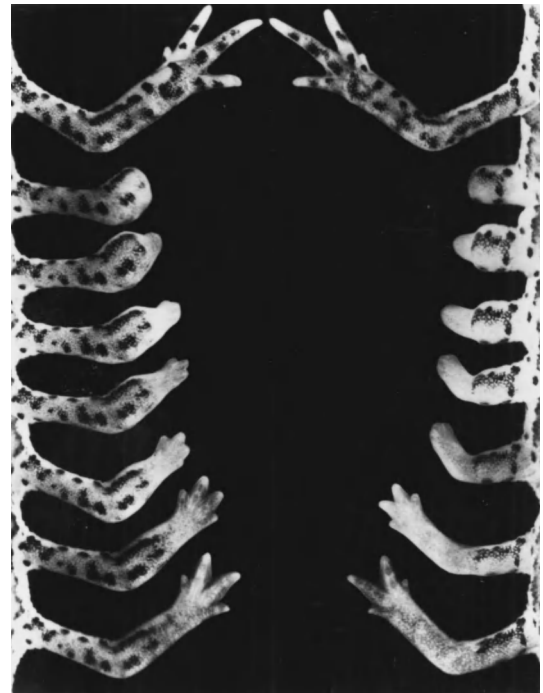

Fig. 1. Amputation of the forelimb of a newt, either through the level of the radius and ulna (left) or through the humerus (right), begins a process of regeneration that includes several morphological stages and results in an anatomically complete and fully functional new limb. The new four-digit limb formed by regeneration will continue to grow until the size of the original limb is attained. (*From R. J. Goss, Principles of Regeneration, Academic Press, 1969*)

not known, but experimental evidence suggests that endogenous currents are important for initiation of the limb regeneration process. *See* BIOPOTENTIALS AND IONIC CURRENTS.

Tissue interactions. The population of undifferentiated cells formed in the aftermath of amputation injury grows and develops the missing limb components through interactions with other limb tissues. As with limb development in embryos, the process of amphibian limb regeneration depends on an inductive influence from the thickened epithelium at the distal end of the growing appendage. In both developmental processes, proliferation of underlying cells is inhibited, and growth of the limb is blocked if the epithelium is removed or prevented from forming. Grafts of full-thickness skin over an amputation site preclude the interaction between epidermis and internal tissues and thereby inhibit regeneration. Deviation of the wound epithelium from its normal location results in asymmetrical outgrowth of the regenerated portion; grafting additional wound epithelia to a limb stump causes supernumerary limbs to develop. The epithelial effect in both embryonic and regenerating limbs is likely to involve release of growth-promoting factors and enzymes to break down extracellular components. The effect of these factors during regeneration is to transform a tissue-repair process into the process of blastema formation and epimorphic regeneration.

Another tissue interaction that is critical for appendage regeneration in amphibians involves the nerves in the limb stump. Cut ends of axons in motor, sensory, and autonomic nerves grow into the developing blastema and are present throughout this tissue during its period of active growth. If this elongation of axons is prevented or curtailed by cutting the nerves at a more proximal level, further growth of the early blastema is inhibited. Limbs of larval salamanders will resume regeneration when the nerves themselves regenerate to distal areas of the limb stump. Moreover, limbs without their normal supply of nerves can regenerate if neurons are grafted into the stump.

The mechanism by which nerves exert their stimulatory influence on regeneration involves secretion by axons of nutritive or growth-triggering factors. Although these factors have not yet been identified with certainty, the effect is clearly not dependent on a specific type of nerve: sensory, motor, and sympathetic nerves can all support blastema formation if present in sufficient quantity. The quantitative requirement is indicated by the observation that a limb will not regenerate unless it contains a certain minimum number of nerve fibers at the level of amputation. The number of nerve fibers necessary to support regeneration varies with the size of the limb and level of amputation. The neural influence is important primarily for mitotic activity of the dedifferentiated cells leading to establishment and growth of the blastema. Removal of nerves from a limb does not inhibit epidermal closure of an amputation site or the onset of internal tissue dissociation, nor does cutting the nerves after a well-developed blastema has

appeared prevent redifferentiation and formation of the replacement limb. Most available evidence, both from analyses of proliferative activity of blastema cells and from studies on proteins transported and released from regenerating axons, is consistent with the view that nerves release factors specifically required for DNA replication in proliferating cells of the early blastema.

Artificial induction of limb regeneration in vertebrates. Various degrees of regenerative growth have been stimulated in certain nonregenerating animals by manipulation of the factors important for regeneration in salamanders. For example, regeneration has been reported in the limbs of adult frogs following amplification of the current leaving the stumps, after localized applications of growth factors, and after augmentation of the nerve supply to the limb. Grafts of neural tissue to limb stumps have also been shown to induce regeneration in lizards.

Most attempts to induce regeneration in warm-blooded vertebrates have been relatively unsuccessful, with the exception of the regeneration of lost fingertips in young children. Restorative growth of the digit depends on how the injury is treated: the wound must be allowed to close itself by simple epithelial growth and contraction, rather than being surgically closed with sutures. The importance of how the amputation wound heals suggests that human fingertip regeneration involves a positive influence from the epidermis on the underlying dissociated tissues, like that normally operative during amphibian limb regeneration.

Morphogenesis during limb regeneration. The manner in which specific limb tissues, such as individual muscles and bones, assume their unique morphological structures and locations during reconstitution of the complete appendage is a major unanswered question in regeneration biology.

Research into such questions of morphogenesis during amphibian limb regeneration is relevant to understanding similar events of pattern formation that occur during embryonic limb development. An experimental tool for analyzing these events was discovered with the observation that retinoids, a class of compounds related to vitamin A, are able to respecify in a predictable manner the positions in which muscles and skeletal elements develop in both embryonic and regenerating vertebrate limbs. Local application of retinoic acid at a specific site on the embryonic limb bud leads to a duplication in the anterior-posterior axis of the limb, so that six digits develop instead of the normal three. Application of retinoic acid during amphibian limb regeneration affects all axes of the developing limb, but the effect on the proximal distal axis, which leads to the duplication of structures present proximal to the level of amputation, has been best studied. For example, amputation through the wrist of a salamander and treatment with vitamin A can result in a blastema at the wrist that will give rise to structures of the upper arm and lower arm, as well as to a complete new hand (**Fig. 2**). Observations that concentration gradients of retinoic acid and binding proteins for
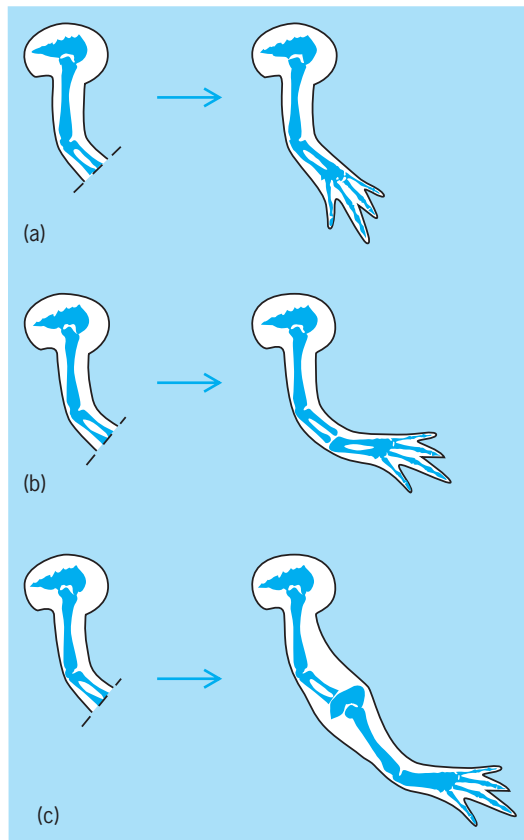
Fig. 2. Effect of a retinoid such as vitamin A on the proximal-distal axis of a regenerated amphibian forelimb. (*a*) Normal regeneration of the forelimb after amputation through the radius and ulna (broken line). (*b*) Regeneration after treatment with a low dose of vitamin A; the duplicate radius and ulna have formed in the regenerated limb. (*c*) Regeneration after treatment with a high dose of vitamin A; the entire forelimb, including the scapula, has been duplicated in the regenerated limb.

retinoic acid exist in developing limb buds suggest that retinoids may be among the endogenous factors, or morphogens, involved in the control of limb pattern formation. *See* PATTERN FORMATION (BIOLOGY).

The apparent involvement of retinoids in the control of limb morphogenesis has provided an opportunity for applying the methods of molecular biology in studies of regeneration. Genes for retinoic acid receptors, proteins to which retinoids bind before exerting their effect on expression of a cell's genetic program, have been cloned from salamanders. These genes are expressed in dedifferentiating cells of the limb stump as the blastema begins to form, and are important in establishing the positional identity of the cells within the new limb, although the mechanism by which retinoic acid and its receptors effect positional information in the regenerating limb remains unclear. Other genes implicated in pattern formation in other developing systems have also been shown to be active in regeneration blastemas, and differences in the rate of expression related to the proximal-distal location of the blastema cells have been observed. The ability to measure expression of such genes by molecular methods, together with the use of retinoid treatments and surgical methods

to manipulate morphogenesis in limbs, should lead to an increased understanding of pattern formation during limb regeneration and development. *See* ANIMAL MORPHOGENESIS; VITAMIN A.

Tail regeneration. A well-developed capacity for tail regeneration is found in larval amphibians, as well as in certain species of adult salamanders. Many lizards can also regenerate tails lost through injury or by autotomy during escape from a predator. Autotomy is produced by a strong contraction of the segmental muscles of the tail, causing one of the vertebrae to come apart along a preformed breakage plane and separating the muscles and other tissues so that the distal portion of the tail is detached from the body.

Whether lost by autotomy or injury, the tail is regenerated by a process superficially similar to that of amphibian limbs. Tissues dissociate locally in the injured area of the stump, during which cells dedifferentiate and begin the period of proliferation that eventually gives rise to the regeneration blastema. Formation of new muscles, axial skeleton, and connective tissue from blastemal cells proceeds in a proximal-to-distal direction. Depending on the species, the morphology and tissue composition of the regenerated tail are abnormal to various degrees. In the lizard, vertebrae are not reconstituted; instead, the regenerated spinal cord is surrounded by a simple tube of flexible cartilage.

An important developmental process unique to tail regeneration is reconstitution of the spinal cord. Cells lining the central neural canal at the severed end of the cord grow out to form a tubular structure that extends into the developing blastema. Neurons and nerve fibers from proximal, intact regions of the spinal cord grow along this tubular outgrowth, and a functional though anatomically abnormal cord is reestablished. In salamanders, but not in lizards, paired spinal ganglia are also formed along the new cord, developing from cells that migrate out of the regenerating cord when it is a simple growing tube of cells.

In all lizards and amphibians (with the apparent exception of larval frogs), an inductive influence from the spinal cord is necessary for tail blastema formation and regeneration of axial structures. If the spinal cord is removed from the site of amputation and prevented from regrowing into that area, regeneration of the tail will not occur. The neurons and other cells around the central canal of the spinal cord appear to be the most important source for this growth-promoting influence. Spinal ganglia and peripheral nerves deviated into the tail stump are unable to replace the spinal cord in allowing growth of a tail regeneration blastema. This property of the spinal cord is similar to the activity of peripheral nerves during limb regeneration, but whether the same chemical factors and cellular interactions are involved is not known. *See* NERVOUS SYSTEM (VERTEBRATE).

Replacement of structures in the eye. Other noteworthy examples of regeneration found primarily in amphibian species involve certain structures of the eye. Surgical removal of the neural retina can be followed by reconstitution of this tissue from the

deeper pigmented layer of the retina. From the new retina a new optic nerve is regenerated. Removal of the iris is followed by its complete replacement, including new muscles and connective tissue, from the edge of the adjacent retina. With various species of salamanders, lens removal initiates a remarkable process by which epithelial cells on the dorsal rim of the iris dedifferentiate and proliferate as a mass of cells that develops into a new lens. This phenomenon is dependent on a neural influence from the retina. Certain larval frogs regenerate a missing lens from epithelium of the cornea, which more closely resembles lens formation in vertebrate embryos. Such transformations of one specialized ocular tissue into another make possible the reconstitution of a fully functional visual system. *See* EYE; VISION.

Regeneration in invertebrates. Protozoa and simple multicellular animals, including sponges, cnidaria (for example, *Hydra*), and flatworms (for example, *Planaria*), display remarkable capacities for regeneration following various experimental manipulations. Regenerative ability in such organisms correlates closely with their capacity to reproduce asexually, most commonly by fission or by budding, and the mechanisms of growth involved in regeneration are often very similar to those of asexual propagation. For example, just as complete ciliated protozoa will develop after fission, which divides the nucleus and organelles between daughter cells, intact individuals will also be reconstituted from fragments of a single organism if the fragment contains a complete set of the genetic material and a portion of the original cell's cortical cytoplasm. Similar rules regarding the importance of the nucleus apply to regenerative processes in all protozoa.

Porifera. Regeneration in sponges can be considered a variant of the method of asexual reproduction used by these primitive multicellular organisms. If part of a sponge is cut away, that portion will usually not grow back to any significant extent. However, if a sponge is mechanically or chemically dissociated into its component cells, a population of reserve cells called amebocytes will reaggregate and give rise to an intact new organism. This regenerative phenomenon is similar to the asexual process of gemmule formation in sponges. In this reproductive procedure, cells are packaged in a special structure (gemmule), which enables them to survive the death of the organism during certain adverse conditions and to emerge later and reconstitute an intact sponge. *See* PORIFERA.

Cnidaria. Among cnidaria, hydroids and sea anemones have considerable regenerative capacity, whereas jellyfish regenerate poorly. Regeneration in fresh-water *Hydra* has been studied extensively. Such animals have a definite anatomical polarity, with a base for attachment to a substrate and a hypostome (mouth) surrounded by tentacles at the opposite end. If a hydra is cut in half, the part containing the hypostome and tentacles will form a new basal half for attachment, and the original basal part will form a mouth and tentacles for a new apical end. A dissected central portion of the animal,

having neither tentacles nor basal structures, will regenerate a new base at one end and tentacles at the other. Even tiny fragments of an individual hydra can regenerate all the parts for complete new organisms. Polarity is maintained not only during regeneration but also during the process that allows individuals to reconstitute themselves when several hydra are grafted surgically end to end.

Regeneration in *Hydra* usually involves growth and differentiation of interstitial cells situated among the cells of the organism's external layer. These cells represent a reserve population important for the steady replacement of certain specialized cells and for asexual reproduction by budding. Polarity appears to be maintained during regeneration and budding by the action of two morphogens, a head activator and a head inhibitor, which are present in gradients along the length of the organism and which together control formation of the hypostome and tentacles. *See* CNIDARIA.

Planaria. Gradients of regulative factors also appear to control regeneration in flatworms, such as *Planaria*. An adult flatworm is highly organized, having a head with eyes and a simple brain, complex digestive, reproductive, and excretory organs, and a tail. Regeneration of complete flatworms from dissected pieces occurs in a polarized fashion, provided a head-tail gradient of morphogens is present (**Fig 3**). New body parts are formed during regen-
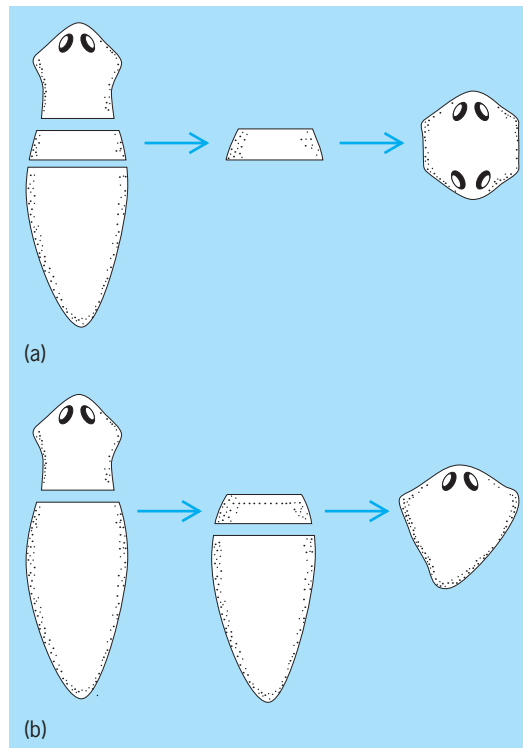

Fig. 3. Relationship of the anterior-posterior gradient to the product of *Planaria* sectioning. (*a*) A thin section of *Planaria* too narrow to establish an anterior-posterior gradient may form two heads facing opposite directions. (*b*) If the posterior cut is delayed, the thin section will form a worm of normal polarity because the delay has allowed an anterior-posterior gradient to be established. (*After R. J. Goss, Principles of Regeneration, Academic Press, 1969*)

eration from a population of reserve cells called neoblasts, which reside in connective tissue below the epidermis. This mechanism of growth in flatworms is similar to that following reproductive fission in these animals. Species incapable of asexual propagation also have poorly developed regenerative capacities. *See* TURBELLARIA.

Annelida. Most annelids, such as the earthworm, can readily regenerate segments after their removal: some species can regenerate whole organisms from any fragment. Like more primitive invertebrates, certain annelids can reproduce asexually by transverse fission. The capacities for fission and for reconstitution from fragments in annelids are remarkable, considering the anatomical complexity of animals in this phylum. Annelids have well-developed muscular systems associated with primitive exoskeletons and efficient digestive systems. The nervous system consists of an anterior brain, complex sense organs, and a highly organized network of motor and sensory nerves. When an earthworm is cut transversely into two parts, the anterior part can regenerate several posterior segments. The ability of the posterior half to regenerate anterior segments is, however, more limited and is absent altogether in some species. Experiments in which components of the nervous system are removed surgically have revealed the importance of a neural influence for segment regeneration, presumably mediated by a growth-stimulating hormone secreted from neural cells. In most species of annelids, cells of the regenerated segments are derived both from populations of reserve cells and from dedifferentiation of existing tissues. *See* ANNELIDA.

Echinodermata. The ability of certain echinoderms, such as starfish, to regenerate missing arms is well known. Cutting such an animal into several pieces results in each piece forming a new organism, a phenomenon that usually requires the presence of at least some of the central portion of the body. Some starfish species, however, can actually regenerate an entire body from the cut end of an isolated arm. Equally remarkable is a regenerative response shown by another echinoderm, the sea cucumber. When this animal is strongly irritated, it undergoes the unusual behavior of eviscerating itself through its anus or through a rupture of its body wall. This phenomenon, which may serve as a protective mechanism, produces a nearly empty sack of skin and muscle, which then proceeds to regenerate all the internal organs, beginning with the digestive tract. *See* ECHINODERMATA.

Arthropoda. The capacity for appendage regeneration is widespread among the many diverse members of the phylum *Arthropoda*. In these complex animals with well-developed exoskeletons and no asexual mode of reproduction, regeneration shows a close correlation with the molting process. Insects regenerate legs and antennae readily during their long larval and pupal stages, but lose this capacity as adults. Crustaceans, such as lobsters, crayfish, and crabs, continue to molt throughout life and retain regenerative ability as adults, regrowing missing appendages at each molt. Many species of crustaceans have capitalized on this regenerative ability by developing the capacity for autotomy of legs caught by predators. Regeneration in other forms of arthropods, including centipedes, millipedes, and spiders, occurs to varying degrees but has not been as well studied as that in crustaceans and insects. *See* ARTHROPODA; DEVELOPMENTAL BIOLOGY.   Anthony Mescher

Bibliography. C. E. Dinsmore (ed.), *A History of Regeneration Research: Milestones in the Evolution of a Science*, 1991; P. Ferretti and J. Geraudie (eds.), *Cellular and Molecular Basis of Regeneration: From Invertebrates to Humans*, 1998; H. G. Gard and M. T. Langaker (eds.), *Scarless Wound Healing*, 2000; L. V. Polazhaev, *Loss and Regulation of Regenerative Capacity in Tissues and Organs of Animals*, 1990; D. L. Stocum, *Wound Repair, Regeneration and Artificial Tissues*, 1994; P. A. Tsonis, *Limb Regeneration*, 1996; I. V. Yannas, *Tissue and Organ Regeneration in Adults*, 2001.

Regolith

The mantle or blanket of unconsolidated or loose rock material that overlies the intact bedrock and nearly everywhere forms the land surface. The regolith may be residual (weathered in place), or it may have been transported to its present site. The undisturbed residual regolith may grade from agricultural soil at the surface, through fresher and coarser weathering products, to solid bedrock several meters or more beneath the surface (see **illus.**). The
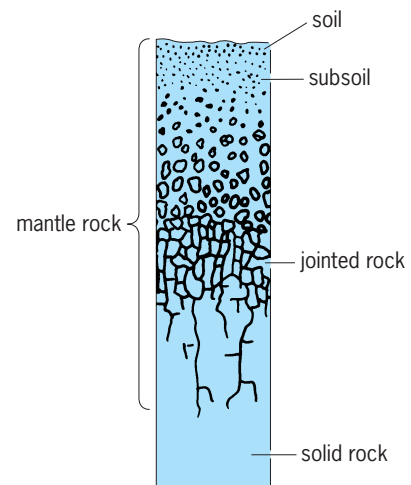

Diagram showing the components of residual regolith.

transported regolith includes the alluvium of rivers, sand dunes, glacial deposits, volcanic ash, coastal deposits, and the various mass-wasting deposits that occur on hillslopes. Scientific results from the Apollo space program show that the lunar surface also has a regolith. This layer of fragmental debris is believed to derive from prolonged meteoritic and secondary fragment impact. *See* SOIL; WEATHERING PROCESSES.
                                        Victor R. Baker

Regular polytopes

The *n*-dimensional analogs of the regular polygons ($n = 2$) and platonic solids ($n = 3$). They are conveniently denoted by their Schläfli symbols $\{p, q, \ldots\}$; for instance, the pentagon, hexagon, octagon, tetrahedron, octahedron are denoted by $\{5\}$, $\{6\}$, $\{8\}$, $\{3,3\}$, $\{3,4\}$. The cube is $\{4,3\}$ because its faces are squares $\{4\}$ and there are three of them at each vertex. The five platonic solids $\{p,q\}$, which are the subject of Euclid's *Elements*, Book XIII, are determined by inequality (1). The numbers of ver-

$$(p - 2)(q - 2) < 4 \qquad (1)$$

tices, edges, faces (V,E,F), as listed in the **table**, can be deduced from the obvious relations $pF = 2E = qV$ with the help of Euler's formula $V - E + F = 2$.

The general polytope [sometimes loosely called a polyhedron (**Fig. 1**) regardless of the number of dimensions] is a finite region of *n*-dimensional space enclosed by a finite number of hyperplanes. When any redundant hyperplanes have been discarded, those that remain contain $(n - 1)$-dimensional polytopes called cells or "facets." For instance, the cells of a polygon are its sides, those of a polyhedron are its faces, and those of a four-dimensional polytope are solids.

The platonic solid $\{p,q\}$ is said to be regular because its faces are regular and its vertices are all surrounded alike. The four-dimensional regular polytope $\{p,q,r\}$ has three-dimensional solid cells $\{p,q\}$, r of which surround each edge; the five-dimensional regular polytope $\{p,q,r,s\}$ has cells $\{p,q,s\}$, s of which surround each plane face; and so on.

The six regular four-dimensional polytopes $\{p,q,r\}$ are determined by inequality (2). The 5-cell $\{3,3,3\}$

$$p - \frac{4}{p} + 2q + r - \frac{4}{r} < 12 \qquad (2)$$

(**Fig. 2a**) may be drawn in perspective as a pentagon
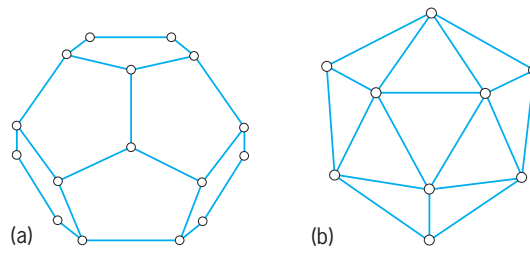
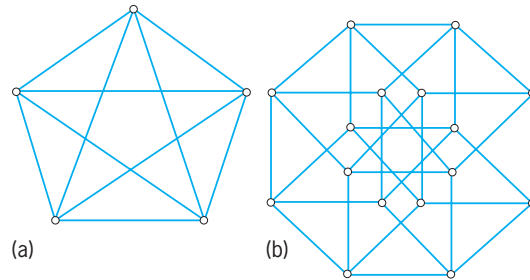Fig. 1. Two regular polyhedrons. (*a*) Dodecahedron $\{5,3\}$. (*b*) Icosahedron $\{3,5\}$.

Fig. 2. Two of the six regular four-dimensional polytopes. (*a*) The 5-cell $\{3,3,3\}$. (*b*) The 8-cell $\{4,3,3\}$.

with its 5 diagonals, though in reality its 10 edges are all equal. The simplest drawing of the 8-cell $\{4,3,3\}$ (Fig. 2*b*) consists of an octagon with a square placed inward on each side. The squares on two alternate sides are easily visualized as two opposite faces of a cube. The 8 such cubes are the cells of the 8-cell.

The regular tetrahedron can be inscribed in the cube, in the sense that the 4 vertices of the former occur among the 8 vertices of the latter. In the same sense, the cube can be inscribed in the dodecahedron, the 16-cell in the 8-cell, the 8-cell in the 24-cell, the 24-cell in the 600-cell, the 600-cell (and also the 5-cell) in the 120-cell.

The vertices of the *n*-dimensional simplex $\{3,3, \ldots,3\}$ consists of $n + 1$ points, all equidistant from one another. Those of the cross polytope $\{3,\ldots,3,4\}$ are at equal distances from the origin in both

Regular polytopes in *n* dimensions						
Polytope	Schläfli symbol	Vertices	Edges	Faces	Solid cells	Hypersolid cells
n=2						
p-gon	$\{p\}$	p	p			
n=3						
Tetrahedron	$\{3,3\}$	4	6	4		
Cube	$\{4,3\}$	8	12	6		
Octahedron	$\{3,4\}$	6	12	8		
Dodecahedron	$\{5,3\}$	20	30	12		
Icosahedron	$\{3,5\}$	12	30	20		
n=4						
5-cell	$\{3,3,3\}$	5	10	10	5	
8-cell	$\{4,3,3\}$	16	32	24	8	
16-cell	$\{3,3,4\}$	8	24	32	16	
24-cell	$\{3,4,3\}$	24	96	96	24	
120-cell	$\{5,3,3\}$	600	1200	720	120	
600-cell	$\{3,3,5\}$	120	720	1200	600	
n > 4						
Simplex	$\{3,3,\ldots,3\}$	$n + 1$	$\frac{1}{2}n(n + 1)$	....		$n + 1$
Hypercube	$\{4,3,\ldots,3\}$	2^n	$2^{n-1}n$	....		$2n$
Cross polytope	$\{3,\ldots,3,4\}$	$2n$	$2n(n - 1)$	....		2^n

directions along the n coordinate axes; thus their coordinates (for a cross polytope of edge $\sqrt{2}$) are the permutations of $(\pm 1,0,\ldots,0)$. Similarly, the 2^n vertices of the hypercube $\{4,3,\ldots,3\}$, of edge 2, are $(\pm 1,\pm 1,\ldots,\pm 1)$. *See* ANALYTIC GEOMETRY; EUCLIDEAN GEOMETRY.   H. S. M. Coxeter

Bibliography. A. Bronosted, *An Introduction to Convex Polytopes*, 1982; H. S. M. Coxeter, *Regular Complex Polytopes*, 2d ed., 1991; H. S. M. Coxeter, *Regular Polytopes*, 2d ed., 1963, reprint 1973; K. Miyazak, *An Adventure in Multidimensional Space: The Art and Geometry of Polygons, Polyhedra, and Polytopes*, 1986; G. M. Ziegler, *Lectures on Polytopes*, 1994.

Regularia

The name given by G. Cuvier in 1817 to an assemblage of echinoids in which the anus and periproct lie within the apical system. The test is globular and preponderantly radially symmetrical, and the ambulacral plates are commonly compound. The group included, in effect, all those echinoids which did not fall in the Irregularia. The Irregularia, however, were shown to be polyphyletic by J. Durham and R. Melville in 1957, and so neither the Irregularia nor the Regularia constitute valid taxa. *See* ECHINOIDEA; IRREGULARIA.   Howard B. Fell

Regulator

A control device designed to maintain the value of some quantity substantially constant. Thus, a temperature regulator is a device designed to maintain the temperature of some environment at a constant value. The value to be maintained can usually be established at any value within the range of the regulator by making an appropriate setting.

A regulated system is a feedback control system employing a regulator to maintain some quantity of the system at a constant value. Another example is the voltage-regulator system of an automobile. *See* CONTROL SYSTEMS.   John A. Hrones

Reheating

The addition of heat to steam of reduced pressure after the steam has given up some of its energy by expansion through the high-pressure stages of a turbine. The reheater tube banks are arranged within the setting of the steam-generating unit in such relation to the gas flow that the steam is restored to a high temperature. Under suitable conditions of initially high steam pressure and superheat, one or two stages of reheat can be advantageously employed to improve thermodynamic efficiency of the cycle. *See* STEAM-GENERATING UNIT; STEAM TURBINE; SUPERHEATER; VAPOR CYCLE.   R. A. Miller
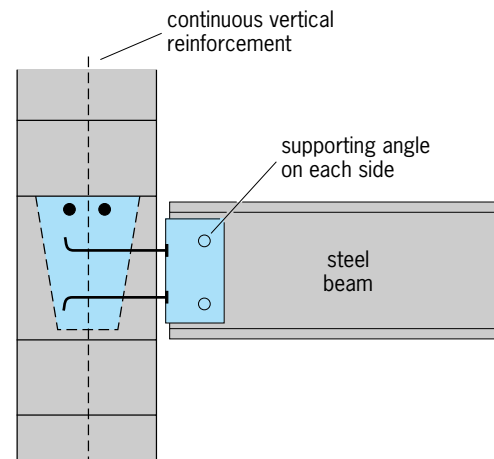
Reinforced concrete

Portland cement concrete containing higher-strength, solid materials to improve its structural properties. Generally, steel wires or bars are used for such reinforcement, but for some purposes glass fibers or chopped wires have provided desired results.

Unreinforced concrete cracks under relatively small loads or temperature changes because of low tensile strength. The cracks are unsightly and can cause structural failures. To prevent cracking or to control the size of crack openings, reinforcement is incorporated in the concrete. Reinforcement may also be used to help resist compressive forces or to improve dynamic properties.

Steel usually is used in concrete. It is elastic, yet has considerable reserve strength beyond its elastic limit. Under a specific axial load, it changes in length only about one-tenth as much as concrete. In compression, steel is more than 10 times stronger than concrete, and in tension, more than 100 times stronger. *See* STEEL.

As reinforcement for concrete, steel may be used in the form of bars or rods, wire, fibers, pipe, or structural shapes such as wide-flange beams (see **illus.**). Bars are the most commonly used form. The size of a bar usually is specified by a number that is about eight times the nominal diameter; sizes range from No. 2, a nominal $\frac{1}{4}$-in. (0.6-cm) bar, to No. 18, a nominal $2\frac{1}{4}$-in. (5.7-cm) bar.

During construction, the bars are placed in a form and then concrete from a mixer is cast to embed them. After the concrete has hardened, deformation is resisted and stresses are transferred from concrete to reinforcement by friction and adhesion along the surface of the reinforcement. The bonding may be improved mechanically by giving reinforcing bars raised surfaces; such bars are known as deformed bars. They are generally manufactured to standard specifications, such as those of the American Society for Testing and Materials (ASTM). Similarly, deformed wires used as reinforcement provide greater bond than smooth wires. Deformed wires

Continuous vertical reinforcement in concrete. (*After C. Beall, Masonry Design and Detailing for Architects, Engineers, and Contractors, 3d ed., McGraw-Hill, 1993*)

are designated by D followed by a number equal to 100 times the nominal area in square inches.

Individual wires or bars resist stretching and tensile stress in the concrete only in the direction in which such reinforcement extends. Tensile stresses and deformations, however, may occur simultaneously in other directions. Therefore reinforcement must usually be placed in more than one direction. For this purpose, reinforcement sometimes is assembled as a rectangular grid, with clips or welded joints at the intersections of the wires or bars. For example, prefabricated wire grids, called welded-wire fabric, often are used for slab reinforcement in highway pavement and in buildings.

Bars, grids, and fabric have the disadvantage that the principal effect of reinforcement occurs primarily in the plane of the layer in which they are placed. Consequently, the reinforcement often must be set in several layers or formed into cages.

Under some conditions, fiber-reinforced concrete is an alternative to such arrangements. The fiber, made of fine steel wires, glass fibers, or plastic threads, is embedded in the concrete in short lengths, often only about 1 in. (2.5 cm). They are added to the concrete mixer along with the other ingredients. To achieve crack control in all directions, the fibers should be uniformly spaced, at close intervals, and randomly oriented throughout the hardened concrete. Such reinforcement also improves concrete tensile strength, ductility, and dynamic properties. *See* COMPOSITE BEAM; CONCRETE; CONCRETE BEAM; CONCRETE COLUMN; CONCRETE SLAB; PRESTRESSED CONCRETE.   Frederick S. Merritt

Bibliography. C. Beall, *Masonry Design and Detailing: For Architects, Engineers, and Contractors*, 4th ed., 1997; K. C. Hover, *Construction with Reinforced Concrete*, 1994; R. R. Schneider and W. L. Dickey, *Reinforced Masonry Design*, 3d ed., 1994.

Relapsing fever

An acute infectious disease characterized by recurring fever. It is caused by spirochetes of the genus *Borrelia* and transmitted by the body louse (*Pediculus humanus humanus*) and by ticks of the genus *Ornithodoros*.

Louse-borne relapsing fever, caused by *Borrelia recurrentis*, is typically epidemic. Its occurrence depends upon ecological and socioeconomic conditions favoring heavy infestations by body lice. Humans are the only reservoir. Epidemics, once widespread on all continents, are rare but still occur in certain parts of South America, Africa, and Asia.

Tick-borne relapsing fevers are endemic. They are more widely distributed throughout the Eastern and Western hemispheres. At least 15 species of *Borrelia* have been recognized as causative agents.

After incubation of 2–10 days, the initial attack begins abruptly with chills, high fever, headache, and pains in muscles and joints, and lasts 2–8 days, ending by crisis. A remission period of 3–10 days is followed by a relapse similar to the initial attack but milder. There may be 4–5 relapses, although occasionally 10 or more have been recorded as a result of antigenic variations in the causative *Borrelia*. Mortality varies from 2 to 5% but may be considerably higher during epidemics.

Diagnosis and treatment. Definite diagnosis depends upon detecting spirochetes in the blood during fevers. They may be seen by dark-field examination of wet blood preparations, or in Giemsa-stained thick and thin blood films. Where symptoms are present but spirochetes are absent or scanty, intraperitoneal inoculation of adult or suckling Swiss mice or rats with blood is useful to amplify the spirochetes. Detection of spirochetes in lice or ticks is by dark-field examination of hemolymph or tissues, and by inoculation of tissue suspensions into suckling and adult Swiss mice or rats.

Ability of borreliae to undergo antigenic phase variations make the serodiagnosis of relapsing fevers difficult. Of the many tests proposed, indirect immunofluorescence with cultured spirochetes gives the most dependable results. *See* IMMUNOFLUORESCENCE.

Chlortetracycline is the most effective antibiotic drug, but penicillin, oxytetracycline, and streptomycin also have therapeutic value. *See* ANTIBIOTIC.

Transmission and prevention. In lice, *B. recurrentis* produces an infection limited to the hemolymph. Thus, transmission does not occur by bite but by contamination of the bite wound with infectious hemolymph of lice that are smashed by scratching. Borreliae may penetrate normal, unbroken skin.

Tick-borne borreliae, on the other hand, produce in their tick vectors a generalized infection, including tissues of salivary glands. Therefore, transmission is via infected saliva or infectious body fluids (coxal fluid) excreted during feeding. In several tick vectors, the presence of spirochetes in ovarian tissues leads to the passage of organisms via eggs (transovarian infection).

The best way to prevent relapsing fever is to control louse and tick populations with effective insecticides and acaricides. *See* MEDICAL BACTERIOLOGY; PEDICULOSIS.   Willy Burgdorfer

Bibliography. M. Kusinitz, *Tropical Medicine*, 1990.

Relative atomic mass

The ratio of the average mass per atom of the natural nuclidic composition of an element to $1/12$ of the mass of an atom of nuclide ^{12}C. For example, $\mu(Cl) = 35.453$. Relative atomic mass replaces the concept of atomic weight. It is also known as relative nuclidic mass. *See* ATOMIC MASS; NUCLIDE.   Thomas C. Waddington

Relative molecular mass

The ratio of the average mass per formula unit of the natural nuclidic composition of a substance to $1/12$ of the mass of an atom of nuclide ^{12}C. For example,

μ(KCl) = 74.555. Relative molecular mass replaces the concept of molecular weight. *See* MOLECULAR WEIGHT; NUCLIDE.   Thomas C. Waddington

Relative motion

All motion is relative to some frame of reference. The simplest laboratory frame of reference is three mutually perpendicular axes at rest with respect to an observer. Such a system is commonly used in the laboratory when various types of motion are being studied. The general effects of other motions to which the system as a whole is subjected are then neglected and the system is said to be isolated. In terms of the frame of reference of an observer some distance from Earth, the laboratory frame of reference would be moving with Earth as it rotates on its axis and as it revolves about the Sun. A simple form of motion in the laboratory frame of reference would appear to be a much more complicated motion in the frame of reference of the distant observer. *See* FRAME OF REFERENCE.

Motion means continuous change of position of an object with respect to an observer. To another observer in a different frame of reference the object may not be moving at all, or it may be moving in an entirely different manner. The motions of the planets were found in ancient times to appear quite complicated in the laboratory frame of reference of an observer on Earth. By transferring to the frame of reference of an imaginary observer on the Sun, Johannes Kepler showed that the relative motion of the planets could be simply described in terms of elliptical orbits. The validity of one description is no greater than the other, but the latter description is far more convenient. *See* PLANET.

Relative velocity. That motion is relative to an observer must have been implicit in the earliest ideas of motion. In the mechanics of Galileo and Isaac Newton these ideas became clarified, and methods were developed for finding the relative velocity of two bodies, each moving with a different velocity. If the velocity of one body is v_1, represented in **Fig. 1** by the magnitude and direction of the vector v_1, and if a second body has a velocity v_2 represented by the vector v_2, then v_3 is the vector to be added to v_1 to make the sum equal to v_2, and, consequently, the vector v_3 is equal to the vector difference $v_2 - v_1$. Therefore, the relative velocity of the second body with respect to the first is that velocity represented in magnitude and direction by the vector v_3, all vectors being drawn to a suitable scale.
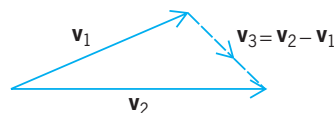

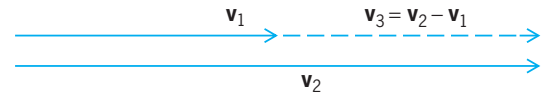Fig. 1. Relative velocity of two bodies moving at an angle with respect to one another.

Fig. 2. Relative velocity of two bodies moving in the same direction; that is, velocities are parallel.
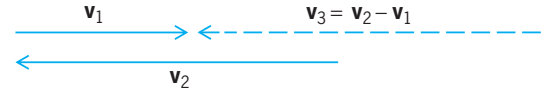

Fig. 3. Relative velocity of two bodies moving in opposite directions; that is, velocities are antiparallel.

In the simplest case the velocities v_1 and v_2 are parallel (**Fig. 2**). The relative velocity of the second body with respect to the first is again v_3, and its magnitude is the difference in the numerical values of v_2 and v_1. When the two velocities are antiparallel (in opposite directions, **Fig. 3**), one is negative with respect to the other, and the magnitude of the relative velocity of v_2 with respect to v_1 is again the vector difference v_3, which is now the numerical sum of the two. To obtain the relative velocity of v_1 with respect to v_2, the arrowhead on v_3 should be reversed. For example, when two automobiles move with velocities of 40 and 60 mi/h (20 and 30 m/s) respectively in the frame of reference of an observer stationed by the roadside, the relative velocity of the two is 20 mi/h (10 m/s). If the automobiles were moving in opposite directions, the relative velocity would be 100 mi/h (50 m/s).

The flight of an airplane illustrates the principles of relative motion. The common statement that the speed of an airplane is, for instance, 300 mi/h (150 m/s) is essentially a meaningless statement since the speed is not designated with any particular frame of reference in mind and the listener must make an assumption that it is perhaps the ground speed in still air. For the airplane to sustain itself with normal lift, it is the airspeed or speed with respect to the air that is important. In a head wind of 100 mi/h (50 m/s) the airplane would have a ground speed of 200 mi/h (100 m/s). In a similar tail wind it would have a ground speed of 400 mi/h (200 m/s). If the wind velocity is at an angle to the direction of flight, the relative velocities must be considered. The velocity with respect to the ground would be the vector sum of the velocity of the plane with respect to the air and the velocity of the wind with respect to the ground. *See* FLIGHT CHARACTERISTICS.

Relative acceleration. Acceleration, like velocity, is relative to the observer's frame of reference. An automobile starting from rest is accelerated with respect to the Earth, but the driver of the car does not see himself or herself or the car accelerated forward. The driver sees objects at rest with respect to the roadway accelerated backward with respect to himself or herself and the car. Persons by the roadside see the driver and the car accelerated forward. In the driver's frame of reference, the car is at rest.

Acceleration is a vector quantity involving both magnitude and direction, just as velocity is. Just as the velocities of two objects may be represented by vectors and their relative velocity obtained by subtracting one vector from the other, so may the accelerations of two bodies be represented by vectors and the relative acceleration of one with respect to the other obtained by taking the vector difference. *See* ACCELERATION.

Relativity. Since, according to Einstein's theory of relativity, the velocity of light is the limiting velocity that any physical object can attain, the addition or subtraction of very large velocities cannot be accomplished by galilean-newtonian methods, and the rules to be followed are derived from relativistic theory. *See* MOTION; RELATIVITY.   Rogers D. Rusk

Relativistic electrodynamics

The study of the interaction between electrically charged particles and electromagnetic fields when the velocities involved are comparable to that of light.

Development. A group of charged particles in motion can be represented by a distribution in charge and distribution in current. During the latter part of the eighteenth century and the early part of the nineteenth century, experiments by C. A. Coulomb, M. Faraday, A. M. Ampère, and others showed that electric and magnetic fields are produced by charge and current distributions. These fields, in turn, act on other charges and currents. The interaction between charges and currents on the one hand and electric and magnetic fields on the other is the topic of study of electrodynamics. This field of study was established as a quantitative and self-contained subject in 1864 when J. C. Maxwell formulated his equations for the electromagnetic field. Maxwell conjectured that a time-varying electric field is equivalent to an electric current in its effect of producing a magnetic field, and named it the displacement current. The inclusion of this displacement current enabled Maxwell to combine all the previously established laws of electromagnetism into a coherent whole in his equations. *See* CLASSICAL FIELD THEORY; DISPLACEMENT CURRENT; MAXWELL'S EQUATIONS.

With the inclusion of the displacement current, the Maxwell equations are relativistically covariant, meaning that they are valid for all velocities, even those approaching the velocity of light. However, the implications of the covariance of the equations were not fully appreciated until A. Einstein formulated the special theory of relativity in 1905. Relativistic electrodynamics was then rapidly developed into a powerful and precise field of physics. It describes and predicts all macroscopic electrodynamic phenomena to the minutest detail and with perfect accuracy, and now forms the foundation on which the entire electrical industry is based. However, its limitations soon became evident when attempts were made to apply it to atomic phenomena: Straightfor-

wardly applied, relativistic electrodynamics failed to explain many of these phenomena, and its predictions frequently disagreed with experimental observations. For these microscopic phenomena, quantum electrodynamics (QED) was developed in the 1930s to replace classical relativistic electrodynamics. In 1967 quantum electrodynamics was further unified by S. Weinberg and A. Salam with the theory of weak interactions to form the electroweak theory. *See* QUANTUM ELECTRODYNAMICS; RELATIVISTIC MECHANICS; RELATIVITY; WEAK NUCLEAR INTERACTIONS.

Types of problems. Electrodynamic problems generally fall into one of two categories:

1. Finding the electromagnetic field produced by prescribed charge and current distributions. For example, one may want to determine the electromagnetic field produced or radiated by a given oscillatory electric current in a transmitting antenna, or the field radiated by an accelerating electron.

2. Finding the effect of a predetermined electromagnetic field on the motion of charges and currents. This is the inverse problem corresponding to that of the receiving antenna or of the motion of charged particles in an accelerator.

All other electrodynamic problems are combinations or iterations of these two basic types. For instance, the scattering of light (electromagnetic radiation) by a charged particle is composed of, first, the incident light shaking the charge and, second, the subsequent emission of the scattered light by the shaken charge. *See* SCATTERING OF ELECTROMAGNETIC RADIATION.

Electromagnetic field. The electromagnetic field produced by a given charge-current distribution is determined by Maxwell's equations. When written in separate space-time form, these are Eqs. (1), where

$$\vec{\nabla} \cdot \vec{E} = 4\pi\rho \qquad\qquad \vec{\nabla} \cdot \vec{B} = 0$$

$$\vec{\nabla} \times \vec{E} = -\frac{1}{c}\frac{\partial \vec{B}}{\partial t} \qquad \vec{\nabla} \times \vec{B} = \frac{1}{c}\frac{\partial \vec{E}}{\partial t} + \frac{4\pi}{c}\vec{J} \qquad (1)$$

$\vec{E}$ and $\vec{B}$ are the electric and the magnetic fields; ρ and $\vec{J}$ are the charge and the current density distributions; $\vec{\nabla}$, given by Eq. (2), is the space differential

$$\vec{\nabla} = \vec{i}\frac{\partial}{\partial x} + j\frac{\partial}{\partial y} + k\frac{\partial}{\partial z} \qquad (2)$$

operator; and $c \cong 3 \times 10^8$ m/s is the speed of light. As customary in this subject, all quantities are expressed in gaussian units. *See* ELECTRICAL UNITS AND STANDARDS; LIGHT.

The fact that these equations are relativistically covariant is demonstrated by showing that they can be written in a four-dimensional (three space dimensions and one time dimension) tensor form, because tensor equations have the same identical form in all frames of reference. For this, the coordinate 4-vector with imaginary time is used, namely $x_\mu = (x, y, z, ict)$ or (x_1, x_2, x_3, x_4). The field 4-tensor is defined by Eq. (3), and the charge-current density 4-vector is

given by Eq. (4), meaning that, for example, F_{12} is

$$F_{\mu\nu} = \begin{pmatrix} 0 & B_z & -B_y & -iE_x \\ -B_z & 0 & B_x & -iE_y \\ B_y & -B_x & 0 & -iE_z \\ iE_x & iE_y & iE_z & 0 \end{pmatrix} \tag{3}$$

$$J_\mu = (\vec{J}, ic\rho) \tag{4}$$

B_z, J_1 is J_x, and J_4 is $ic\rho$. Maxwell's equations (1) can now be written as Eqs. (5), as may be verified by

$$\frac{\partial F_{\mu\nu}}{\partial x_\nu} = \frac{4\pi}{c} J_\mu$$

$$\frac{\partial F_{\mu\nu}}{\partial x_\lambda} + \frac{\partial F_{\lambda\mu}}{\partial x_\nu} + \frac{\partial F_{\nu\lambda}}{\partial x_\mu} = 0 \tag{5}$$

taking special values of the indices. A sum from 1 to 4 is understood to be made on indices repeated in a term, as the index ν in the first of Eqs. (5). *See* CALCULUS OF VECTORS.

If the antisymmetric field tensor is expressed in terms of the 4-potential $A_\mu = (\vec{A}, i\phi)$ by Eq. (6),

$$F_{\mu\nu} = \frac{\partial A_\nu}{\partial x_\mu} - \frac{\partial A_\mu}{\partial x_\nu} \tag{6}$$

the second of Eqs. (5) is identically satisfied. Equation (6), however, does not define A_μ uniquely, and in addition the Lorentz condition, Eq. (7), can

$$\frac{\partial A_\nu}{\partial x_\nu} = 0 \tag{7}$$

be imposed. The first of Eqs. (5), then, becomes Eq. (8). It can easily be verified that Eq. (8) and

$$\frac{\partial^2 A_\mu}{\partial x_\nu^2} = \left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)A_\mu = -\frac{4\pi}{c} J_\mu \tag{8}$$

the Lorentz condition are indeed consistent with the continuity equation for the 4-current, Eq. (9). Equa-

$$\frac{\partial J_\mu}{\partial x_\mu} = 0 \tag{9}$$

tion (8) shows that the potential wave $A_\mu(\vec{x}, t)$ generated by the current source $J_\mu(\vec{x}', t')$ travels with a phase velocity equal to the velocity of light. This led Maxwell to infer that light is an electromagnetic wave. Furthermore, electromagnetic waves span a very broad frequency spectrum from a few hertz for alternating-current waves, to kilohertz and megahertz radio waves, 10^{10} Hz for microwaves, 10^{15} Hz for visible light, 10^{20} Hz for x-rays, to 10^{23} Hz for gamma rays and beyond. *See* ELECTROMAGNETIC RADIATION; EQUATION OF CONTINUITY; POTENTIALS.

Electromagnetic waves always propagate at the speed of light. Only the motion of the sources which are massive materials can be slow compared to the speed of light. Thus, contrary to the existence of the newtonian mechanics as the low-velocity approximation of the relativistic mechanics, it is impossible to derive a nonrelativistic approximation

for Maxwell's equations. Indeed, it is through the investigation and understanding of the transformation properties of the Maxwell equations that Einstein was led to the discovery of the special relativity theory.

Transformation. The covariance of the theory states that the electromagnetic field must transform by the Lorentz transformation from one reference frame to another. This transformation property is illuminating. For example, a pure electric field at rest will appear as an electric field plus a magnetic field when observed in a moving frame. This can be shown by writing the Lorentz transformation from one frame, K, to another, K', moving relative to K at velocity $c\beta$ along the z axis, as Eq. (10), where the coeffi-

$$x'_\mu = a_{\mu\nu}x_\nu \tag{10}$$

cients $a_{\mu\nu}$ are given by Eq. (11), and γ is defined by Eq. (12). Transforming $F_{\mu\nu}$ from K to K' by the

$$a_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \gamma & -i\gamma\beta \\ 0 & 0 & -i\gamma\beta & \gamma \end{pmatrix} \tag{11}$$

$$\gamma \equiv (1 - \beta^2)^{-1/2} \tag{12}$$

tensor transformation rule $F'_{\mu\nu} = a_{\mu\lambda}a_{\nu\rho}F_{\lambda\rho}$ yields Eqs. (13). This shows that the electric and the mag-

$$\begin{aligned} E'_x &= \gamma(E_x - \beta B_y) & B'_x &= \gamma(B_x + \beta E_y) \\ E'_y &= \gamma(E_y + \beta B_x) & B'_y &= \gamma(B_y - \beta E_x) \\ E'_z &= E_z & B'_z &= B_z \end{aligned} \tag{13}$$

netic fields are just different space-time projections of a more general object, electromagnetic field. *See* LORENTZ TRANSFORMATIONS; SPACE-TIME.

Energy and momentum. The electromagnetic field contains both momentum and energy, which is why electromagnetic waves can be used to transmit messages or power. The momentum and energy content of an electromagnetic field is given by the energy-momentum 4-tensor of Eq. (14), where the double

$$T_{\mu\nu} = \frac{1}{4\pi}\left[F_{\mu\nu}F_{\lambda\nu} + \frac{1}{4}\delta_{\mu\nu}F_{\lambda\sigma}F_{\lambda\sigma}\right]$$

$$= \left(\begin{array}{c|c} \vec{\vec{T}} & -ic\vec{g} \\ \hline -ic\vec{g} & u \end{array}\right) \tag{14}$$

overarrow signifies a 3-tensor, $\delta_{\mu\nu} = 1$ when $\mu = \nu$ and $\delta_{\mu\nu} = 0$ when $\mu \neq \nu$. This is expressed in separate space and time components by Eqs. (15);

$$\vec{\vec{T}} = \frac{1}{4\pi}\left[\vec{E}\vec{E} + \vec{B}\vec{B} - \frac{1}{2}\vec{\vec{I}}(E^2 + B^2)\right]$$
$$= \text{energy} - \text{momentum } 3 - \text{tensor}$$

$$\vec{g} \frac{1}{4\pi c}\vec{E} \times \vec{B} = \frac{1}{c^2}\vec{S} = \text{momentum density} \tag{15}$$

$$u = \frac{1}{8\pi}(E^2 + B^2) = \text{energy density}$$

here $\vec{S}$ is called the Poynting vector. Conservation of energy and momentum is expressed by Eq. (16),

$$\frac{\partial T_{\mu\nu}}{\partial x_\nu} = \frac{1}{c} F_{\mu\nu} J_\nu \qquad (16)$$

which can be derived from Eqs. (5) and (14). The expression on the right side is the 4-force density exerted by the charge-current density J_ν on the field. Hertz carried out an experiment in 1887 which succeeded in demonstrating the transmission of energy across space on a radio wave produced by an oscillatory spark discharge, consistent with this aspect of Maxwell's theory. *See* CONSERVATION OF ENERGY; CONSERVATION OF MOMENTUM; ENERGY; POYNTING'S VECTOR.

Solution of wave equation. To obtain the electromagnetic field, it is necessary to solve Eq. (1) for $\vec{E}$ and $\vec{B}$ or, equivalently, Eq. (8) for A_μ and the solution of Eq. (8) requires the Green's function of the equation. This is given by Eq. (17), where R is given by Eq. (18),

$$G(\vec{x}, t; \vec{x}', t') = \frac{\delta\left(t' + \dfrac{R}{c} - t\right)}{R} \qquad (17)$$

$$R = |\vec{x} - \vec{x}'|$$

= distance between source point $\vec{x}'$

and field point $\vec{x}$ (18)

and the Dirac δ-function specifies that at the source point $\vec{x}'$ the only values that matter are at the retarded time t' given by Eq. (19). Hence the general solution

$$t' = t - \frac{R}{c} \qquad (19)$$

for A_μ is given by Eq. (20), where the brackets on the

$$A_\mu(\vec{x}, t) = \frac{1}{c} \int d^3x' \left[\frac{J_\mu(\vec{x}', t')}{R}\right]_{\text{ret}} \qquad (20)$$

right emphasize that t' is given by Eq. (19). The retarded time t' assures causal behavior, such that the effect of the source propagates only with velocity c. This now gives the general and complete solution of all problems of finding the electromagnetic field produced by prescribed charge and current distributions. With a given charge-current distribution J_μ, Eq. (20) yields the 4-potential, A_μ which in turn gives the electromagnetic field $\vec{E}$ and $\vec{B}$ through Eqs. (6) and (3). Other interesting physical quantities such as the momentum and energy distributions are obtained by substituting $\vec{E}$ and $\vec{B}$ in Eq. (15). *See* GREEN'S FUNCTION; WAVE EQUATION.

Radiation from general localized source. The most interesting special case of the electromagnetic field is the radiation, or the faraway electromagnetic field, generated by a localized oscillating source. If only a single frequency component, ω, of the source is chosen [that is, if $J\mu$ has the form $J_\mu(\vec{x}, t) = J_\mu(\vec{x})e^{-iwt}$], and if it is assumed that the wavelength c/ω is much larger than the size of the source but much smaller than the distance to the field point, then the expanded form obtained for the 3-vector potential $\vec{A}$ is

given by Eq. (21), with $\vec{A}_m$ given by Eq. (22), where

$$\vec{A}(\vec{x}) = \sum_{m=0}^{\infty} \vec{A}_m(\vec{x}) \qquad (21)$$

$$\vec{A}_m(\vec{x}) = \frac{e^{ikr}}{cr} \frac{(-ik)^m}{m!} \int \vec{J}(\vec{x}')(\vec{n} \cdot \vec{x}')^m \, d^3x' \qquad (22)$$

$r \equiv |\vec{x}|$, $\vec{n} = \vec{x}/r$ and $k \equiv \omega/c$. The integral contains the electric $2(m+1)$-pole moment and the magnetic $2m$-pole moment of the source. Thus, A_0 gives the electric dipole radiation. This is the case of the simple dipole antenna. The term $\vec{A}_1$ gives the electric quadrupole and the magnetic dipole radiation. Its radiated power is lower than that of $\vec{A}_0$ by an order of (source size)^2/(wavelength)^2. *See* ANTENNA (ELECTROMAGNETISM); ELECTROMAGNETIC WAVE TRANSMISSION; MULTIPOLE RADIATION.

This part of relativistic electrodynamics clearly plays a central role in telecommunications, and in addition forms the basis for the development of radar, waveguides, radio telescopes, and so forth. *See* RADAR; RADIO; RADIO TELESCOPE; WAVEGUIDE.

Radiation from moving charge. In the case of a single charge e moving with given velocity $c\vec{\beta}$ along the given trajectory $\vec{x}_e(t)$, the 4-current density is given by Eq. (23), where $\beta_\nu = (\vec{\beta}, i)$. With this J_ν, Eq. (20)

$$J_\nu(\vec{x}, t) = ec\beta_\nu\delta[\vec{x} - \vec{x}_e(t)] \qquad (23)$$

gives, in separate space-time form, Eqs. (24), where κ is defined by Eq. (25). Here $R(t)$ denotes

$$\vec{A}(\vec{x}, t) = e\left[\frac{\vec{\beta}}{\kappa R}\right]_{\text{ret}}$$

$$\phi(\vec{x}, t) = e\left[\frac{1}{\kappa R}\right]_{\text{ret}} \qquad (24)$$

$$\kappa \equiv \frac{d}{dt}\left[t + \frac{R(t)}{c}\right] \qquad (25)$$

$|x - x_e(t)|$ and so now is a function of time. The significance of the square brackets on the right is that the functions of t are to be evaluated at the retarded time t' found from Eq. (26). These are called

$$t' = t - \frac{1}{c}R(t') \qquad (26)$$

Liénard-Wiechert potentials and give in the radiation zone (far away, where the fields depend on $1/R$) the fields of Eqs. (27), showing that a charged particle

$$\vec{E}(\vec{x}, t) = \frac{e}{c}\left[\frac{\vec{n}}{\kappa^3 R} \times \{(\vec{n} - \vec{\beta}) \times \dot{\vec{\beta}}\}\right]_{\text{ret}} \qquad (27)$$

$$\vec{B} = [\vec{n}]_{\text{ret}} \times \vec{E}$$

radiates when it is accelerating ($\dot{\vec{\beta}} \neq 0$). At this point $\vec{n}$ indicates a unit vector in the direction $\vec{x} - \vec{x}_e(t)$.

For a charged particle moving in a circle with radius ρ, the resulting radiative power is given by Eq. (28), and the energy loss per revolution is given

by Eq. (29). For high-energy electrons ($\beta \cong 1$) this has the numerical value given by Eq. (30).

$$P = \frac{2}{3}\frac{e^2 c}{\rho^2}\beta^4\gamma^4 \qquad (28)$$

$$\Delta E = \frac{2\pi\rho}{c\beta}P = \frac{4\pi}{3}\frac{e^2}{\rho}\beta^3\gamma^4 \qquad (29)$$

$$\Delta E \text{ (keV)} = 88.5\frac{[E \text{ (GeV)}]^4}{\rho \text{ (meter)}} \qquad (30)$$

In a circular accelerator, for example, a synchrotron, the electrons are deflected by dipole magnets to travel in a circular orbit. While traveling in this orbit, they emit the radiation given by Eq. (28). This synchrotron radiation was first observed and identified in 1948. At the time, it was considered a nuisance because it causes energy loss while one is trying to accelerate the electrons. Almost 20 years later it was realized that, because of its very high brightness, very broad spectrum (10^{-3} to 10^5 eV, or from infrared through visible and ultraviolet to soft and hard x-rays), very small source size, and special polarization, synchrotron radiation makes a very useful research tool for studying all types of atomic, molecular, solid, and surface structure problems. There are now about 20 electron storage rings with beam energies ranging from several hundred MeV to over 10 GeV operated for synchrotron radiation research. In addition, another 20 such facilities are being planned or in construction. *See* SYNCHROTRON RADIATION.

Charged particle motion. The problem of finding the motion of a charged particle in a specified electromagnetic field is more a mechanical problem, and hence has been largely formulated and solved in relativistic mechanics. All that is required is the "force." The relativistic equation of motion of a charged particle acted on by a given electromagnetic field is Eq. (31), where m is the rest mass of the particle;

$$\frac{dp_\mu}{d\tau} = \frac{e}{mc}F_{\mu\nu}p_\nu \qquad (31)$$

$d\tau = dt/\gamma$ is the proper time differential; p_μ, given by Eq. (32), is the 4-momentum; and the expression

$$p_\mu = m\gamma c\beta_\mu = \left(\vec{p}, \frac{iE}{c}\right) \qquad (32)$$

on the right side of Eq. (31) is the 4-force exerted on the particle by the field. Equation (32) can be derived from the relativistic lagrangian of Eq. (33).

$$L = \frac{1}{\gamma}\left(-mc^2 + \frac{e}{mc}p_\mu A_\mu\right) \qquad (33)$$

See LAGRANGE'S EQUATIONS.

In the separated space-time form, the equation of motion becomes Eqs. (34) and (35). The expression

$$\frac{d\vec{p}}{dt} = e(\vec{E} + \vec{\beta}\times\vec{B}) \qquad (34)$$

$$\frac{dE}{dt} = ec\vec{\beta}\cdot\vec{E} \qquad (35)$$

on the right side of Eq. (34), the momentum equation, is the force exerted on the particle by the field and is known as the Lorentz force. The right-side expression of Eq. (35), the energy equation, is the rate of energy imparted to the particle by the field. With a prescribed electromagnetic field, solving Eqs. (34) and (35) for the particle motion is straightforward and gives precise particle orbits at velocities arbitrarily close to the velocity of light. In high-energy particle acclerators, as used for research in particle physics, electron velocities within 10^{-10} of the velocity of light have been attained.

In addition to the application to particle accelerators, the solution of this second category of relativistic electrodynamic problems is indispensable to the design of magnetic confinement fusion machines such as the tokamak and the mirror machine, magnetohydrodynamic devices, and other plasma devices. The theory is also useful for describing the aurora borealis, the Van Allen belt, and other phenomena involving interplanetary and interstellar particle motions. *See* AURORA; MAGNETOSPHERE; NUCLEAR FUSION; PARTICLE ACCELERATOR; PLASMA (PHYSICS); SOLAR WIND; VAN ALLEN RADIATION.

Self-consistent charge-field theory. It is very tempting to try combining the formulations of the two categories described above to form a consistent picture of charged particles interacting among themselves through their own electromagnetic field. For a single electron this could give a total description of its electromagnetic behavior (the theory of the electron). But while the two parts of relativistic electrodynamics give quantitatively correct predictions to very high degrees of accuracy for macroscopic phenomena, the combined self-consistent theory when applied to microscopic phenomena runs into a variety of insurmountable mathematical and physical difficulties. For example, the electron orbiting the proton in a hydrogen atom does not radiate despite its acceleration. The electromagnetic energy of the electron computed from the theory is much greater than the energy corresponding to its mass. The coupled equations also give an unphysical self-accelerating solution according to which an electron could spontaneously accelerate from rest to infinite velocity. All these difficulties indicate that there is some fundamental defect in the theory making it invalid when applied to microscopic phenomena. This was resolved by the advent of quantum (relativistic) electrodynamics. The additional ingredient in quantum electrodynamics is the appreciation of the minimal effect of the action of making a measurement which becomes significant or even dominant for microscopic phenomena. *See* QUANTUM MECHANICS.   Lee C. Teng

Bibliography. D. F. Griffiths, *Introduction to Electrodynamics*, 3d ed., 1998; J. D. Jackson, *Classical Electrodynamics*; 3d ed., 1998; E. Konopinski, *Electromagnetic Fields and Relativistic Particles*, 1981; L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, 4th ed., 1976; H. C. Ohanian,

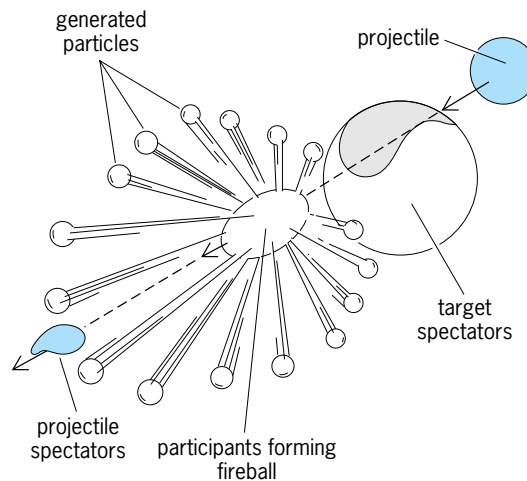*Classical Electrodynamics*, 1988; P. B. Visscher, *Fields and Electrodynamics*, 1988.

High-energy nuclear collision. The participant nucleons in the projectile and target transiently form a hot and dense fireball, while the spectator nucleons remain relatively undisturbed. The fireball will subsequently expand and emit a large number of particles, primarily pions and kaons.

Relativistic heavy-ion collisions

Collisions between heavy atomic nuclei at relative velocities close to the speed of light. These high-energy nuclear collisions are usually divided into two different domains, relativistic and ultrarelativistic collisions, depending on whether the kinetic energy per nucleon (the generic name for protons and neutrons) is either close to the rest mass of the nucleon (relativistic collisions) or much larger than the nucleon rest mass (ultrarelativistic collisions).

Nuclear and hadronic matter. By utilizing high-energy nuclear collisions, it is possible to study nuclear matter under conditions of very high temperatures and densities. The most common form of nuclear matter, at least under terrestrial conditions, is found in the atomic nucleus, which consists of protons and neutrons bound together by the strong nuclear force. If nuclear matter is heated up to temperatures comparable to the rest mass of the pion, it becomes a mixture of nucleons, pions, and various other particles, collectively denoted hadrons. Under these circumstances, nuclear matter is referred to as hadronic matter. *See* HADRON; NEUTRON; NEUTRON STAR; NUCLEAR STRUCTURE; PROTON.

Quark-gluon plasma. According to the quantum chromodynamics (QCD) theory, all hadrons are bound states of a set of more fundamental entities called quarks. The quarks are confined within the hadrons by the exchange of gluons. Quantum chromodynamics calculations using the most powerful computers available show that if hadronic matter is further heated or compressed to very high densities it will undergo a phase transition into a new phase of matter, called the quark-gluon plasma. In this phase the hadrons will lose their identity, and the quarks and gluons will be deconfined within volumes much larger than the typical hadron volume of 0.1–0.5 cubic femtometer. Quantum chromodynamics calculations indicate that the phase transition will occur at a critical density around 5–10 times the normal nuclear matter density of approximately 0.2 nucleon/fm^3, or at a critical temperature around 150 MeV. These conditions should have been fulfilled during the first few microseconds of the big bang, and might be attained in the core of high-mass neutron stars. Strong interest in ultrarelativistic heavy-ion collisions is caused by the realization that these collisions seem to be the only way to create a quark-gluon plasma under laboratory conditions. *See* BIG BANG THEORY; GLUONS; QUANTUM CHROMODYNAMICS; QUARK-GLUON PLASMA; QUARKS.

Space-time evolution. When two nuclei collide at high energies, some of the nucleons in each nucleus, called spectators, will continue their motion unaffected, while other nucleons, called participants, will strike one or several nucleons in the other nucleus (see **illus.**). In the overlap volume a hot and dense fireball will develop. If the temperature or density of the fireball becomes larger than the critical values, a quark-gluon plasma will be created with an estimated lifetime of 1–5 × 10^{-23} s. The fireball will start to expand and cool, and the quarks in the plasma will eventually be reconfined into a large number of hadrons (hadronization). After further expansion the hadrons will cease interacting with each other (freeze out) and leave the collision zone without further mutual interactions.

Global variables. The probability of creating the quark-gluon plasma is largest in central collisions, where the collision is head-on or close to it. It is therefore important to distinguish these central collisions from peripheral collisions. This characterization is best done by detecting global variables like the transverse energy or the charged-particle multiplicity. The transverse energy is a measure of how much energy has been redirected from the initial beam direction into the directions transverse to the beam. Central collisions will generate the largest values of transverse energy. Values as large as 50% of the total available center-of-mass energy have been observed in the most central collisions of 200 GeV/nucleon sulfur nuclei with lead nuclei. Such large values of transverse energy show that the nuclei are not transparent to each other but that the participating nucleons are being significantly slowed down. This high nuclear stopping power implies that a large fraction of the participants' initial kinetic energy is transformed into the production of many new particles at high temperatures. Based on transverse energy measurements, the highest fireball energy density attained is estimated to be approximately 10 times the density of normal nuclear matter.

The charged-particle multiplicity, which is the number of charged particles emerging from the nuclear collision, can also be used as a centrality indicator. The charged-particle multiplicity will, in addition, carry information on the amount of entropy generated during the collision. Experimentally, the

transverse energy and the charged-particle multiplicity have been shown to be very strongly correlated.

Theoretical models based on the assumption that collisions between two nuclei can be viewed as a superposition of a large number of binary collisions between the individual nucleons in the nuclei have proven to be very successful in estimating the observed values of the transverse energy and the charged-particle multiplicities. String models are in especially wide use. In these models the kinetic energy of two colliding nucleons is transformed into a string-shaped gluon field stretching between the quarks in the nucleons as they move past each other. The string will then subsequently be stretched and snap in several pieces, with a quark and an antiquark situated at opposite ends of each string fragment. Each string fragment will eventually evolve into a hadron.

Signatures of the quark-gluon plasma. In the search for the quark-gluon plasma, a fundamental problem is that even if the plasma is created in the early phases of the collisions, the subsequent hadronization and scattering of the hadrons before freeze-out might mask any traces of the plasma. In order to circumvent this problem, many plasma signatures have been proposed. The most promising signature is based on the conjecture that certain particles consisting of a charm quark and an anticharm quark (such as the J/ψ and ψ' particles) will be dissolved in the plasma and will not be regenerated in the later phases of the collision. This suppression in the production of the J/ψ particle has been observed. *See* J/PSI PARTICLE.

Another important property of the quark-gluon plasma in chemical equilibrium will be the high strange and antistrange quark content. Most of these strange quarks will be bound in strange particles (kaons, lambda, hyperons, and so forth) during the later stages of the collision. An unusually high content of these strange particles could therefore also be a signature for the creation of the quark-gluon plasma. Several experiments have actually detected unusually high yields of strange particles. *See* ELEMENTARY PARTICLE; NUCLEAR REACTION; STRANGE PARTICLES.   Soren P. Sorensen

Bibliography. G. F. Bertsch, Searching for the quark-gluon plasma, *Science*, 265:480–481, 1994; B. Schwarschild, Nuclear matter in extremis sought with ultrarelativistic heavy ion beams, *Phys. Today*, 41(3):17–21, March 1988.

Relativistic mechanics

An extension of newtonian mechanics conforming to the principles of special relativity. Suitably generalized, its results are also incorporated in the general theory. The energy-momentum conservation laws of relativistic mechanics enter in the development of relativistic quantum mechanics, and they find important application in high-energy physics.

Of Newton's laws of motion, it is the third law whose relativistic formulation presents particular challenge. Its application is simplest in the idealized context of strictly continguous action, whether in continuous media or in discrete particle systems—the only case discussed here. Without this assumption, one must cope with the problem of retarded action at a distance, on which some interesting exploratory work has been done. *See* NEWTON'S LAWS OF MOTION; RELATIVITY.

Equations of motion of a particle. The first correct relativistic equation of motion was that of a charged particle in an electromagnetic field developed by Max Planck in 1906. The restriction on the applied force was later removed by Hermann Minkowski, who was also the first to use the four-dimensional space-time formalism systematically. This formalism greatly facilitates a concise discussion of the subject, and will be employed here from the start. *See* RELATIVISTIC ELECTRODYNAMICS; SPACE-TIME.

The relativistic generalization of Newton's second law of motion for a particle, Eq. (1), must satisfy two

$$ma = f \tag{1}$$

conditions: it must include the newtonian equation as a limiting case for small particle speeds; it must be covariant (that is, retain its form) under Lorentz transformations, which is automatically assured when it is expressed in four-vector form. As to Newton's first law, its relativistic extension is immediate, because it serves only to define the family of inertial frames relative to which the second law is enunciated, and that family plays an identical role in the special theory of relativity. *See* LORENTZ TRANSFORMATIONS.

To arrive at the relativistic generalization of Eq. (1), it is first rewritten, using the particle momentum, $\mathbf{p} = m\, d\mathbf{x}/dt$, as $d\mathbf{p}/dt = \mathbf{f}$ (which is actually a closer representation of the law as stated in Newton's *Principia*). One then replaces dx, dy, dz by dx^α ($\alpha = 0, 1, 2, 3; x^0 \equiv ct$), m by the rest mass m_0 (that is, the mass of the particle as measured in the reference frame in which it is instantaneously at rest), and dt by its four-scalar extension $d\tau = dt\sqrt{1 - (v^2/c^2)}$. Retaining, as is customary, the symbol p for the generalization of $\mathbf{p}$, so that $p^\alpha = m_0 dx^\alpha/d\tau$ ($\alpha = 0,1,2,3$), the resulting four-vector equation assumes the form of Eq. (2), which may be called Minkowski's equation of motion.

$$\frac{dp^\alpha}{d\tau} = F^\alpha \qquad (\alpha = 0, 1, 2, 3) \tag{2}$$

The significance of Eq. (2) becomes apparent when the time t is reintroduced, resulting in Eqs. (3) and (4), where v is the instantaneous speed

$$\frac{dp^\alpha}{dt} = f^\alpha$$

$$p^\alpha = m\frac{dx^\alpha}{dt} \equiv mv^\alpha \qquad (\alpha = 0, 1, 2, 3) \tag{3}$$

$$m = \gamma m_0$$

$$f^\alpha = \frac{F^\alpha}{\gamma} \qquad (\alpha = 0, 1, 2, 3) \tag{4}$$

$$\gamma \equiv 1/\sqrt{1 - (v^2/c^2)}$$

of the particle. Since $\gamma = 1 + v^2/2c^2 \ldots$ it follows that

when v^2/c^2 can be neglected, Eqs. (3) for $\alpha = 1, 2, 3$ coincide indeed with the corresponding newtonian equations. Moreover, it is F^α and not f^α that are the components of a four-vector. This four-vector generalization of the newtonian force is known as the Minkowski force.

The meaning of Eqs. (3) for $\alpha = 0$ can be established with the aid of the four-scalar equation (5)

$$(p^0)^2 - \mathbf{p}^2 = m_0^2 c^2 \qquad (5)$$

derivable from Eqs. (4) and (3), where $\mathbf{p} \equiv (p^1, p^2, p^3)$ (not to be confused with the previous $\mathbf{p}$). Differentiating this equation with respect to t and using Eqs. (3), one finds: $d(mc^2)/dt = cf^0 = \mathbf{f} \cdot \mathbf{v}$. Hence the quantitydefined by Eq. (6) is an energy, and the

$$E = cp^0 = mc^2 \qquad (6)$$

four-vector p, the four-momentum, is thus also commonly called the energy-momentum vector. By Eqs. (5) and (6), it follows that E can also be expressed in terms of the "relativistic" $\mathbf{p}$ as in Eq. (7). In addition

$$E = \sqrt{m_0^2 c^4 + c^2 \mathbf{p}^2} \qquad (7)$$

tion to E, which is usually called the total energy, it is useful to introduce a relativistic kinetic energy (vanishing with v): $m_0 c^2 (\gamma - 1)$, where γ is defined in Eqs. (4). The expansion of γ shows that when v^2/c^2 can be neglected, this expression reduces indeed to the newtonian kinetic energy $mv^2/2$.

When $m_0 \neq 0$, Eqs. (4) and (6) show that for $v = 0$ the energy E has the nonvanishing value $E_0 = m_0 c^2$. This rest energy is thus an intrinsic energy of a particle associated entirely with its rest mass. The universality of this inertia-of-energy result, as established principally by Albert Einstein, and extensively confirmed by experiment, constitutes one of the most far-reaching conclusions of the theory of relativity.

Another important conclusion from Eqs. (4), which accords with other deductions from the relativistic principles, is that a particle of nonvanishing rest mass may approach but can never attain the speed c. On the other hand, the speed of a particle of vanishing rest mass, such as a photon, can only be c; otherwise, contrary to theory and experience, the particle's energy and momentum would be zero, as is seen, for instance, by taking the limit $m_0 \to 0$ in Eqs. (3). Equation (7) shows, moreover, that the energy and momentum of such a particle are related by the simple equation $E = c|\mathbf{p}|$; and also, that when $m_0 \neq 0$ this relation holds as an ultrahigh-energy approximation ($m_0 c \ll |\mathbf{p}|$).

Special systems of particles. Under the restrictive assumption that the particles of the system are involved only in spatially and temporally localized interactions, it follows easily that Newton's third law can be formulated in a relativistically covariant fashion. Moreover, if the system is closed, that is, subject to no external action, then as in the case of newtonian mechanics the total energy and momentum of the system are conserved quantities: $\Sigma E = $ constant,

$\Sigma \mathbf{p} = $ constant. Or, in four-vector notation Eq. (8) is satisfied.

$$P^\alpha \equiv \Sigma p^\alpha = \text{constant} \qquad (8)$$

$(\alpha = 0, 1, 2, 3; \text{ summation over the particles})$

It is possible to extend this result to allow for the (quantum) processes of the creation and annihilation of particles, so that the number of particles of the system need not be conserved, and the summation symbol in Eq. (8) refers to the particles existing at the instant under consideration. In particular, for a particle decay process or a localized binary collision of any type, Eq. (8) can be written more explicitly as Eq. (9), with $m = 1$ or 2, respectively. An early

$$\sum_{i=1}^{m} p_i^\alpha \text{ (initial)} = \sum_{j=1}^{n} p_j^\alpha \text{ (final)} \qquad (9)$$

and historically very important application of this formula was made by Arthur Compton in 1922 to the collision of an x-ray photon and a practically free electron. The extended formula has been applied to numerous high-energy particle reactions, and it appears to represent one of the most firmly established laws in physics. *See* COMPTON EFFECT.

As in the treatment of low-energy collisions, so also in the treatment of high-energy particle reactions it is helpful to deal not only with the laboratory frame but also with the center-of-mass (or center-of-momentum) frame of the mechanical system under consideration. The latter is defined by the condition that the spatial part of the total four-momentum of the system vanishes. In addition to the usual advantages of the center-of-mass frame, it is, in a practical sense, especially useful in the case of very high-energy collisions. With the aid of the Lorentz transformations connecting the components of the total four-momentum P defined by Eq. (8) in the laboratory and center-of-mass frames, it can be shown that if instead of bombarding some type of particles, which are effectively stationary in the laboratory, by a beam of similar particles, one directs two such beams against each other so that for the latter system the laboratory serves as a center-of-mass frame, the energy released in the collision is substantially greater in the second case, the advantage increasing with the rise in particle energy. This has been utilized in positron-electron and proton-proton colliding-beam devices called storage rings. *See* PARTICLE ACCELERATOR.

The center-of-mass frame has also independent interest. If it is denoted by S_0, and S is an arbitrary inertial frame, it can be shown from Lorentz transformations of the total four-momentum that $\mathbf{P} = \gamma P_{(0)}^0 \mathbf{v}/c = P^0 \mathbf{v}/c$, where $\mathbf{v}$ represents the velocity of S_0 relative to S, and P is defined in Eq. (8), P_0^α being the components of this four-vector in S_0. It follows that one can consider the quantity $M_0 \equiv P_{(0)}^0/c$ as the rest-mass of the system, and that $M \equiv P^0/c = \gamma M_0$, $\mathbf{P} = M\mathbf{v}$. The latter relations are identical with the appropriate relations in Eqs. (3) and (4) for a single particle of rest mass M_0 and velocity $\mathbf{v}$ relative to S. One can show, further, that $M_0 = \Sigma m_0 + \Sigma K_0/c^2$, the sums

being taken over the particles of the system, with m_0 and K_0 representing, respectively, the rest mass and the relativistic kinetic energy of a particle relative to S_0. The last term is an interesting manifestation of the inertia of energy.   H. M. Schwartz

Bibliography. A. Einstein, *The Meaning of Relativity*, 5th ed., 1956; R. A. Mould, *Basic Relativity*, 1994; R. Resnick and D. Halliday, *Basic Concepts in Relativity*, 1992; W. Rindler, *Introduction to Special Relativity*, 2d. ed, 1991; H. M. Schwartz, *Introduction to Special Relativity*, 1968, reprint 1977; E. Tocaci, *Relativistic Mechanics, Time and Inertia*, 1984.

Relativistic quantum theory

The quantum theory of particles which is consistent with the special theory of relativity, and thus can describe particles moving arbitrarily close to the speed of light. It is now realized that the only satisfactory relativistic quantum theory is quantum field theory; the attempt to relativize the Schrödinger equation for the wave function of a single particle fails, as described below. However, with a change of interpretation, relativistic wave equations do correctly describe some aspects of the motions of particles in an electromagnetic field. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM FIELD THEORY; QUANTUM MECHANICS; RELATIVITY.

Invariance. The Schrödinger equation for the wave function $\psi(\mathbf{r},t)$ of a particle [$|\psi(\mathbf{r},t)|^2$ is the density of probability of finding the particle at point $\mathbf{r}$ at time t] is Eq. (1), where E is the energy operator $i\hbar(\partial/\partial t)$, $\mathbf{p}$ is the momentum operator $-i\hbar\nabla$, $H(\mathbf{p},\mathbf{r})$ is the classical hamiltonian, and $\hbar$ is Planck's constant divided by 2π. For a nonrelativistic free particle, $H = \mathbf{p}^2/2m$. The naive way to relativize Eq. (1) would be to use the relativistic hamiltonian, Eq. (2).

$$E\psi = H(\mathbf{p},\mathbf{r})\psi \qquad (1)$$

$$H = \sqrt{(mc^2)^2 + \mathbf{p}^2 c^2} \qquad (2)$$

Although Eq. (1) with (2) gives the correct relativistic relation between frequency and wave number, and hence between energy and momentum, the equation itself is not relativistically invariant, essentially because E and $\mathbf{p}$ do not occur in it in a similar manner. A concrete difficulty is this: Suppose that at $t = 0$ the particle is localized at $\mathbf{r} = 0$, that is, $\psi(\mathbf{r},0) = 0$ for $\mathbf{r} \neq 0$. Then it can be shown that, according to Eq. (1) with (2), at any later time $\psi(\mathbf{r},t) \neq 0$ for all $\mathbf{r}$. But, according to relativity, the particle could not have traveled faster than light, speed c, so that $\psi(\mathbf{r},t)$ should be zero for $|\mathbf{r}| > ct$. *See* SYMMETRY LAWS (PHYSICS).

Klein-Gordon equation. An equation without the above defects is the so-called Klein-Gordon equation, Eq. (3). But $\varphi(\mathbf{r},t)$, which satisfies this equation, can-

$$E^2\varphi = [(mc^2)^2 + \mathbf{p}^2 c^2]\varphi \qquad (3)$$

not not be a wave function, for two related reasons: (i) Equation (3) is second-order in $\partial/\partial t$, and so not

only $\varphi(\mathbf{r},0)$ but also $\partial\varphi/\partial t$ is needed to determine the future values of φ; (ii) the only possible density of a conserved quantity formed from φ is of the form shown in relation (4). But this cannot be a proba-

$$\rho \propto \varphi^* E\varphi - \varphi E\varphi^* \qquad (4)$$

bility density, because it is not positive definite (it changes sign when φ is replaced by φ^*).

But ρ, in relation (4), can be interpreted as a charge density (when multiplied by a unit charge e); φ is then to be interpreted as a matrix element of a field operator Φ of a quantized field whose quanta are particles with mass m and charge e or $-e$ and zero spin. [Equation (3) is obeyed by $\Phi(\mathbf{r},t)$, and hence by any of its matrix elements.] The same is true in an electromagnetic field, where relations (3) and (4) are to be modified as shown in Eq. (5), where $\phi(\mathbf{r},t)$ and $\mathbf{A}(\mathbf{r},t)$

$$E \rightarrow E - e\phi \qquad \mathbf{p} \rightarrow \mathbf{p} - e\mathbf{A} \qquad (5)$$

are the scalar and vector potentials, respectively, of the electromagnetic field. Eigenstates of φ in a static field can be found in the usual way, which yields the energy levels of a charged spinless particle, say a π^- meson in the electric field of a nucleus, correctly except for the effects of radiative corrections (the electromagnetic interactions of the particle with itself and other virtually present particles) and of any internal structure of the particle. *See* EIGENFUNCTION; ENERGY LEVEL (QUANTUM MECHANICS).

Dirac equation. P. A. M. Dirac found a relativized form of Eq. (1), Eq. (6), which is both linear in E and

$$E\psi = [\beta mc^2 + \alpha \cdot \mathbf{p}c]\psi \qquad \rho \propto \psi^*\psi \qquad (6)$$

has a positive definite density form ρ, where β and α are constants which obey Eqs. (7). Applying the

$$\alpha_i \alpha_j + \alpha_j \alpha_i = 0 \qquad i \neq j$$
$$\alpha_i \beta + \beta \alpha_i = 0, \quad \alpha_i^2 = 1, \beta^2 = 1 \quad i, j = 1, 2, 3 \qquad (7)$$

equation a second time, Eq. (8) is obtained, that is,

$$E^2\psi = [\beta mc^2 + \alpha \cdot \mathbf{p}c]^2\psi$$
$$= [(mc^2)^2 + p^2 c^2]\psi \qquad (8)$$

the Klein-Gordon equation, which assures that the energy and momentum of the particle are correctly related. Obviously the four constants β and α_i cannot be numbers; however, they can be 4×4 matrices, and ψ is then a four-component object called a Dirac spinor. *See* MATRIX THEORY.

If plane wave solutions of Dirac's equation (6) are considered, then $\mathbf{p}$ is now a number. Taking Eq. (6) as an eigenequation for E, four eigenstates are found (because H is a 4×4 matrix), two with

$$E = +\sqrt{(mc^2)^2 + p^2 c^2}$$

and two with

$$E = -\sqrt{(mc^2)^2 + p^2 c^2}$$

the interpretation of the two positive energy states is that they are the two spin states of a particle with spin $\frac{1}{2}\hbar$; in fact an operator representing the spin angular momentum can be constructed out of α.

But the two negative energy states are an embarrassment; even a particle that was initially in a positive energy state would quickly make radiative transitions down through the negative energy states. Dirac's solution was to observe that if the particle described by ψ obeyed the Pauli principle (as the electron or any other spin-$^1/_2\hbar$ particle must), then it can be supposed that all the negative energy states are already filled with particles, thus excluding any more. There are still four so-called single-particle states for a given momentum $\mathbf{p}$: the two spin states of a particle with positive energy, and the two states obtained by removing a negative energy particle (of momentum $-\mathbf{p}$). These last states (hole states) have positive energy and a charge opposite the charge of the particle. The hole is in fact the antiparticle; if the particle is an electron, the hole is a positron. *See* ANTIMATTER; ELECTRON-POSITRON PAIR PRODUCTION; POSITRON.

With the filling up of the negative energy states, the system in question is no longer a single-particle system, and ψ, just as in the Klein-Gordon case, no longer can be interpreted as a wave function but must be interpreted as a matrix element of a field operator Ψ. Equation (6), with the modifications in Eq. (5), gives correct results for the eigenstates of a charged spin-$^1/_2\hbar$ particle in an electromagnetic field, except for the effects of radiative corrections and internal structure (compositeness); the latter is present only in hadrons such as the proton, but absent for the electron and muon. Particularly significant is the fact that the correct magnetic moment (Dirac moment) is given for the electron and muon (up to radiative corrections). *See* ELECTRON SPIN; ELEMENTARY PARTICLE; HADRON.

For a massless spin-$^1/_2\hbar$ particle, β does not appear in Dirac equation (6), and the α_i, satisfying Eqs. (7) with β omitted, can be 2×2 matrices; this is known as the Weyl equation. The particle, and likewise the antiparticle, then have only one spin state. But it can be shown that the particle has spin $^1/_2\hbar$; the particle (antiparticle) can only have its spin parallel (antiparallel) to its momentum. In an inverted coordinate system these statements would reverse, so that the two-component Weyl equation is not invariant to space inversion (parity). This "economical" version of the Dirac equation can be used to describe massless neutrinos. More fundamentally, the traces of chiral symmetry seen in weak interaction theory indicate that all spin-$^1/_2\hbar$ fundamental particles (leptons and quarks) satisfy the Weyl equation; the fact that the physical particles do not, except for massless neutrinos, is a consequence of the interactions of the particles. *See* LEPTON; NEUTRINO; QUARKS; SYMMETRY LAWS (PHYSICS).

Still another two-component variant of the Dirac equation is the Majorana equation, which describes a self-conjugate spin-$^1/_2\hbar$ particle. Self-conjugate means that the antiparticle is identical to the particle. The particle must be uncharged; it has only two intrinsic states, spin up and spin down. No particles are known to satisfy the Majorana equation, although massive neutrinos might.

Higher-spin particles. Wave equations for higher spin particles also exist. A well-known example is Maxwell equations for a (massless) spin-$1\hbar$ particle. The equations become more complicated as the spin becomes higher, and they will not be written here. When the spin is equal to or greater than $^1/_2\hbar$, the wave equation is not uniquely fixed by the particle's mass and spin; those higher electromagnetic moments which are allowed by the value of the spin (for spin $^1/_2\hbar$, magnetic dipole; for spin $1\hbar$, electric quadrupole also; and so forth) are arbitrary. However, if the particles are meant to be fundamental (point) particles, then the condition of renormalizability should be imposed, that is, the condition that radiative corrections be finite and calculable. In that case, the magnetic dipole moment of a spin-$^1/_2\hbar$ particle is fixed to be the Dirac value (before radiative corrections). A spin-$1\hbar$ particle must either be completely neutral (no electromagnetic interaction at all) or else be massless and be coupled as is a gauge particle. (Such particles can acquire mass through spontaneous symmetry breaking, as do the w and z bosons, for example, according to the standard model. Their magnetic dipole and electric quadrupole moments are predictable.) Finally, particles of spin $^3/_2\hbar$ or greater are forbidden to have any interactions at all. *See* MAXWELL'S EQUATIONS; QUANTUM ELECTRODYNAMICS; RELATIVISTIC MECHANICS; RENORMALIZATION; STANDARD MODEL; SYMMETRY BREAKING; WAVE EQUATION.   Charles J. Goebel

Relativity

A general theory of physics, primarily conceived by Albert Einstein, which involves a profound analysis of time and space, leading to a generalization of physical laws, with far-reaching implications in important branches of physics and in cosmology. Historically, the theory developed in two stages. Einstein's initial formulation in 1905 (now known as the special, or restricted, theory of relativity) does not treat gravitation; and one of the two principles on which it is based, the principle of relativity (the other being the principle of the constancy of the speed of light), stipulates the form invariance of physical laws only for inertial reference systems. Both restrictions were removed by Einstein in his general theory of relativity developed in 1915, which exploits a deep-seated equivalence between inertial and gravitational effects, and leads to a successful "relativistic" generalization of Isaac Newton's theory of gravitation.

Special Theory

The key feature of the theory of special relativity is the elimination of an absolute notion of simultaneity in favor of the notion that all observers always measure light to have the same velocity, in vacuum, c, independently of their own motion. The impetus for the development of the theory arose from the theory of electricity and magnetism developed by J. C. Maxwell. This theory accounted for all observed phenomena involving electric and magnetic fields and also predicted that disturbances in these fields would propagate as waves with a definite speed, c, in vacuum. These electromagnetic waves predicted

in Maxwell's theory successfully accounted for the existence of light and other forms of electromagnetic radiation. However, the presence of a definite speed, c, posed a difficulty, since if one inertial observer measures light to have velocity c, it would be expected that another inertial observer, moving toward the light ray with velocity v with respect to the first, would measure the light to have velocity $c + v$. Hence, it initially was taken for granted that there must be a preferred rest frame (often referred to as the ether) in which Maxwell's equations would be valid, and only in that frame would light be seen to travel with velocity c. However, this viewpoint was greatly shaken by the 1887 experiment of A. A. Michelson and E. W. Morley, which failed to detect any motion of the Earth through the ether. By radically altering some previously held beliefs concerning the structure of space and time, the theory of special relativity allows Maxwell's equations to hold, and light to propagate with velocity c, in all frames of reference, thereby making Maxwell's theory consistent with the null result of Michelson and Morley. *See* ELECTROMAGNETIC RADIATION; LIGHT; MAXWELL'S EQUATIONS.

Simultaneity in prerelativity physics. The most dramatic aspect of the theory of special relativity is its overthrowing of the notion that there is a well-defined, observer-independent meaning to the notion of simultaneity. To explain more precisely what is meant by absolute simultaneity, the following terminology will be introduced: An event is a point of space at an instant of time. Since it takes four numbers to specify an event—one for the time at which the event occurred and three for its spatial position—it follows that the set of all events constitutes a four-dimensional continuum, which is referred to as space-time.

A space-time diagram (**Fig. 1**) is a plot of events in space-time, with time, t, represented by the vertical axis and two spatial directions (x, y) represented by the horizontal axes. (The z direction is not shown.) For any event A shown in the diagram, there are many other events in this diagram—say, an event B—having the property that an observer or material body starting at event B can, in principle, be present at event A. (The world line—that is, path in
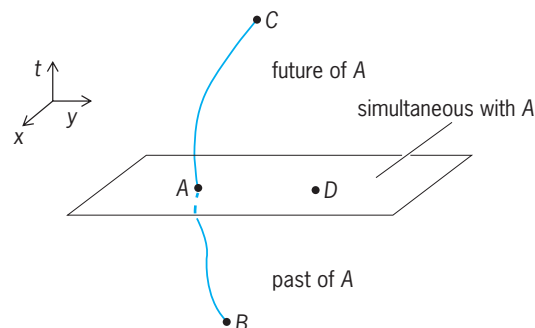
Fig. 1. Space-time diagram illustrating the causal relationships with respect to an event, *A*, in prerelativity physics. Event *B* lies to the past of *A*, event *C* lies to the future of *A*, and event *D* is simultaneous with *A*.

space-time—of an observer going from B to A can be illustrated in the diagram.) The collection of all such events constitutes the past of event A. Similarly, there are many events—say, an event C—having the property that an observer or material body starting at event A can, in principle, be present at event C. These events constitute the future of A. Finally, there remain some events in space-time which lie neither to past nor future of A. In prerelativity physics, these events are assumed to make a three-dimensional set, and are referred to as the events which are simultaneous with event A. This notion of simultaneity in prerelativity physics is described as being absolute, because it does not depend upon any constructions associated with a particular observer and all observers agree upon which events are simultaneous with event A.

In both prerelativity physics and special relativity, an inertial observer is one who is not acted upon by any external forces. In both theories it is assumed that any inertial observer, $\mathcal{O}$ can build a rigid cartesian grid of meter sticks, all of which intersect each other at right angles. Observer $\mathcal{O}$ may then label the points on this cartesian grid by the coordinates (x, y, z) representing the distance of the point from $\mathcal{O}$ along the three orthogonal directions of the grid. Observer $\mathcal{O}$ may carry the grid without rotating it, so that each point on the grid also undergoes inertial motion. A clock may then be placed at each grid point. In prerelativity physics, these clocks may be synchronized by requiring that they start simultaneously with each other (using the absolute notion of simultaneity described above). Any event in space-time may then be labeled by the four numbers t, x, y, z as follows: The numbers x, y, z assigned to the event are the spatial coordinates of the grid point at which the event occurred, whereas t is the time of the event as determined by the synchronized clock situated at that grid point. This labeling of events is referred to as the global inertial coordinate system associated with observer $\mathcal{O}$. *See* FRAME OF REFERENCE.

It is of interest to compare the coordinate labelings given to events in space-time by two inertial observers, $\mathcal{O}$ and $\mathcal{O}'$, who are in relative motion. The relationship occurring in prerelativity physics is called a galilean transformation. In the simple case where $\mathcal{O}'$ moves with velocity v in the x direction with respect to $\mathcal{O}$, and these observers meet at the event A labeled by $(t, x, y, z) = (t', x', y', z') = (0, 0, 0, 0)$, with the axes of the grid of meter sticks carried by $\mathcal{O}'$ aligned (that is, not rotated) with respect to those of $\mathcal{O}$, the transformation is given by Eqs. (1). The

$$t' = t \tag{1a}$$

$$x' = x - vt \tag{1b}$$

$$y' = y \tag{1c}$$

$$z' = z \tag{1d}$$

galilean transformation displays in an explicit manner that the two inertial observers, $\mathcal{O}$ and $\mathcal{O}'$, agree upon the time labeling of events and, in particular,
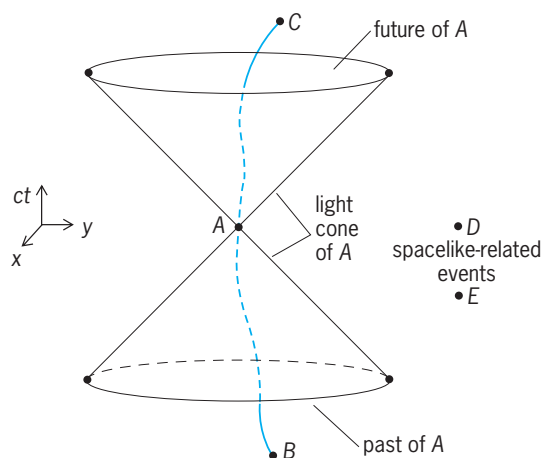
Fig. 2. Space-time diagram illustrating the causal relationships with respect to an event, *A*, in special relativity. Event *B* lies to the past of *A*, event *C* lies to the future of *A*, and events *D* and *E* are spacelike-related to *A*.

agree upon which events are simultaneous with a given event.

Causal structure in special relativity. The structure of space-time assumed in prerelativity physics as described above is both mathematically and logically self-consistent. Remarkably, however, nature does not appear to conform to this structure. Most dramatically, the causal structure of space-time in prerelativity physics (Fig. 1) is not correct. Rather, in special relativity there is a different causal relationship between an arbitrary event *A* and other events in space-time (**Fig. 2**). As in prerelativity physics, there are many events, *B*, which lie to the past of *A*; that is, an observer or material body starting at *B* can reach event *A*. There also are many events which lie to the future of *A*. However, there is now a much larger class of events which lie neither to the past nor to the future of *A*. These events are referred to as being spacelike-related to *A*.

The most striking feature of this causal structure (Fig. 2) is the absence of any three-dimensional surface of simultaneity. Indeed, the closest analog to the surface of simultaneity in prerelativity physics is the double-cone-shaped surface that marks the boundaries of the past and future of event *A*. This surface comprises the paths in space-time of all light rays which pass through event *A*, and for this reason it is referred to as the light cone of *A*. Thus, the statement that the events lying to the future of *A* are contained within the light cone of *A* is equivalent to the statement that a material body present at event *A* can never overtake a light ray emitted at event *A*. In special relativity, the light cone of an event *A* replaces the surface of simultaneity with event *A* as the absolute, observer-independent structure of space-time related to causality.

As in prerelativity physics, it is assumed in special relativity that an inertial observer, $\mathcal{O}$, can build a rigid grid of meter sticks, place clocks at the grid points, and label events in space-time by global inertial coordinates (t, x, y, z). The only difference from the procedure used for the construction of the sim-

ilar coordinates in prerelativity physics is that the synchronization of clocks is now nontrivial, since the causal structure of space-time no longer defines an absolute notion of simultaneity. Nevertheless, any pair of clocks in $\mathcal{O}$'s system can be synchronized—and thereby all of $\mathcal{O}$'s clocks can be synchronized—by having an assistant stationed half-way between the clocks send a signal to the two clocks in a symmetrical manner. This synchronization of clocks allows $\mathcal{O}$ to define a notion of simultaneity; that is, $\mathcal{O}$ may declare events A_1 and A_2 to be simultaneous if time readings, t_1 and t_2, of the synchronized clocks at events A_1 and A_2 satisfy $t_1 = t_2$. However, since there is no longer any absolute notion of simultaneity defined by the causal structure of space-time, there is no guarantee that a different inertial observer, $\mathcal{O}'$, will agree with $\mathcal{O}$ on the simultaneity of events. Indeed, events judged by $\mathcal{O}$ to be simultaneous ($t_1 = t_2$) will, in general, be judged by $\mathcal{O}'$ to be nonsimultaneous ($t_1' = t_2'$). *See* LORENTZ TRANSFORMATIONS.

The key assumptions of special relativity are encapsulated by the following two postulates.

Postulate 1: The laws of physics do not distinguish between inertial observers; in particular, no inertial observer can be said to be at rest in an absolute sense. Thus, if observer $\mathcal{O}$ writes down equations describing laws of physics obeyed by physically measurable quantities in her global inertial coordinate system (t, x, y, z), then the form of these equations must be identical when written down by observer $\mathcal{O}'$ in his global inertial coordinates (t', x', y', z').

Postulate 2: All inertial observers (independent of their relative motion) must always obtain the same value, c, when they measure the velocity of light in vacuum. In particular, the path of a light ray in space-time must be independent of the motion of the emitter of the light ray. Furthermore, no material body can have a velocity greater than c.

It is a direct consequence of postulate 2 that two inertial observers, $\mathcal{O}$ and $\mathcal{O}'$, who are moving with respect to each other must disagree over the notion of simultaneity. The precise relationship between the labeling of events in space-time by the global inertial coordinate systems of two inertial observers, $\mathcal{O}$ and $\mathcal{O}'$, in special relativity is given by the Lorentz transformation formulas. In the simple case where $\mathcal{O}'$ moves with velocity v in the x direction with respect to $\mathcal{O}$ and crosses $\mathcal{O}$'s world line at the event labeled by $(t, x, y, z) = (t', x', y', z') = (0, 0, 0, 0)$ with spatial axes aligned, the Lorentz transformation is given by Eqs. (2).

$$t' = \frac{t - \frac{xv}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{2a}$$

$$x' = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{2b}$$

$$y' = y \tag{2c}$$

$$z' = z \tag{2d}$$

Equation (2*a*) shows explicitly that $\mathcal{O}$ and $\mathcal{O}'$ disagree over simultaneity.

One immediate consequence of the disagreement between $\mathcal{O}$ and $\mathcal{O}'$ over simultaneity is the failure of the addition of velocities law of prerelativistic physics to hold. If $\mathcal{O}$ determines that $\mathcal{O}'$ moves (with respect to $\mathcal{O}$) with velocity v_1 in the x direction, and if $\mathcal{O}'$ similarly determines that another inertial observer $\mathcal{O}''$ moves (with respect to $\mathcal{O}'$) with velocity v_2 in the x' direction, then velocity v of $\mathcal{O}''$ with respect to $\mathcal{O}$ as determined by $\mathcal{O}$ can be computed by composing the Lorentz transformations between $\mathcal{O}$ and $\mathcal{O}'$ and between $\mathcal{O}'$ and $\mathcal{O}''$ to get the Lorentz transformation between $\mathcal{O}$ and $\mathcal{O}''$, from which the velocity v can be read out. Equation (3) is thereby obtained. When either v_1 or v_2 is much less than c,

$$v = \frac{v_1 + v_2}{\left(1 + \dfrac{v_1 v_2}{c^2}\right)} \tag{3}$$

this reduces to the law $v = v_1 + v_2$ that would be obtained from the galilean transformation, Eq. (1).

Space-time geometry. A key question both in pre-relativity physics and in special relativity is what quantities, describing the space-time relationships between events, are observer independent. Such quantities having observer-independent status may be viewed as describing the fundamental, intrinsic structure of space-time.

In prerelativity physics, the time interval $\Delta t \equiv t_2 - t_1$ between two events, labeled by an inertial observer $\mathcal{O}$ as (t_1, x_1, y_1, a_1) and (t_2, x_2, y_2, z_2), is observer independent. In other words, although a different inertial observer $\mathcal{O}'$ would attach different labels (t_1', x_1', y_1', z_1') and (t_2', x_2', y_2', z_2') to these two events, they always satisfy $\Delta t' = \Delta t$, where $\Delta t' \equiv t_2' - t_1'$. Similarly, in prerelativity physics any two inertial observers, $\mathcal{O}$ and $\mathcal{O}'$, would agree upon the spatial interval

$$\sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2}$$

between two simultaneous events ($\Delta t = 0$).

It has already been seen that in special relativity the time interval, Δt, between two events is no longer observer independent. Furthermore, since different inertial observers disagree over simultaneity, the spatial interval between two simultaneous events is not even a well-defined concept, and cannot be observer independent. Remarkably, however, in special relativity, all inertial observers agree upon the value of the space-time interval, I, between any two events, where I is defined by Eq. (4).

$$I = (\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 - c^2(\Delta t)^2 \tag{4}$$

What is most remarkable about this formula for I is that it is very closely analogous to the formula for squared distance in euclidean geometry. Indeed, for ordinary euclidean geometry on a four-dimensional space, the cartesian coordinates (a_1, a_2, a_3, a_4) may be introduced to label the points. Many different cartesian coordinate systems may be introduced by making rotations of the axes of the original sys-

tem, so the labels a_1, a_2, a_3, a_4 themselves have no intrinsic significance. However, the squared distance, D_2, between two points, given by the four-dimensional version of the pythagorean theorem, Eq. (5), is independent of the choice of cartesian

$$D^2 = (\Delta a_1)^2 + (\Delta a_2)^2 + (\Delta a_3)^2 + (\Delta a_4)^2 \tag{5}$$

coordinate system, in exactly the same way as I is independent of the choice of global inertial coordinate system.

The only mathematical difference between Eqs. (4) and (5) is The minus sign occurring in the last term in Eq. (4). This minus sign is of considerable importance, since it distinguishes between the notions of time and space in special relativity. Nevertheless, this minus sign turns out not to be a serious obstacle to the mathematical development of the theory of lorentzian geometry based upon the space-time interval, I, in a manner which parallels closely the development of euclidean geometry based upon D_2. In particular, notions such as geodesics (straightest possible lines) can be introduced in lorentzian geometry in complete analogy with euclidean geometry. The Lorentz transformation between the two inertial observers is seen from this perspective to be the mathematical analog of a rotation between two cartesian frames in euclidean geometry.

The formulation of special relativity as a theory of the lorentzian geometry of space-time is of great importance for the further development of the theory, since it makes possible the generalization which describes gravitation. The lorentzian geometry defined by Eq. (4) is a flat geometry, wherein initially parallel geodesics remain parallel forever. The theory of general relativity accounts for the effects of gravitation by allowing the lorentzian geometry of space-time to be curved. *See* SPACE-TIME.

Consequences. The theory of special relativity makes many important predictions, the most striking of which concern properties of time. One such effect, known as time dilation, is predicted directly by the Lorentz transformation formula (2*a*). If observer $\mathcal{O}$ carries a clock, then the event at which $\mathcal{O}$'s clock reads time τ would be labeled by her as (τ, 0, 0, 0). According to Eq. (2*a*), the observer $\mathcal{O}'$ would label the event as (t', x', 0, 0) where t' is given by Eq. (6). Thus, $\mathcal{O}'$ could say that a clock carried by $\mathcal{O}$

$$t' = \frac{\tau}{\sqrt{1 - \dfrac{v^2}{c^2}}} > \tau \tag{6}$$

slows down on account of $\mathcal{O}$'s motion relative to $\mathcal{O}'$. Similarly, $\mathcal{O}$ would find that a clock carried by $\mathcal{O}'$ slows down with respect to hers. This apparent disagreement between $\mathcal{O}$ and $\mathcal{O}'$ as to whose clock runs slower is resolved by noting that $\mathcal{O}$ and $\mathcal{O}'$ use different notions is simultaneity in comparing the readings of their clocks. *See* CLOCK PARADOX.

The decay of unstable elementary particles provides an important direct application of the time dilation effect. If a particle is observed to have a decay lifetime T when it is at rest, special relativity predicts

that its observed lifetime will increase according to Eq. (6) when it is moving. Exactly such an increase is routinely observed in experiments using particle accelerators, where particle velocities can be made to be extremely close to c. *See* PARTICLE ACCELERATOR.

An even more striking prediction of special relativity is the clock paradox: Two identical clocks which start together at an event A, undergo different motions, and then rejoin at event B will, in general, register different total elapsed time in going from A to B. This effect is the lorentzian geometry analog of the mundane fact in euclidean geometry that different paths between two points can have different total lengths.

Another effect predicted by special relativity which is closely related to the time dilation effect is the relativistic Doppler shift for light (or other electromagnetic) waves. It can be derived directly from the Lorentz transformation formulas (2a) and (2b). *See* DOPPLER EFFECT.

Many other ramifications of special relativity arise directly from the assumption that the laws of physics cannot prefer any global inertial coordinate system (postulate 1) and, thus, in particular, cannot make use of any notion of simultaneity. In prerelativity physics, the momentum three-vector of a particle, $\vec{p} = m\vec{v}$, appears in many laws. However, from a space-time perspective, $\vec{p} = (p^x, p^y, p^z)$ points in a spatial direction; that is, it is tangent to a surface of simultaneity. A law of physics involving $\vec{p}$ alone could not satisfy postulate 1. However, $\vec{p}$ could appear in laws of physics if it were the spatial components of a space-time four-vector, $p^\mu = (p^t, p^x, p^y, p^z)$. The only natural choice of such a space-time vector is given by Eq. (7), where m is the rest mass of the par-

$$p^\mu = mu^\mu \qquad (7)$$

ticle and u^u denotes the tangent vector to the curve describing the world line of the particle in space-time, with this curve parametrized by the time on a clock carried by the particle. The time component, p^t, of p^u in the global inertial coordinate system of observer $\mathcal{O}$ then has the natural interpretation of representing the energy, E, divided by c, of the particle as measured by $\mathcal{O}$; that is $pt = E/c$. Equation (7) then yields Eq. (8) for the energy and momentum of a par-

$$E = \frac{mc^2}{\sqrt{1 - \dfrac{v^2}{c^2}}} \qquad (8a)$$

$$\vec{p} = \frac{m\vec{v}}{\sqrt{1 - \dfrac{v^2}{c^2}}} \qquad (8b)$$

ticle in special relativity. *See* ENERGY; MOMENTUM; RELATIVISTIC MECHANICS.

The most striking feature of Eq. (8) is that an energy given by Eq. (9) must be assigned to a particleat

$$E = mc^2 \qquad (9)$$

rest. This contrasts sharply with the situation in prerelativity physics, where the energy is undefined up to the addition of an arbitrary constant and, by convention, this constant is normally chosen to make the energy of a particle at rest be zero. The existence of the relationship (9) implies that rest mass should be viewed as a form of energy, and thereby suggests the possibility that rest mass could be converted to other forms of energy. The fact that this possibility can be realized is perhaps the most significant ramification of the theory of special relativity. *See* REST MASS.   Robert H. Wald

General Theory

General relativity is the geometric theory of gravitation developed by Einstein in the decade before 1915. It is a generalization of special relativity, and includes the classical gravitational theory of Newton as its limit for weak gravitational fields and low velocities. The most important applications of the theory are to cosmology and to the structure of neutron stars and black holes. Due to astronomical observations since the 1990s, general relativity is now in the forefront of fundamental physics research. *See* GRAVITATION.

Need for a relativistic theory of gravity. One of the basic results of special relativity is that no physical effect can propagate with a velocity greater than that of light c; nature has a universal speed limit. However, classical gravitational theory describes the gravitational field of a body everywhere in space as a function of its instantaneous position, which is equivalent to the assumption that gravitational effects propagate with an infinite velocity. Thus, special relativity and classical gravitational theory are inconsistent, and a relativistic theory of gravity is necessary.

Principle of equivalence. Dating to Galileo, it is a fundamental question as to why bodies of different mass fall with the same acceleration in a gravitational field. This is now called the universality of free fall or the weak equivalence principle, and has been well tested experimentally. The best tests at present involve measuring the distance to the Moon with extreme accuracy by using laser beams bounced from reflectors on the lunar surface; these have been used to verify the principle to better than a part in 10^{12}, and this should be improved in the near future.

The universality of free fall follows from classical Newtonian gravitational theory since both the gravitational force on a body and its inertial resistance to acceleration are proportional to its mass. Thus the mass cancels out of the equations for the motion. However, this is not really an explanation but only an ad hoc description.

A deeper and more natural explanation occurred to Einstein. If an observer in the gravitational field of the Earth and another in an accelerating elevator or rocket in free space (**Fig. 3**) both drop a test body, they will both observe it to accelerate relative to the floor. According to classical theory, the Earth-based observer would attribute this to a gravitational force, while the elevator-based observer would attribute it to the accelerated floor overtaking the uniformly moving body. For both observers the motion is identical, and in particular the acceleration is independent
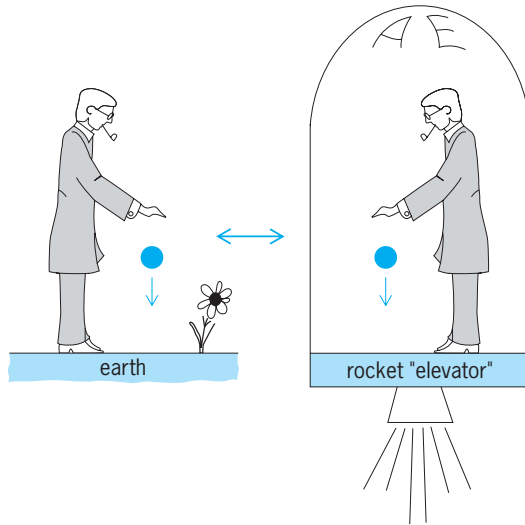
Fig. 3. Einstein's elevator, illustrating the principle of equivalence. The elevator acceleration is equal to *g*, the acceleration of gravity.

of the mass of the test body. Einstein elevated this fact to a general principle, called the Einstein principle of equivalence. This states that on a local scale all physical effects of a gravitational field are indistinguishable from the physical effects of an accelerated reference frame. This profound principle is a cornerstone of the theory of general relativity. From the point of view of the principle of equivalence, it is obvious why the motion of a test body in a gravitational field is independent of its mass. But the principle applies not only to gravitational effects but to all physical phenomena, and thereby has profound consequences (discussed below).

The principle of equivalence is true in general relativity, but not in modifications or in other gravity theories; for example, string theory suggests a violation of very roughly a part in 10^{17}. For the next decade a satellite test of the equivalence principle (STEP) is planned, which may achieve an accuracy of about 10^{18} and thereby provide a rare experimental test of string theory. *See* SUPERSTRING THEORY.

Tensors and Einstein's field equations. The observational fact that the motion of a body in a gravitational field is independent of its mass suggested to Einstein that gravity is fundamentally a geometric phenomenon, and that bodies should follow a special path determined only by geometry (see below). Because of this, gravity is naturally described by the mathematics of four-dimensional geometry. This geometry involves tensors, which are a simple generalization of vectors, and systems of tensor equations that are manifestly independent of the coordinate system. *See* TENSOR ANALYSIS.

Suppose that the points in four-dimensional space-time are labeled by two essentially arbitrary systems of coordinates x^μ and x'^μ (the indices μ and ν range from 0 to 3) and that the two coordinate systems are related to each other by a differentiable set of four functions, $x'^\mu = x'^\mu(x^\alpha)$ and $x^\nu = x^\nu(x'^\mu)$. Then contravariant tensors of rank 0, rank 1, rank 2, and

so forth are defined as sets of 1, 4, 16, and so forth numbers or functions whose values at the same point in the two systems of coordinates are related by the transformation Eqs. (10). Covariant tensors are similarly defined by Eqs. (11).

$$\text{rank 0:} \quad \phi' = \phi$$

$$\text{rank 1:} \quad \eta'^\mu = \frac{\partial x'^\mu}{\partial x^\alpha}\eta^\alpha \tag{10}$$

$$\text{rank 2:} \quad T'^{\mu\nu} = \frac{\partial x'^\mu}{\partial x^\alpha}\frac{\partial x'^\nu}{\partial x^\beta}T^{\alpha\beta}$$

$$\text{rank 1:} \quad \sigma'_\mu = \frac{\partial x^\alpha}{\partial x'^\mu}\sigma_\alpha$$

$$\text{rank 2:} \quad \omega'_{\mu\nu} = \frac{\partial x^\alpha}{\partial x'^\mu}\frac{\partial x^\beta}{\partial x'^\nu}\omega_{\alpha\beta} \tag{11}$$

In these equations the indices that appear twice in an expression are to be summed over; this is the Einstein summation convention. Tensors of rank 0 are also called scalars or invariants since they have the same value in all coordinate systems, and are thus of great importance in the description of physical quantities. Tensors of rank 1 are also called four-vectors.

The space-time of relativity contains one rank-2 tensor of particular importance, called the metric tensor $g_{\mu\nu}$, which is a generalization of the Lorentz metric of special relativity, introduced in Eq. (4). The metric tensor completely describes the geometry of space-time. Nearby points in space-time, also called events, which are separated by coordinate distances dx^μ, have an invariant physical separation whose square, called the line element, is defined by Eq. (12).

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu \tag{12}$$

This quantity is a generalization of the space-time interval in special relativity. When ds^2 is positive, it represents the square of the time interval between nearby events (multiplied by c^2) measured by an observer who moves at approximately constant velocity in such a way as to be present at both events. [A space with a quadratic metric form as in Eq. (12) is called a Riemann space.] *See* RIEMANNIAN GEOMETRY.

In a tensor equation one tensor of a given type is set equal to another of the same type, for example, $T^\mu = S^\mu$. In a different coordinate system, denoted by a prime, the definition of a tensor implies that both sides transform in the same way, so that the partial derivatives of the transformation cancel from the equation and it has the same form as in the original system, for example, $T'^\mu = S'^\mu$. Tensor equations are thus called form-invariant, or covariant.

The field equations of general relativity are tensor equations for the metric tensor. They will be presented very briefly. A Christoffel symbol (not a tensor) is first defined by Eq. (13); here $g^{\mu\tau}$ is the inverse of the matrix $g_{\mu\tau}$.

$$\left\{ \begin{matrix} \mu \\ \alpha\beta \end{matrix} \right\} = \frac{1}{2}g^{\mu\sigma}\left(\frac{\partial g_{\sigma\alpha}}{\partial x^\beta} + \frac{\partial g_{\sigma\beta}}{\partial x^\alpha} - \frac{\partial g_{\alpha\beta}}{\partial x^\sigma} \right) \tag{13}$$

The Riemann tensor (or curvature tensor) is then defined in terms of the Christoffel symbols by Eq. (14):

$$R^{\alpha}_{\mu\beta\nu} = \frac{\partial}{\partial x^{\nu}} \left\{ \begin{matrix} \alpha \\ \beta\mu \end{matrix} \right\} - \frac{\partial}{\partial x^{\mu}} \left\{ \begin{matrix} \alpha \\ \beta\nu \end{matrix} \right\}$$
$$+ \left\{ \begin{matrix} \alpha \\ \tau\nu \end{matrix} \right\} \left\{ \begin{matrix} \tau \\ \beta\mu \end{matrix} \right\} - \left\{ \begin{matrix} \alpha \\ \tau\beta \end{matrix} \right\} \left\{ \begin{matrix} \tau \\ \mu\nu \end{matrix} \right\} \quad (14)$$

This tensor plays a central role in the geometric structure of space-time; if it is zero, the space-time is termed flat and has no gravitational field; if nonzero, the space-time is termed curved and a gravitational field is present. In terms of the Ricci tensor, which is the Riemann tensor summed over $\alpha = \beta$, the Einstein field equations for empty space-time are given in Eq. (15):

$$R^{\alpha}_{\mu\alpha\nu} \equiv R_{\mu\nu} = 0 \qquad \text{(empty space)} \qquad (15)$$

The field equations are a set of 10 second-order partial differential equations, since $R^{\mu\nu}$ is a four-by-four symmetric tensor and has 10 independent components; the equations are to be solved for the metric tensor in a coordinate system that we are free to choose. Since the curvature of this space-time corresponds to the intrinsic presence of a gravitational field, the idea of a force field in classical gravitational theory is replaced in general relativity by the geometric concept of curved space-time. *See* DIFFERENTIAL EQUATION.

In a nonempty region of space-time the field equations (15) must be modified to include a tensor that represents the matter or energy which is the source of gravity, called the energy-momentum tensor $T_{\mu\nu}$. The modified equations are given in Eq. (16), where G is the gravitational constant, equal to 6.67×10^{-11} N $\cdot$ m$^2 \cdot$ kg^{-2}: [For empty space, $G_{\mu\nu} = 0$, which can be shown to imply $R_{\mu\nu} = 0$ also, so this is consistent with Eq. (15).]

$$G_{\mu\nu} \equiv R_{\mu\nu} - {}^1/_2 g_{\mu\nu} R^{\alpha}_{\alpha}$$
$$= -\frac{8\pi G}{c^4} T_{\mu\nu} \qquad \text{(nonempty space)} \qquad (16)$$

The tensor $G_{\mu\nu}$, called the Einstein tensor, represents the space-time geometry, and the tensor $T_{\mu\nu}$ represents the mass or energy that produces gravity. These equations automatically imply the conservation of energy and momentum, which is an extremely important result. Moreover, in the limiting case when the fields are weak and velocities small compared to c, the equations reduce to the equations of classical Newtonian gravity.

Cosmological term. The field equations were given in the form of Eq. (16) by Einstein in 1916. However, they can be consistently generalized by the addition of another term on the left side, which he called the cosmological term, $\Lambda g_{\mu\nu}$. The more general equations are given in Eq. (17).

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = -\frac{8\pi G}{c^4} T_{\mu\nu} \qquad (17)$$

The constant Λ is called the cosmological constant, and has the dimension of an inverse length squared. The extra term corresponds to a repulsive force between bodies that increases with distance.

Einstein introduced the cosmological term in 1917 in order to balance the usual attractive force of gravity and obtain static models of the universe, since it was then believed that the universe was static. When it was discovered by E. P. Hubble in 1929 that the universe is expanding, as evidenced by the Doppler shifts of distant galaxies, Einstein abandoned the cosmological term. However, beginning in 1998 measurements of the distance and recession velocity of very distant supernovae indicated that the cosmological term, or something that behaves just like it, is present and is causing the expansion of the universe to accelerate. The observed value of the constant is about (10^{10} light-years)$^{-2}$, so the cosmological term is only relevant on the cosmological scale. *See* COSMOLOGICAL CONSTANT; SUPERNOVA.

In the more modern view the cosmological term may be taken to the right side of the field equations and interpreted as the energy-momentum tensor of the vacuum; the energy density of the vacuum is $\rho_v = \Lambda c^4/8\pi G$. Just such a vacuum energy is predicted by quantum field theory, but its magnitude is, very roughly, an astounding 120 orders of magnitude greater than allowed by observation. This nonsensical result (with related questions) is called the cosmological constant problem (discussed below).

The energy of the vacuum, in whatever guise, is now generally referred to as dark energy (discussed below). Its magnitude, as inferred from observation, is about 3.5 GeV/m^3, and it comprises about 73% of the cosmic energy density; although it constitutes the dominant stuff of the universe, its nature is unknown, which is one of the biggest unresolved questions in physics today. *See* COSMOLOGY; DARK ENERGY.

Motion of test bodies. Test bodies are objects with very small mass and size and no internal structure. In general relativity the path of a test body in space-time is a generalized straight line called a geodesic; a geodesic may be thought of as a curve that is "everywhere parallel to itself," or as an "extremal curve," either the shortest or longest curve between two given points. For example, a straight line is a geodesic in Euclidean space, and a great circle is a geodesic on the surface of the Earth. In relativity a geodesic is the longest curve between two points in space-time, with the distance being defined as the integral of ds along the curve. In the special case of no gravitational field, with the metric equal to the Lorentz metric of special relativity, a geodesic is a simple straight line.

The scenario for gravitational influence in general relativity is that matter curves space in its vicinity, in accord with the Einstein equations, and in this curved space test bodies move on geodesics. The motion is thereby clearly independent of the mass of the test body, in accord with the principle of equivalence.

General relativity theory possesses an extraordinary property regarding geodesics: because the field

equations are nonlinear (unlike those of classical theory), the motion of a test body in a gravitational field cannot be specified independently of the field equations, since the body has mass and contributes a small amount to the field. Indeed the field equations are so restrictive that the equations of motion may be derived from them and need not be treated as a postulate.

Schwarzschild's solution. A very important solution of the field equations was obtained by K. Schwarzschild in 1916, surprisingly soon after the inception of general relativity. This solution represents the field in empty space around a spherically symmetric body such as the Sun. It is the basis for a relativistic description of the solar system and most of the experimental tests of general relativity to date. In spherical coordinates r, θ, ϕ and time coordinate $x^0 = ct$ the solution is represented by the line element given by Eq. (18), which is valid only outside the body:

$$ds^2 = \left(1 - \frac{2GM}{c^2 r}\right) c^2 dt^2 - \left(1 - \frac{2GM}{c^2 r}\right)^{-1} dr^2$$
$$- r^2(d\theta^2 + \sin^2\theta d\phi^2) \quad (18)$$

Here M is the mass of the body. In the limit of $M = 0$ this is the line element of special relativity written in spherical coordinates. This solution is the relativistic analog of the classical gravitational potential field of a spherical body, $\phi = -GM/r$. The quantity in each of the parentheses becomes zero at a radius $2GM/c^2$, the Schwarzschild radius, and infinite at the origin, which is a singular point.

Gravitational redshift. Electromagnetic radiation of a given frequency emitted in a gravitational field will appear to an outside observer to have a lower frequency; that is, it will be redshifted. The formula for the redshift can be derived from the principle of equivalence without the use of the Einstein field equations or the Schwarzschild solution, so redshift experiments test the basic idea of relativity but do not test the field equations. The most accurate test of the redshift to date was performed using a hydrogen maser on a rocket. Comparison of the maser frequency with Earth-based masers gave a measured redshift in agreement with theory to about 1 part in 10^4. *See* GRAVITATIONAL REDSHIFT.

Perihelion shift of Mercury. The geodesic equations of motion can be solved for the orbit of a planet, considered as a test body in the Schwarzschild field of the Sun. As should be expected, the planetary orbit obtained is very similar to the ellipse of classical theory. However, according to relativity the ellipse rotates very slowly in the plane of the orbit so that the perihelion, the point of closest approach of the planet to the Sun, is at a slightly different angular position for each orbit. This shift is extremely small for all the planets of the solar system, but it is greatest for Mercury, whose perihelion advance is predicted to be 43 seconds of arc in a century. This agrees with the discrepancy between classical theory and observation that was well known for many years before the discovery of general relativity. Results of

very precise planetary and asteroid perihelion position measurements using radar are consistent with general relativity theory. *See* CELESTIAL MECHANICS; RADAR ASTRONOMY.

Deflection of light. The principle of equivalence suggests an extraordinary phenomenon of gravity. Light crossing the Einstein elevator (Fig. 3) horizontally will appear to be deflected downward in a parabolic arc because of the upward acceleration of the elevator. The same phenomenon must therefore occur for light in the gravitational field of the Sun; that is, it must be deflected toward the Sun. A calculation of this deflection gives $1.75''$ for the net deflection of starlight grazing the edge of the Sun. The first measurements of this effect in 1919 verified the qualitative correctness of the deflection phenomenon. Modern measurements, made by tracking quasars with radio telescopes as they pass near or behind the Sun, find the deflection to be within 0.1% of the value predicted by general relativity.

In 1936 Einstein observed that if two stars were exactly lined up with the Earth, the more distant star would appear as a ring of light, distorted from its point appearance by the lens effect of the gravitational field of the nearer star. It was soon pointed out that a very similar phenomenon was much more likely to occur for entire galaxies instead of individual stars. Many examples of such gravitational lens systems have now been found. Some involve a quasar at a very large distance, nearly aligned with a normal galaxy nearer to us. Much effort is being given to gravitational lenses as tools to make absolute measurements of the distance to quasars, to measure the Hubble constant, and to seek and study the nonluminous dark matter. *See* DARK MATTER; GRAVITATIONAL LENS; QUASAR; RADIO ASTRONOMY.

Radar time delay. In the curved space around the Sun the distance between points in space, for example between two planets, is not the same as it would be in flat space. In particular, the round-trip travel time of a radar signal sent between the Earth and another planet will be increased by the curvature effect when the Earth, the Sun, and the planet are approximately lined up. Using radar ranging to planets and spacecraft, the time delay has been found to agree with the predictions of general relativity to an accuracy of better than 0.1%

Precession of a gyroscope. Relativity theory predicts that a spinning gyroscope in orbit around the Earth will precess by a small amount. Most of this relativistic precession is due to the curvature of space-time described by the Schwarzschild metric in Eq. (18); it is called the geodetic precession, and amounts to about 6.6 arcseconds per year for a low Earth orbit.

However, there is an additional smaller precession of great interest. For weak gravitational fields the Einstein field equations resemble Maxwell's equations of electromagnetism, and contain two types of fields: one resembles the electric field and is called the gravito-electric field, since it is due simply to the presence of matter; the other resembles the magnetic field and is called the gravito-magnetic

field, since it is due to matter in motion, that is, a mass current. The Schwarzschild metric describes the gravito-electric field of the Earth. The spin of the Earth produces a mass current which causes a gravito-magnetic field, and this gives rise to further precession of a gyroscope in orbit. Effects of the gravito-magnetic field are generally referred to as Lense-Thirring effects (following the work of J. Lense and H. Thirring in 1918) or as frame dragging effects. For an appropriately oriented low polar orbit and gyroscope spin direction the Lense-Thirring precession is about 0.042 arcseconds per year, and is orthogonal to the geodetic precession, so the two precessions may be separately measured. *See* MAXWELL'S EQUATIONS.

The existence of the gravito-magnetic field may be quite reliably inferred from the existence of the gravito-electric field and the Lorentz transformation, but before 2006 had never been directly verified in a clean experiment. The Gravity Probe B (GP-B) spacecraft was launched in April 2004 and has made a high-precision measurement of the geodetic precession and the first direct measurement of the Lense-Thirring precession—and thus the gravito-magnetic field of the Earth. Data taking was completed in 2005, and data analysis and release of results should be completed in 1 or 2 years. *See* GYROSCOPE.

Neutron stars. Pulsars are astronomical sources of electromagnetic radiation emitted in a series of precisely repeated pulses at a rate ranging from about 1 per second to over 100 per second. They are generally believed to be rotating neutron stars, the residual cores of heavy stars left behind after supernova explosions.

For ordinary stars, such as the Sun, the gravitational field is sufficiently weak that classical gravitational theory is a good approximation for studying the internal structure. This is not true for neutron stars, which contain a core composed largely of neutrons, and a few other elementary particles, at densities of about $2 \times 10^{14} \text{g/cm}^3$, greater than nuclear density; neutron stars pack more than a solar mass into a radius of only about 10 km (6 mi). The gravitational field in these stars can thus be quite large, and it is necessary to describe them using hydrostatic equations derived from the Einstein field equations in the presence of matter. These are called the Tolman-Oppenheimer-Volkov (TOV) equations.

One of the most remarkable theoretical predictions, found from the TOV equations, is an upper limit for the mass of a neutron star, above which it cannot be stable. This limit is similar to the well-known white dwarf mass limit of 1.4 solar masses, found by S. Chandrasekhar. One reason for neutron star instability above the limit is that, according to general relativity, pressure in the star is a source of gravity and contributes to the downward force of gravity. The exact value of the mass limit depends on the theoretical equation of state assumed, but is about a few solar masses. Many neutron stars have been observed; the masses cluster around 1.4 solar masses and none is much greater. For example, the

binary pulsar (discussed below) consists of a pulsar and a neutron star companion, both with about 1.4 solar masses. Many astronomical x-ray sources are powered by the strong gravitational fields of neutron stars, according to current theory, and this is presently a very active field of astronomy. *See* NEUTRON STAR; PULSAR; X-RAY ASTRONOMY.

Binary pulsar. The binary pulsar PSR 1913+16, discovered by R. Hulse and J. Taylor in 1974, is a pulsar in orbit about a neutron star companion. The orbital period is about 8 h, the orbital radius is about 10^6 km (6×10^5 mi), both stars have about 1.4 solar masses, and the pulse period is about 59 ms. Both stars move at about 300 km/s (186 mi/s), and the system is clean in that it has few nongravitational complications. Timing of its pulses to an accuracy of about a part in 10^{10} has been used to obtain much information about the system. Its orbital period is obtained from the Doppler shift of the pulses as the pulsar moves toward and away from us, which in turn gives the periastron advance, which is over $4°$ per year. Many of the general relativistic effects measured in the solar system have also been measured for the binary pulsar, and all are consistent with general relativity.

Perhaps the most interesting feature of the binary pulsar is that for over three decades of observation the period has decreased precisely as predicted by general relativity for energy loss to gravitational radiation (discussed below). The data agree with theory to better than 1%, which is convincing evidence for the existence of gravitational radiation.

Gravitational radiation. Gravitational radiation is analogous to electromagnetic radiation; whereas electromagnetic radiation is emitted by charges in accelerated motion, gravitational radiation is emitted by masses in accelerated motion. Such motion produces ripples or waves in the gravitational field that propagate at the velocity of light. The usual method of studying these ripples is with linearized general-relativity theory; in this theory the metric tensor is expressed as the constant Lorentz metric of special relativity plus a small perturbation term representing small gravitational ripples in an almost flat space. The forces exerted by plane gravitational waves are at right angles to the direction of propagation of the wave; that is, they are transverse (**Fig. 4**). As with electromagnetic waves, there are two polarizations possible for gravitational waves; the force field of one polarization is rotated $45°$ to the other. *See* ELECTROMAGNETIC RADIATION; PERTURBATION (MATHEMATICS); POLARIZATION OF WAVES.

Most relativists consider the binary pulsar observations to be convincing evidence for the existence of gravitational radiation, but direct detection is still of the utmost importance. There are many expected astrophysical sources of gravitational radiation. These include violent events such as supernova explosions, the gravitational collapse of stars, and the collisions of neutron stars and black holes. Waves from even such violent events are expected to produce only a fractional length distortion of 10^{-20} or less when they reach the Earth.
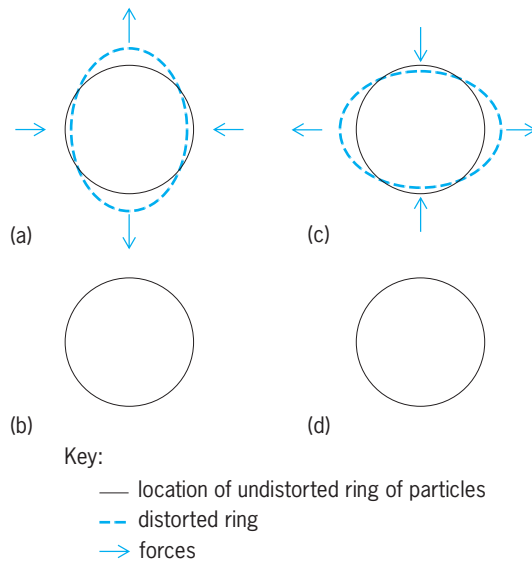
Key:
—— location of undistorted ring of particles
- - - distorted ring
→ forces

Fig. 4. The forces and distortions produced by a gravitational wave. (a) Begin cycle. (b) $^1/_4$ cycle (forces equal zero). (c) $^1/_2$ cycle (forces reverse). (d) $^3/_4$ cycle (forces equal zero).

Detection of astrophysical gravitational waves is difficult, but in principle the operation of real detectors is nearly as simple as shown in Fig. 4. The most sensitive modern detectors use laser interferometers; for example, the Laser Interferometer Gravitational-wave Observatory (LIGO) is a pair of L-shaped interferometers about 4 km (2.5 mi) in size, in which beams of laser light measure the difference in length between the two legs. The laser light bounces back and forth to give an effective length much greater than 4 km. For a typical expected wave the effective displacement is only about the size of a nucleus, but is measurable. LIGO has only been in operation since 2002, and no waves have yet been detected as of 2006, but as sensitivity is increased with various upgrades the first detection may be expected in a few years. *See* LIGO (LASER INTERFEROMETER GRAVITATIONAL-WAVE OBSERVATORY).
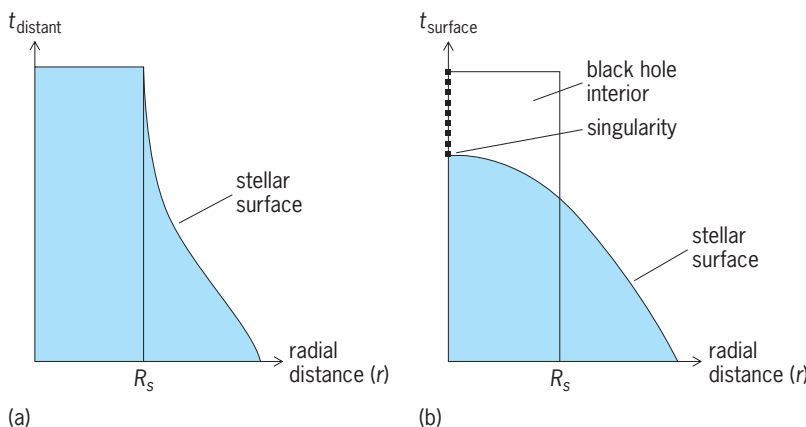

Fig. 5. Gravitational collapse. (a) From the point of view of a distant observer, the surface of a collapsing star approaches the Schwarzschild radius R_s or black hole surface. (b) As seen by an observer riding on its falling surface, it collapses through the black hole surface to reach zero size at a singularity.

Detection of gravitational radiation may open a new observational window that could be as important as the windows opened by neutrino astronomy and gamma-ray astronomy, which have changed our view of the universe. With gravitational radiation it might be possible to observe such things as the interiors of exploding supernovas and gamma-ray bursters, and see directly back to the early stages of the universe, none of which are accessible with light. *See* GAMMA-RAY BURSTS; GRAVITATIONAL RADIATION.

Black holes. If the core of an exploding supernova is heavier than the mass limit for a neutron star, as discussed above, there is no known stable end state. In 1937 J. R. Oppenheimer and H. Snyder predicted that such a core, when viewed from outside, will asymptotically collapse toward the Schwarzschild radius, $R_s = 2GM/c^2$. At that radius the redshift is infinite so no light can escape; effectively, time on the surface stops. This state is now known as a black hole, a name introduced by J. A. Wheeler. The prediction was verified in the 1970s when astrophysical x-ray sources were discovered. Many x-ray sources are clearly very compact objects with masses too large to be neutron stars, and black holes are the obvious candidates. One example is the source GRO J 1655-40 with 7.02 ± 0.22 solar masses. Moreover, most galaxies are believed to contain supermassive black holes of millions or billions of solar masses, and the huge energy output of quasars and radio galaxies are explained in terms of matter falling into black holes and emitting radiation.

Astronomical observations are limited to black hole exteriors, but theorists are also interested in their surfaces and interiors. To an outside observer the surface of a collapsing star (or stellar core) asymptotically approaches the Schwarzschild or black hole radius on a time scale of typically 10^{-5} s. Since the star is filled with burnt-out stellar material, the interior is not described by the Schwarzschild solution, which is only valid for the empty space outside the star. Thus there is no singularity inside the star.

However an observer sitting on the surface of the collapsing star measures a different time from that of an exterior observer (**Fig. 5**). For the surface observer the surface of the star passes through the Schwarzschild radius and continues on to the origin, where the density becomes infinite and a space-time singularity forms. At the Schwarzschild radius the surface observer's time differs by an infinite amount from that of a distant outside observer; the surface observer passes the Schwarzschild radius only after an infinite external time. In passing the surface of the black hole at the Schwarzschild radius, the surface observer enters the black hole interior, a region of space-time that is not accessible to any outside observer. Neither particles nor light can pass outward through the black hole surface. The surface thus acts as a one-way membrane or horizon, and outside observers cannot see inside.

If a collapsing star has angular momentum, the asymptotic black hole state is a generalization of the Schwarzschild solution to rotating systems,

discovered by R. Kerr. This solution is believed to be the general final asymptotic state for collapsing stars with angular momentum but no electric charge. A rotating black hole, described by the Kerr solution, has a gravito-magnetic field due to its angular momentum.

Gravitational singularities are a serious theoretical problem; they can produce paradoxical effects such as a flux of unlimited energy. But the singularity inside a black hole is inside a horizon and cannot affect the exterior world, which is thereby protected from such anomalies. R. Penrose and S. Hawking have conjectured a "cosmic censorship" principle, such that in relativity all singularities are similarly hidden by a horizon. Other theorists believe that singularities should not occur if quantum effects are taken into account, making cosmic censorship irrelevant.

Unlike an astrophysical black hole as discussed above, an "eternal" black hole is a theoretical object described by the Schwarzschild geometry everywhere in space-time. It is a region of empty space, containing a singularity, which has existed forever and was not formed by the collapse of a star. Part of the space-time is a black hole interior, part is a time-reversed black hole called a white hole, and part is a second exterior region. Eternal black holes have "wormholes" that connect the exterior regions, and which might allow time travel. Real black holes, formed by the collapse of matter in the past, do not have such regions, and there is no reason to expect that eternal black holes exist in nature.

Quantum effects change the above classical view of a black hole. In particular J. Bekenstein and Hawking (and many others) have shown that black holes have an intrinsic temperature, and radiate energy like a blackbody at that temperature. Thus they are not truly black and are not totally stable, but can evaporate away (as discussed below). *See* BLACK HOLE; GRAVITATIONAL COLLAPSE; HEAT RADIATION.

Cosmology. Observations of the last few decades have revolutionized the field of cosmology and put general relativity in the forefront of fundamental physics. But they have also raised many questions, and made it clear that we do not understand the dominant constituents of the universe.

The geometric viewpoint of relativity provides an elegant description of the universe on the cosmological scale of billions of light years. On this scale the visible material of the universe appears to be spatially isotropic and homogeneous; that is, it is substantially the same in all directions and at all points in space. (On a scale of less than 100 million light years there is much clumping of matter into galaxies, clusters of galaxies, and "great walls" and filaments of galactic clusters.) It is thus natural to assume that the large-scale geometry of the universe is isotropic and homogeneous. *See* GALAXY, EXTERNAL; UNIVERSE.

Consistent with this assumption there are essentially three kinds of geometry possible for the three-dimensional space of the universe (assuming a simple topology). These are illustrated by two-dimensional analogs, which are the surfaces of a plane, a sphere, and a saddle (**Fig. 6**). They are la-
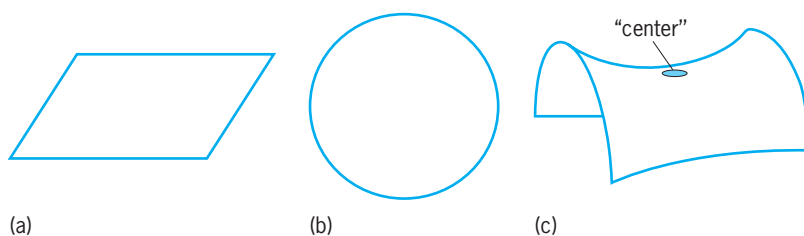
Fig. 6. Two-dimensional analogs of the three-dimensional space of the universe. All are homogeneous and isotropic. (*a*) Flat space (*k* = 0). (*b*) Positively curved space (*k* = 1). (*c*) Negatively curved space (*k* = −1). This space is like the circled center region everywhere.

beled with a curvature parameter k and called flat ($k = 0$), positively curved ($k = 1$), or negatively curved ($k = -1$). The negatively curved space is not really possible to visualize, but is everywhere analogous to the center of the two-dimensional saddle. Of these three geometries only the positively curved one is finite and closed, the others being open and having infinite extent.

Expansion of the universe may be visualized by picturing grid lines on the analog surfaces, with galaxies considered as particles of the cosmic fluid attached at the intersections. These move apart as the surfaces expand. In the case of the sphere the radius becomes larger with expansion. This picture emphasizes the role of coordinates in relativity as only markers of position: the galaxies remain at rest in the coordinate system but move apart physically as the universe expands.

The metric that most conveniently describes these geometries is called the FRW metric, after A. Friedmann (who first used it in 1922), H. P. Robertson and A. G. Walker; it is given in Eq. (19)

$$ds^2 = c^2 dt^2 - a(t)^2$$
$$\times [dr^2/(1 - kr^2) - r^2(d\theta^2 + \sin^2\theta\, d\phi^2)] \quad (19)$$

in terms of spherical coordinates. The scale function $a(t)$ has the dimension of length, r is a dimensionless radial coordinate, and t is a universal time.

A primary task of theoretical cosmology is to determine the scale function, but the curvature parameter k must be determined from observation. If the FRW metric in Eq. (19) is substituted into the Einstein field equations (17), two equations for the scale function result, often called the cosmological equations (20);

$$-\left(\frac{8\pi G}{c^4}\right)\rho = \Lambda - 3\left[\frac{k}{a^2} + \frac{\dot{a}^2}{c^2 a^2}\right],$$

$$\left(\frac{8\pi G}{c^4}\right)p = \Lambda - \left[\frac{k}{a^2} + \frac{\dot{a}^2}{c^2 a^2} + \frac{2\ddot{a}}{c^2 a}\right] \quad (20)$$

ρ denotes the density of the material of the universe, considered to be a fluid, p denotes its pressure, and the dot denotes a time derivative.

These equations must be supplemented by a relation between the density and pressure called the equation of state. The equation of state is often assumed to be a linear relation $p = w\rho$, with constant w; $w = 1/3$ corresponds to radiation (or hot matter in rapid thermal motion); $w = 0$ corresponds to cold

matter (with negligible pressure and slow thermal motion); and the rather curious case of $w = -1$ corresponds to vacuum or dark energy, which is equivalent to a cosmological constant.

Various assumptions about the nature of the cosmic fluid and the values of Λ and k may be made, and the cosmological equations solved accordingly. The favored combination for the present era is that the universe is spatially flat ($k = 0$), Λ is nonzero, and the fluid is cold matter with negligible pressure; this is the "concordance model" discussed below. Then the solution for the scale function is given in Eq. (21).

$$a(t) = A \, \sinh^{2/3}\left((\sqrt{3\Lambda}/2)ct\right)$$
$$A = \text{arbitrary constant} \quad (21)$$

For early times the scale function is very small, and for late times it approaches accelerated exponential expansion. However, this solution is not valid for very early times when the universe contained radiation and high-density material.

Observational cosmology relates the scale function $a(t)$ and the parameters k and Λ and ω to the real universe. Hubble's law for the observed recession velocity of galaxies is $v = HL$, where L is the distance to the galaxy and H is the "Hubble constant"; H is actually a parameter that varies slowly with time and is related to the scale function by $H = \dot{a}/a$. Its measured value at present is $H = 71 \pm 4$ km/(s Mpc), where 1 megaparsec (Mpc) = 3.3 million light years 3.1×10^{19} km, in the units favored by astronomers; H^{-1} is a rough measure of the lifetime of the universe, and is 14×10^9 years. It is clear from Hubble's law that the material of the universe must have been very dense and hot about 14×10^9 years ago. *See* HUBBLE CONSTANT.

The curvature parameter k is related to the total energy density of the universe, including the energy of the vacuum: if the density is greater than a critical value given by $\rho_c = 3H^2c^2/8\pi G \cong 8.3 \times 10^{-10}$ J/m^3, $\cong 5.2$ GeV/m^3, then the universe is positively curved ($k = 1$); if it is equal to critical, then the universe is spatially flat ($k = 0$); if it is less than critical, then the universe is negatively curved ($k = -1$). Diverse measurements of the density, deviations from linearity in the Hubble expansion from recent supernova data, and measurements of the cosmic background radiation or cosmic microwave background from the *Wilkinson Microwave Anisotropy Probe (WMAP)* have led astronomers to the present concordance model. This is spatially flat ($k = 0$), with about 73% of the cosmic energy being vacuum, about 23% cold dark matter of unknown nature, and only about 4% ordinary baryonic matter. The scale function is that in Eq. (21). *See* COSMIC BACKGROUND RADIATION; WILKINSON MICROWAVE ANISOTROPY PROBE.

The nature of the dark matter and the vacuum energy, now generally called dark energy, are both unknown. One leading candidate for the dark matter is the neutralino, a leptonic partner of the photon suggested by supersymmetry theory (SUSY), but never observed in the laboratory. It is one type of

WIMP (weakly interacting massive particle). There has as yet been no direct detection of dark matter particles. There are many speculative theories of the dark energy, which acts like vacuum energy or the cosmological constant. For example, it may be a slowly varying self-interacting scalar field that permeates the universe, analogous to the inflaton field of inflationary cosmology theory (discussed below). None of the theories of dark energy is generally accepted. *See* INFLATIONARY UNIVERSE COSMOLOGY; SUPERSYMMETRY; WEAKLY INTERACTING MASSIVE PARTICLE (WIMP).

In summary, cosmology is now in a somewhat paradoxical state. The general properties of the present universe are understood rather well, in terms of the concordance model of cosmology, but almost nothing is known about its main ingredients, the dark matter and dark energy.

Early universe. According to observation and general relativity, the universe began very small and very hot with a big bang, and expanded to become larger and cooler. While the present universe is dominated by dark energy and dark matter, the presence of the cosmic microwave background indicates that radiation and a hot plasma of electrons and nuclei dominated at earlier times, before about 100,000 years after the big bang, when the scale function was about 1,000 times smaller than at present. The cosmic microwave background is itself the cooled residual radiation from that time; it has been moving largely decoupled from interaction with matter since about 300,000 years after the big bang, when the electrons and nuclei in the previously existing plasma combined to form transparent neutral atoms. *See* BIG BANG THEORY.

The radiation era was in fact cosmically fairly simple, with a scale function proportional to $t^{1/2}$, and is fairly well understood theoretically. Rather early in the radiation era, at a minute or so, the nuclei of the light elements like helium and lithium formed, essentially by condensing out of a previously existing plasma of nucleons and electrons. The theory of this nucleosynthesis is quite successful at explaining the presently observed abundance of the light elements (70% hydrogen, 28% helium, 2% other nuclei), which gives us confidence in our understanding of the cosmology of the radiation era. *See* NUCLEOSYNTHESIS.

Most cosmologists believe that before the radiation era the universe went through a period of exponentially rapid expansion, or inflation, by a large factor of about e^{60} (10^{26}), during which time it was dominated by material that behaved much like a large cosmological constant, often called the inflaton field. Inflation is used to explain the present homogeneity of the universe and the isotropy of the cosmic microwave background. It also provides a possible explanation of the flatness (or near flatness) of the present universe ($k = 0$), the small fluctuations of about 10^{-5} in the temperature of the cosmic microwave background, and the clumpiness of the present universe in the form of galaxies and galactic clusters.

General relativity and quantum theory. In general relativity the space-time continuum is dynamical

and is the dominant entity. But in quantum theories such as quantum electrodynamics and quantum chromodynamics it serves as the unchanging arena in which objects interact. Moreover, in general relativity it is implicitly assumed that space-time distances can be measured to any desired degree of accuracy, whereas in quantum theory observables such as the position of an electron generally involve an intrinsic uncertainty. Thus it is not surprising that combining general relativity and quantum theory is difficult, and that there is as yet no generally accepted quantum theory of gravity and space-time. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM CHROMODYNAMICS; QUANTUM ELECTRODYNAMICS; QUANTUM FIELD THEORY; QUANTUM MECHANICS; RELATIVISTIC QUANTUM THEORY; STANDARD MODEL; UNCERTAINTY PRINCIPLE.

Indeed it is clear that the concept of a space-time continuum as used in general relativity must break down at a sufficiently small distance scale; this follows heuristically from general ideas of quantum theory. To measure a distance L an observer would use light of a wavelength λ smaller than L, and the measurement would have an intrinsic uncertainty of about λ (according to general principles of optics). But light has energy that creates a gravitational field by distorting space-time in the region being measured, which leads to a further uncertainty in the measurement. The minimum energy of the light is that of one photon $E = hc/\lambda$ (where h is Planck's constant), the effective photon mass is $M = E/c^2 = h/c\lambda$, and the fractional distortion of the space-time is very roughly $GM/c^2 L \approx Gh/c^3\lambda L$, which may be inferred from the Schwarzschild metric of Eq. (18). Thus the total uncertainty is about $\lambda + (Gh/c^3)/\lambda$. (At the desired level of approximation it is permissible to ignore factors like 2π and to make no distinction between h and $\hbar = h/2\pi$.) This expression has a minimum value $L_p = (Gh/c^3)^{1/2}$, which is called the Planck length. The Planck length is some 20 orders of magnitude smaller than distances presently measured in high-energy physics experiments. The classical concept of distance is widely believed to lose meaning at the Planck scale, and perhaps the intrinsic idea of distance itself does so as well. *See* PLANCK'S CONSTANT.

Max Planck recognized in 1899 that a basic scale of mass, length, and time is set by the constants G, c, and $\hbar$; these combine uniquely to give the Planck scale quantities in Eqs. (22). To the quantities in

$$L_P = \left(\frac{G\hbar}{c^3}\right)^{1/2} = 1.61 \times 10^{-35} \text{ m (Planck length)}$$

$$T_P = \left(\frac{G\hbar}{c^5}\right)^{1/2} = 5.37 \times 10^{-44} \text{ s (Planck time)}$$

$$M_P = \left(\frac{\hbar c}{G}\right)^{1/2} = 2.17 \times 10^{-8} \text{ kg (Planck mass)}$$
$$(22)$$

Eq. (22) a Planck energy may be added, defined by $E_p = M_p c^2 = (\hbar c^5/G)^{1/2} = 1.22 \times 10^{19}$ GeV in units appropriate to high-energy physics.

While gravitational theory needs to include quantum effects at the Planck scale the converse is also true, that quantum theories such as quantum electrodynamics and quantum chromodynamics need to include gravitational effects at the Planck scale. This is because the gravitational interaction becomes roughly comparable in strength to the electromagnetic and strong (quantum chromodynamics) forces at the Planck scale. The electric force between two particles with charges e is given classically by $F_{el} = e^2/r^2 = \hbar c\alpha/r^2$, where $\alpha = e^2/\hbar c \cong 1/137$ is the fine-structure constant. The gravitational force between two particles of energy E is given roughly by $F_g = G(E/C^2)^2/r^2$, so the two forces become equal at about $E \approx \sqrt{\hbar c^5/G}\sqrt{\alpha}$. To about an order of magnitude $\sqrt{\alpha} \approx 1$, so this energy is approximately the Planck energy. Thus, to describe physics at the Planck scale (the realm of small distances, short times, and high energy), a unified quantum treatment of gravity and space-time is required. Two such realms are the behavior of the hot dense material of the early universe, and the behavior of virtual particles of arbitrarily high energy in quantum field theory. Of course, laboratory experiments or astronomical observations at the Planck scale are difficult to even imagine. *See* GRAVITY.

Despite the lack of an accepted quantum theory of gravity, some interesting results have been obtained involving gravity and the quantum. The simplest way to combine ideas of gravity and quanta is to assume that quantum field theory is approximately valid in a classical curved background space-time obtained from the field equations of general relativity. The best-known example of this involves the thermodynamics of black holes. Bekenstein showed that for consistency in the interaction of radiation with black holes a black hole should have an entropy proportional to its area. This appeared to be inconsistent with the black hole property of absorbing radiation but not emitting it; but Hawking applied quantum field theory in the curved space-time of a black hole and showed that a black hole should indeed emit radiation (from very near its surface) as if it were a black body with a temperature given (in units of energy) by $T_{BH} = \hbar c/8\pi GM$. Hawking radiation makes black hole thermodynamics consistent, and is widely accepted by theorists, although it has as yet not been verified by observation.

However, it is not consistent to have a classical gravitational field coexisting with other quantized force fields. The uncertainty principle could then be subverted, in principle, by gravitational measurements. One approach to a true quantum theory of gravity is to ignore the geometric interpretation of gravity and develop the theory in analogy with quantum electrodynamics and other successful quantum field theories. To do this the metric tensor is written as the sum of the Lorentz metric of flat space-time plus a quantity $h_{\mu\nu}$ that describes the gravitational field in a flat background space-time. Then $h_{\mu\nu}$ may be quantized in the standard way, and Feynman diagrams and amplitude rules obtained for a perturbative quantum gravity theory. The perturbative

theory is successful until diagrams with internal loops are encountered. Divergent integrals associated with these diagrams cannot be reinterpreted and removed as they are in quantum electrodynamics; that is, the theory is not renormalizable, or involves an infinite number of renormalization constants, depending on the viewpoint of the theorist. *See* RENORMALIZATION.

The cosmological constant problem, mentioned previously, has perplexed theorists for decades. It is the result of combining gravity and quantum ideas, although its resolution may not require either a theory of quantum gravity or the Planck scale. In quantum field theory the vacuum is the lowest energy state of the field, which is not in general zero. Indeed it is generally given by a divergent integral, which cannot be removed by renormalization like other divergent integrals; the energy of the vacuum does have observable gravitational effects, equivalent to a cosmological constant as already discussed. If the upper limit of the divergent integral is taken to be the Planck energy instead of infinity, which is quite reasonable, the energy of the vacuum turns out to be approximately the Planck density, E_p/L_p^3, which is to be expected on dimensional grounds. But this has the nonsensical value 10^{124} GeV/m^3, which is about 124 orders of magnitude larger than allowed by observation. There has been no convincing resolution of this paradox, which is considered by many to be the most troublesome problem in theoretical physics.

Many interesting attempts have been made to develop a complete quantum gravity theory. The best known is superstring theory, which purports to describe all particles and all forces, including gravity, in terms of one-dimensional segments of string in a space of 10 or more dimensions. Another popular approach is called loop quantum gravity, and recasts the variables and mathematical structure of general relativity in a way that is more amenable to quantization. Neither of these theories has produced definitive predictions that may be tested by experiment. More drastic theoretical ideas involve constructing space-time in a way that is consistent with quantum theory, from so-called spin networks or from a quantum fluid ether. Probably the most drastic idea is that space-time as a fundamental entity simply does not exist, and only emerges as an approximation at scales well above the Planck scale. *See* QUANTUM GRAVITATION.   Ronald J. Adler

Bibliography. R. J. Adler, M. Bazin, and M. M. Schiffer, *Introduction to General Relativity*, McGraw-Hill, 1975; R. J. Adler, Gravity, Chap. 3 in G. Fraser (ed.), *The New Physics for the Twenty-first Century*, Cambridge, 2005 (in press); H. Bondi, *Relativity and Common Sense*, 1964, reprint, Dover, 1980; A. Einstein, *Relativity: The Special and the General Theory: The Masterpiece Edition*, Pi Press, 2005; N. D. Mermin, *Space and Time in Special Relativity*, 1968, reprint, Waveland Press, 1989; C. Misner, K. Thorne, and J. Wheeler, *Gravitation*, W. H. Freeman, 1973; B. Schutz, *A First Course in General Relativity*, Cambridge, 1985; C. M. Will, *Theory and Experiment in Gravitational Physics*, Cambridge, 1993; P. J. E. Peebles, *Principles of Physical Cosmology*, Princeton, 1993; E. W. Kolb and M. S. Turner, *The Early Universe*, Westview, 1993.

Relaxation time of electrons

The characteristic time for a distribution of electrons in a solid to approach or "relax" to equilibrium after a disturbance is removed. A familiar example is the property of electrical conductivity, in which an applied electric field generates an electron current which relaxes to an equilibrium zero current after the field is turned off. The conductivity of a material is directly proportional to this relaxation time; highly conductive materials have relatively long relaxation times. The closely related concept of a lifetime is the mean time that an electron will reside in a given quantum state before changing state as a result of collision with another particle or intrinsic excitation. This lifetime is related to equilibrium properties of the material, whereas the relaxation time relates to the thermal and electrical transport properties. The average distance that an electron travels before a collision is called the mean free path. Although typical collision times in metals are quite short (on the order of 10^{-14} s at room temperature), mean free paths range from about 100 atomic distances at room temperature to 10^6 atomic distances in pure metals near absolute zero temperature. Considering the very dense packing of atoms in a solid, these surprisingly long electron path lengths are analogous to the unlikely event that a rifle bullet might travel for miles through a dense forest without hitting a tree. The detailed explanation of the electron mean free path in metals is a major success of the modern theory of solids.

Transport property. A relaxation time appears in the simplest expression for the transport property of electrical conductivity, which states that the electrical conductivity equals the product of the relaxation time, the density of conduction electrons, and the square of the electron charge, divided by the electron effective mass in the solid. *See* BAND THEORY OF SOLIDS; ELECTRICAL CONDUCTIVITY OF METALS.

The conduction process is a steady-state balance between the accelerating force of an electric field and the decelerating friction of electron collisions which occur on the time scale of the relaxation time. This process may be described in terms of the probability distribution function for the electrons, which depends on the electron momentum (proportional to the wave vector, **k**, which labels the quantum state of the electrons), the position, and the time. Viewed in **k**-space, the entire distribution will shift from an equilibrium state under the influence of a perturbation such as an electric field (**Fig. 1**). For example, in the ground state, the collection of occupied electron states in **k**-space is bounded by the Fermi surface centered at the origin, while in an electric field this region is shifted. Because of electron collisions with impurities, lattice imperfections, and vibrations (also called phonons), the displaced surface may
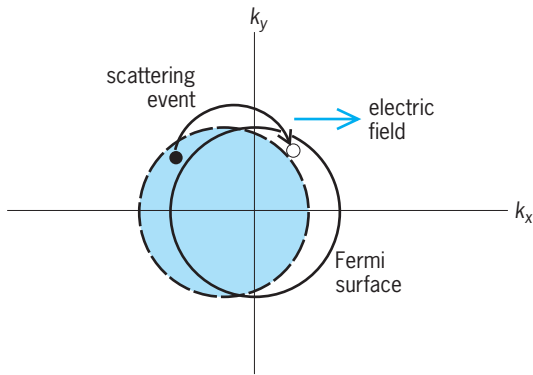
Fig. 1. Effect of the electric field on electron distribution in a solid, viewed in k-space, where k is the electron wave vector. The shaded area indicates occupied states in distribution which result when the field is applied.

sulting in screening of the cores and coherent transmission of the waves over large distances. Any disturbance to the lattice periodicity tends to destroy this wave phenomenon, resulting in lower transmission or energy loss. At room temperature, the conductivity is usually limited by scattering from lattice vibrations (phonons). At lower temperatures, vibrations are greatly reduced, but the conductivity is still limited by scattering from impurities and imperfections. *See* ELECTRICAL RESISTIVITY; FREE-ELECTRON THEORY OF METALS; QUANTUM MECHANICS.

be maintained in a steady state in an electric field. These collisions also restore the equilibrium distribution after the field is turned off, and the relaxation time is determined by the rate at which the shifted distribution returns to equilibrium. Specifically, the contribution of collisions to the rate of change of the shifted distribution after the field is turned off equals the difference between the shifted and equilibrium distributions divided by the relaxation time. This statement is referred to as the relaxation-time approximation, and is a simple way of expressing the role of collisions in the maintenance of thermodynamic equilibrium. The details of the various collision mechanisms are lumped into the parameter of the relaxation time. For example, in the case of mixed scattering by impurities and phonons, the inverse of the relaxation time can be determined from the sum of collision rates as the sum of the inverses of the electron-impurity and electron-phonon scattering times. *See* CRYSTAL DEFECTS; FERMI SURFACE; LATTICE VIBRATIONS; PHONON.

Constructive interference. In pure metals at low temperatures, the long mean free path of conduction electrons results from their large velocity (on the order of 10^6 m/s near the Fermi surface) and relatively long relaxation time, on the order of 10^{-9} s. From a practical standpoint, this is what makes metals useful as electrical conductors even at room temperature, where a relaxation time on the order of 10^{-14} s and a mean free path (equal to the product of the velocity and the relaxation time) on the order of 10^{-8} m (or about 100 atomic distances) is typical.

As compared to poorly conducting solids or insulators, an excess of so-called free electrons in a metal contributes to the long mean free paths. From a quantum mechanical standpoint, the free-electron wave function readjusts in a perfectly periodic atomic lattice to avoid the atomic ion cores and spend most of the time in the spaces between. In the analogy of the rifle fired into a forest, the bullet will not travel far, but the sound of the gunshot can, because the sound waves bend around the trees in a constructive manner. In a perfectly periodic lattice, the electron waves scatter constructively from the atomic ion cores, re-

Fig. 2. Scattering of conduction electrons from lattice vibrations in copper. (*a*) Fermi surface and Brillouin zone of copper. (*b*) Dependence of the scattering coefficient on the direction of the conduction electron wave vector, according to theoretical calculations and measurements using the radio-frequency size effect (RFSE) and the cyclotron resonance of surface Landau levels (SLL). (*After J. F. Koch and R. E. Dozema, The anisotropy of thermal scattering on the Fermi surface of metals, in Proceedings of the 14th International Conference on Low Temperature Physics, pp. 314–338, 1975*)

In the superconductive state, observed in many metals and certain complex compounds at sufficiently low temperatures, impurities, imperfections, and lattice vibrations become completely ineffective in retarding current flow, leading to persistent electric currents even after the driving electric field is removed. This state of resistanceless conduction is described as an ordered quantum state of pairs of electrons resulting from lattice vibrations which deform the ion core potential in such a way as to provide an attractive electron-electron interaction. The relaxation time thus becomes infinite (as nearly as can be measured), a rare macroscopic manifestation of a quantum effect. *See* SUPERCONDUCTIVITY.

Local measurements. A direct measurement of electron lifetimes for specific electron states in **k**-space is possible. The techniques work best at low temperatures and generally involve some kind of resonance in space (for example, radio-frequency size effect), time (cyclotron resonance), or energy (de Haas–van Alphen effect). The richness of detail made possible by such measurements has greatly aided the understanding of the interactions between conduction electrons, impurities, lattice imperfections, and vibrations. *See* CYCLOTRON RESONANCE EXPERIMENTS; DE HAAS-VAN ALPHEN EFFECT.

For example, measurements of conduction electron scattering from lattice vibrations reveal that the strongest interaction occurs when the electron's wave vector (proportional to the inverse wavelength) is nearly commensurate with the lattice periodicity (**Fig. 2**). For example, the wave vectors of the conduction electrons in copper lie near the Fermi surface (Fig. 2a), and the scattering is strongest for wave vectors on the Fermi surface in the vicinity of the [111] axis (Fig. 2b). In this direction, the electron wavelength coincides with the lattice periodicity at points where the Fermi surface touches the Brillouin zone (Fig. 2a). *See* BRILLOUIN ZONE; CRYSTAL STRUCTURE; CRYSTALLOGRAPHY.

Measurements of impurity scattering have shown that the symmetry character of both the impurity and the host are important in scattering. The scattering is strongest along directions where the electronic band structure symmetry is most nearly similar and where a "resonant" scattering is most nearly achieved. Thus studies of impurity scattering rates allow the symmetry of the scattering to be deduced. For example, the scattering of copper conduction electrons from nickel impurity atoms is largest near the "belly" region of the Fermi surface on the [110] axis (Fig. 2a). In this direction, the symmetry of the conduction electron wave function is very $d-$like, establishing that the scattering is likely due to resonance with the nickel $d-$band.

Measurements of conduction electron scattering from lattice dislocations (a disturbance of lattice periodicity which permits metal deformability) show that the measured collision lifetime differs by more than three orders of magnitude, depending upon the extent to which a particular type of measurement is sensitive to the small-angle scattering. For a dislocation "forest" in which 1 out of every 10,000 atoms in a row is out of place, the scattering rate is about 10^{11} s^{-1} in the de Haas–van Alphen effect, but only 10^8 s^{-1} in resistivity, and intermediate in the radio-frequency size effect. These differences arise because small-angle scattering is a small perturbation to current flow, whereas for quantum-effect measurements even a tiny scattering event can destroy constructive interference of the incident and scattered electron waves.

Energy relaxation. At higher temperatures and in disordered or polycrystalline metals, such refined **k**-space measurements are difficult, if not impossible, to perform. Even transport measurements of relaxation times are complicated by mixed scattering from defects, phonons, and so forth. However, direct time-resolved measurements of inelastic electron scattering are possible by means of ultrashort-duration (10^{-14} s to 10^{-12} s) laser pulses. In simple metals, a rapid increase in the electron temperature is produced by absorption of a laser pulse whose duration is shorter than the time required for hot electrons to equilibrate with the surrounding lattice temperature. Various methods are used to monitor the relaxation of this nonequilibrium temperature differential, including photoemission of the hot electrons into energy analyzers, and optical measurements of the heating-induced reflectivity change. In such measurements, a sequentially time-delayed probing laser pulse is used to sample the electron temperature or energy distribution. *See* OPTICAL PULSES; PHOTOEMISSION; REFLECTION OF ELECTROMAGNETIC RADIATION.

Not only can the effect of inelastic electron-phonon collisions be observed, but the early evolution of a nonequilibrium electron energy distribution can also be resolved by time-resolved photoemission (**Fig. 3**). Soon after heating of the conduction electrons by a laser pulse, the nonequilibrium nature of the electron system is evidenced by a high-energy distribution extending beyond the best-fit equilibrium Fermi distribution. The energy distribution then relaxes back to an equilibrium electronic temperature via electron-electron and electron-phonon
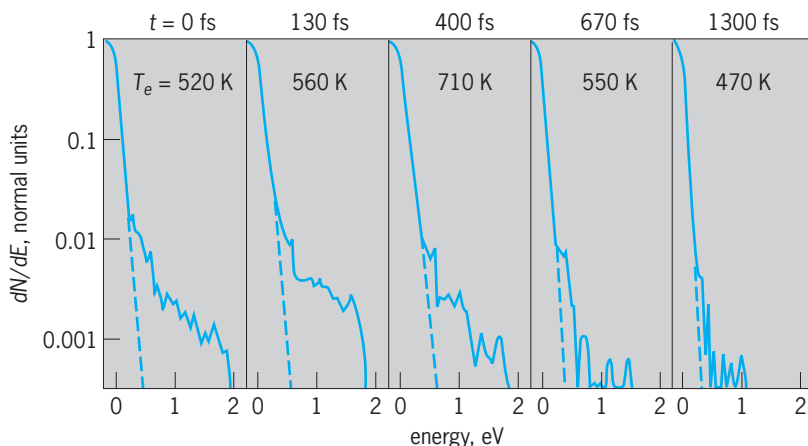

Fig. 3. Electron energy distribution function (density of states) versus energy at five time delays after absorption of an 180-femtosecond laser pulse by a gold film. The broken line is the best Fermi distribution fit at the electron temperature shown. (*After W. S. Fann et al., Electron thermalization in gold, Phys. Rev. B, 46(20):13592–13595, 1992*)

scattering. Subsequent electron-phonon scattering cools the electron distribution to the lower lattice temperature. When analyzed by suitable models, time-resolved optical measurements can yield values for both the electron-electron and electron-phonon energy relaxation rate parameters.

Semiconductor measurements. A measure of the relaxation time may be extracted from the expression for the conductivity stated above by combining the conductivity with a measurement of the Hall coefficient. It turns out that the product of the conductivity, the Hall coefficient, and the speed of light equals the mobility, which is a measure of the speed of response and power dissipation in semiconductor devices, and is thus an important figure of merit for evaluating the performance of integrated circuits. The mobility, in turn, equals the product of the relaxation time and the ratio of the electron charge to effective mass. *See* HALL EFFECT; INTEGRATED CIRCUITS.

The behavior of semiconductors is strongly influenced by another lifetime, the minority-carrier recombination time, which is the mean time an electron can survive before recombining with a hole in *p*-type material (or vice versa). Recombination times range from 10^{-4} to 10^{-9} s, many orders of magnitude longer than the transport relaxation time discussed above. They are measured by methods such as photocurrent decay, the characteristic time for the excess current created by a light pulse to decay. The dynamics of minority carrier transport may be studied in some detail by injecting excess carriers into a material with an applied electric field and observing the decay of the current pulse downstream. Ultrafast optical methods are also applied to the measurement of a variety of relaxation processes in semiconductors. This is especially relevant to gallium arsenide (GaAs) and related materials, which offer the prospect of very high speed devices for communications and computers. *See* MICROWAVE SOLID-STATE DEVICES.

The key to inventing the junction (or bipolar) transistor was fabricating a base region thin enough that a minority carrier injected from the adjacent emitter could survive recombination long enough to reach the collector. Recombination also limits solar-cell efficiency, since the electron-hole pairs must survive long enough to contribute to the current at the cell terminals. Recombination centers or traps may be impurities, lattice imperfections, or surface states. A clean silicon surface contains one dangling bond per atom which acts as a trap. Fabrication of metal-oxide-semiconductor (MOS) transistors and integrated circuits includes passivation of these surface states during oxide formation. An internal version of surface states is the dangling bond on lattice imperfections such as grain boundaries and dislocations. This is why semiconductor devices are ordinarily fabricated on single crystals of high perfection. The use of amorphous or polycrystalline materials, particularly in large-area solar cells, depends on methods that neutralize the trapping action of dangling bonds and thereby increase minority-carrier recombination time. *See* ELECTRON-HOLE RECOMBINATION; GRAIN BOUNDARIES; JUNCTION TRANSISTOR; SEMICONDUCTOR; SOLAR CELL; TRANSISTOR; TRAPS IN SOLIDS.   Gary L. Eesley

Bibliography. N. W. Ashcroft and N. D. Mermin, *Solid State Physics,* 1976; M. Ya Azbel', M. I. Kaganov, and I. M. Lifshitz, Conduction electrons in metals, *Sci. Amer.*, 228(1):88–98, 1973; R. H. Bube, *Electrons in Solids: An Introductory Survey*, 3d ed., 1992; M. L. Cohen and J. R. Chelikowsky, *Electronic Structure and Optical Properties of Semiconductors*, 2d ed., 1989; G. Grimvall, *The Electron-Phonon Interaction in Metals*, 1981; S. M. Sze, *Physics of Semiconductor Devices*, 2d ed., 1981.

Relay

An electromechanical, solid-state, or digital device operated by variations in the input that, in turn, operate or control other devices connected to the output. Relays are used in a wide variety of applications throughout industry, such as in telephone exchanges, digital computers, motor and sequencing controls, and automation systems. Highly sophisticated relays are utilized to protect electric power systems against trouble and power blackouts as well as to regulate and control the generation and distribution of power. In the home, relays are used in refrigerators, automatic washing machines and dishwashers, and heating and air-conditioning controls. Although relays are generally associated with electrical circuitry, there are many other types, such as pneumatic and hydraulic. Input may be electrical and output directly mechanical, or vice versa. *See* ELECTRIC PROTECTIVE DEVICES; SWITCHING SYSTEMS (COMMUNICATIONS).

Relays using discrete solid-state components, operational amplifiers, or microprocessors can provide more sophisticated designs. For industrial, commercial, or residential control, they are used particularly in applications where the relay and associated equipment are packaged together. The basic operation may be complex, but frequently it is similar or equivalent to the units described. For industrial or utility power systems they are used separately in conjunction with appropriate switching devices. *See* AMPLIFIER; MICROPROCESSOR; OPERATIONAL AMPLIFIER.

Basic classifications. A basic, simplified block diagram of a relay is shown in **Fig. 1**. The actual physical embodiment of the blocks varies widely, and one or more blocks may not exist or may be combined in practice. Classifying electrical-type relays by function and somewhat in the order of increasing complexity, there are (1) auxiliary relays, (2) monitoring relays, (3) regulating relays, (4) reclosing, synchronism check, and synchronizing relays, and (5) protective relays.

While the application and protection principles are the same, the relay units may be electromechanical, solid-state, or based on digital microprocessors. Originally, all relays were electromechanical. Electromechanical relays are still used in many
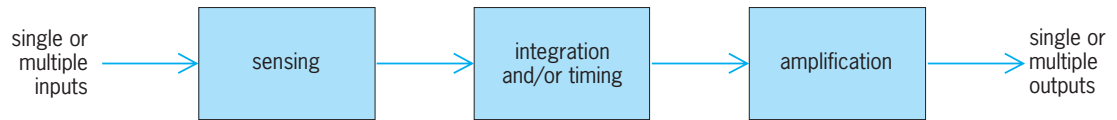
Fig. 1. Basic block diagram of a relay.

applications because of their relative simplicity, long life, and proven high reliability. However, in the late 1960s, solid-state components such as transistors and diodes began to be used. In this rapidly changing technology, operational amplifiers were used briefly. Relays based on digital technology are now preferred, particularly for protective relays, and are replacing the electromechanical or solid-state relays as existing applications or facilities are modified or new ones are installed.

Electromechanical relays. Relays of this type can be illustrated by several common basic units. An auxiliary relay (clapper or telephone type) is shown in **Fig. 2***a*. The sensing unit is the electric coil; when the applied current or voltage exceeds a threshold value, the coil activates the armature, which operates either to close the open contacts or to open the closed contacts. Relays of this type are used as contact multipliers, and to provide the necessary contact outputs for operating the trip coils of circuit breakers for solid-state digital-type relays. Other applications are to isolate circuits or to use a low-power input to control higher-power input.

The plunger type (Fig. 2*b*) operates on the principle of electromagnetic attraction. This relay is composed of a coil, plunger, and set of contacts. When current I flows in the coil, a force is produced that causes the plunger to move and close the relay contacts. Both the clapper and plunger types can be designed to operate on either direct or alternating current. *See* ELECTROMAGNET; SOLENOID (ELECTRICITY).

The balanced polar unit (Fig. 2*c*) is used as a monitoring relay. The unbalanced type is shown in Fig. 2*d*. A permanent magnet normally polarizes the pole faces on either side of the armature, and direct current is applied to the coil that is around the armature. At nominal values of current, the armature can be adjusted to float in the air gap between the pole pieces by the magnetic shunts, but when this current decreases or increases significantly the armature is magnetized and moves to one of the pole pieces to close the contacts. Thus this relay monitors the current level in an electric circuit. The sensing, integrating, and amplifying functions indicated in the block diagram of Fig. 1 are combined in the single mechanical element and its contacts for this polar unit.

The relay unit known as the induction-disk type (Fig. 2*e*) is widely used in regulating and protective relays. With alternating current or voltage applied to the main coil, transformer action induces current in the secondary circuit connected to the upper poles. Fluxes produced by the currents flowing in the upper pole circuit induce eddy currents in the rotor disk. Interaction between rotor eddy currents and

the flux from the lower pole produces mechanical torque to rotate the disk, which is of the order of 3 in. (8 cm) in diameter, thus closing the contacts. This is the split-phase motor principle, where two out-of-phase fluxes produce torque in a rotor. The upper pole may be supplied from another source to permit comparison of two quantities. A spring automatically resets the disk after the relay has operated. By use of the principles of electromagnetic attraction and electromagnetic induction, protective relays can be built to respond to all abnormal conditions that may occur. *See* EDDY CURRENT; ELECTROMAGNETIC INDUCTION; INDUCTION MOTOR; TRANSFORMER.

By design of the coils, damping magnet, and spring, a variety of inverse time overcurrent and under- or overvoltage characteristics are produced. With reference to Fig. 1, the sensing unit is approximated by the coil, the armature and its gap by the timing unit, and the contacts by the amplifier.

For regulating, the output is arranged to adjust or regulate the input within prescribed limits; for monitoring, the output is used to alarm or to shut down the input equipment.

Reclosing, synchronism check, and synchronizing relays provide an automatic sequence of operations. Good examples of these are the timer control relays on automatic washing machines and dishwashers, which provide the necessary program in the various washing operations.

Another widely used basic relay unit is the cylinder or cup type (Fig. 2*f*). The operation is similar to a two-phase induction motor. With out-of-phase currents, voltages, or current and voltage applied to the two windings, flux is produced to rotate the thin-walled rotor in the form of a cylinder or cup. Contacts are attached to the rotor. The contacts and stops permit only a very small movement of the rotor, just enough to provide opening or closing of the contacts. This unit is used in directional sensing and in distance-type relays.

Solid-state relays. The growing complexity of industrial control and modern power systems has resulted in a need for relays, and particularly protective relays, with a higher level of performance and more sophisticated characteristics. This has been made possible by the development of semiconductors and other associated components, which can be utilized in relay designs. These are referred to as solid-state or static relays. All of the functions and characteristics of the electromechanical relays can be performed by solid-state devices, either as discrete components or integrated circuits. They may be divided into two categories: analog circuits that are either fault-sensing or measuring circuits, and digital logic circuits for operation on logical variables. *See* INTEGRATED CIRCUITS; LOGIC CIRCUITS.
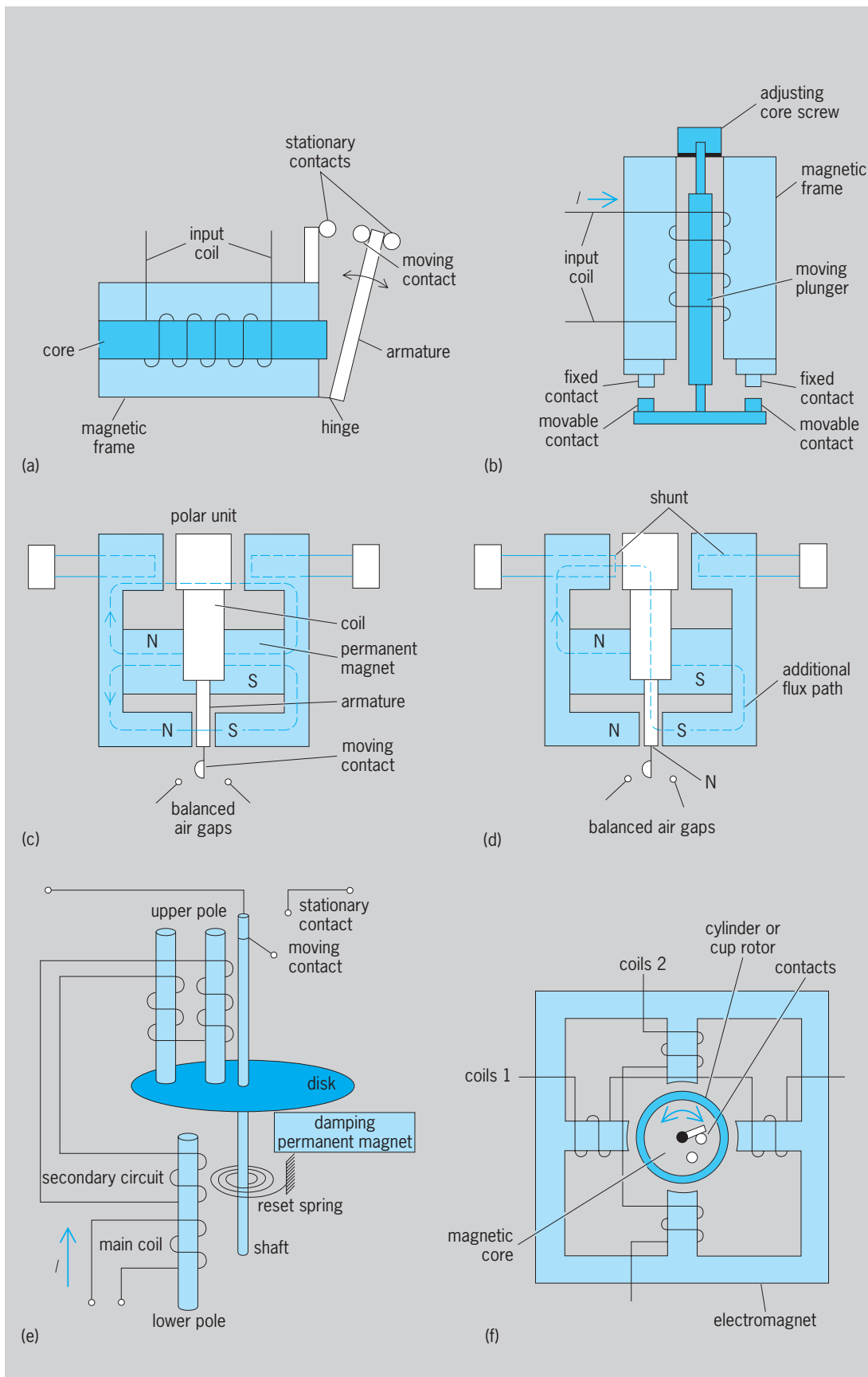
Fig. 2. Electromechanical relays. (*a*) Clapper or telephone type. (*b*) Plunger type. (*c*) Balanced polar unit. (*d*) Unbalanced polar unit. (*e*) Induction-disk type. (*f*) Cylinder or cup type (top view). (*Parts a, b, and f after J. L. Blackburn, Protective Relaying, Principles and Application, Marcel Dekker, 1987*)
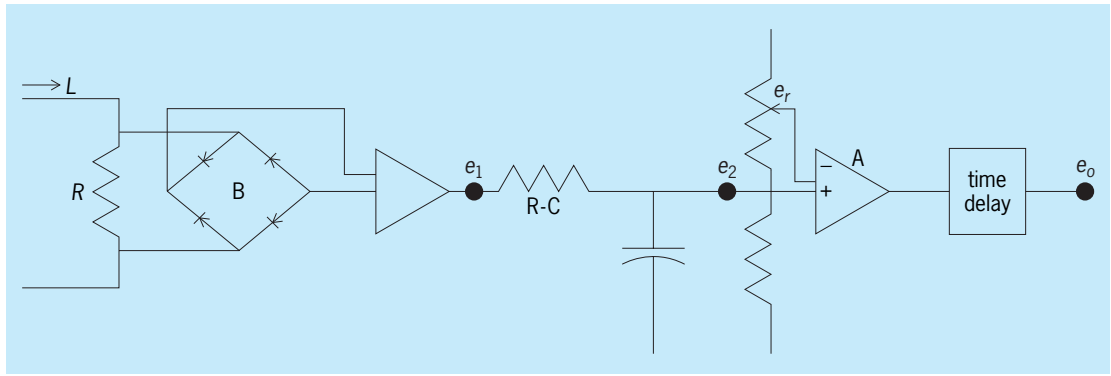
Fig. 3. A possible circuit configuration for a solid-state overcurrent relay. (*After S. H. Horowitz and A. G. Phadke, Power System Relaying, Research Studies Press, 2d ed., 1995*)

In the circuit shown in **Fig. 3**, the input current I is passed through the resistive shunt R, full-wave rectified by the bridge rectifier B, filtered to remove the ripple by the R-C filter, and applied to a high-gain summing amplifier A. The other input of the summing amplifier is supplied with an adjustable reference voltage e_r. When the input on the positive input of the summing amplifier exceeds the reference setting, the amplifier goes high. This step change is delayed by a time-delay circuit, in order to provide immunity against transient signals. By making the time-delay circuit adjustable, and making the amount of the delay depend on the magnitude of the input current, an instantaneous or a time-delay overcurrent relay characteristic can be obtained.

Fig. 4. Typical block diagram of a microprocessor relay. (*After Institute of Electrical and Electronics Engineers, Power Engineering Society, Microprocessor Relays and Protection Systems, Report 88EH0269-1-PWR, 1988*)

Microprocessor relays. Early computer relays were designed to replace existing protective functions, including transmission line, transformer, and bus protection. Some relays employed microprocessors to make the relaying decisions, using digitized analog signals, while others continued to use analog concepts to make relaying decisions and digital techniques for the necessary logic and auxiliary functions. However, both these types of digital relay provided two major advantages: (1) the ability of the relay to diagnose itself, a capability that could be obtained with an analog relay, if at all, only with great effort, cost, and complexity; and (2) a communication capability that allowed the relay to warn system operators when it was not functioning properly, permitted remote diagnostics and possible correction, and provided local and remote readout of its settings and operations. As research on digital relays continued, and confidence was gained, another dimension was added to the reliabilty of the protective system. An inherent feature of the software-dominated digital relay is its ability to adapt itself to changing system conditions in real time. These changes can be initiated by local inputs or by signals sent from a central computer. The digital relay is likely to become the preferred protective system due to the self-diagnostic capability, the hierarchical nature, and the data-sharing abilities of microprocessors, and their ability to adapt settings and other characteristics to actual system conditions in real time.

The major functional blocks for a microprocessor relay are shown in **Fig. 4**. Some of the blocks shown may not be required in certain relay designs.

The analog input subsystem translates the power-system information of current, voltage, frequency, temperature, and so forth, as required for relay fault detection and operation. It provides protection from power-system transients and overvoltages, converts currents to equivalent voltage signals, reduces voltages levels, and prevents high-frequency components from distorting the fault quantities.

The analog interface provides the analog-to-digital conversion and the multiplexing hardware. The status of the power-system circuit breakers, associated isolators, relay targets, and voltage levels is supplied to the relay through the digital input subsystem. The
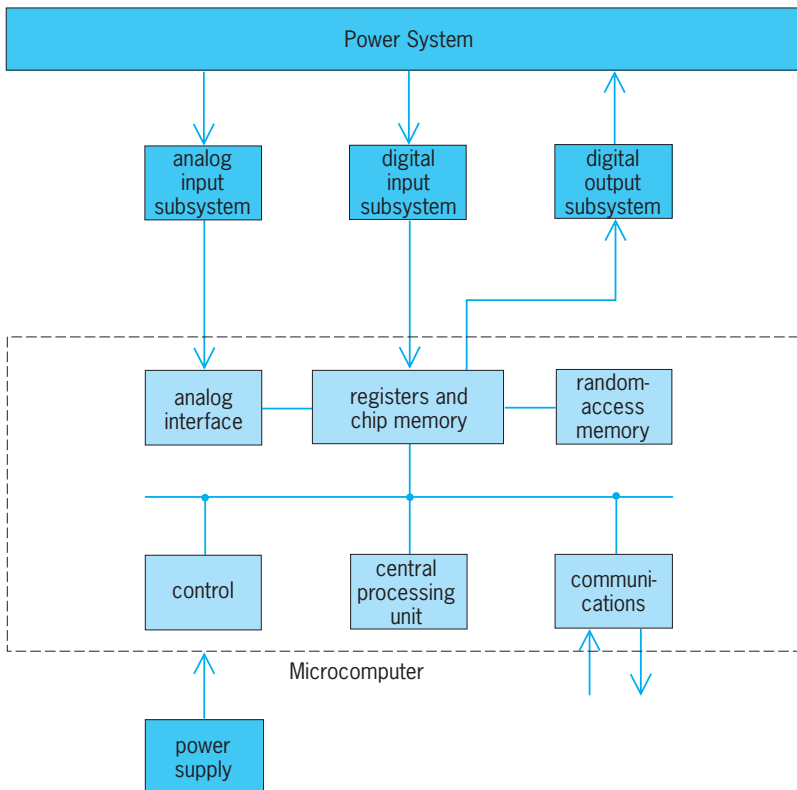
relay program, settings, and logic are stored in chip memory. The control, central processing unit, and registers operate together to execute the program via chip and random-access memory. The relay output is via the digital output subsystem.

Where the relay is part of a pilot relaying system, the communications block provides the required channel information both to and from the remote terminals of the protected line. The random-access memory provides storage of pre- and postfault or disturbance information with appropriate time information. This is accessed through the digital output subsystem for system analysis and fault location.

J. Lewis Blackburn; Stanley H. Horowitz

Bibliography. J. L. Blackburn, *Protective Relaying, Principles and Applications*, 2d ed., Marcel Dekker, 1997; W. A. Elmore (ed.), *Protective Relaying: Theory and Applications*, 2d ed., Marcel Dekker, 2004; S. H. Horowitz and A. G. Phadke, *Power System Relaying*, 2d ed., Research Studies Press Ltd, Taunton, Somerset, England, 1995; Institute of Electrical and Electronics Engineers, Power Engineering Society, *Microprocessor Relays and Protection Systems*, Report 88EH0269-1-PWR, 1988; C. R. Mason, *The Art and Science of Protective Relaying*, Wiley, New York, 1955; L. P. Singh, *Digital Protection: Protective Relaying from Electromechanical to Microprocessor*, Wiley, 1994.

Reliability, availability, and maintainability

Reliability is the probability that an engineering system will perform its intended function satisfactorily (from the viewpoint of the customer) for its intended life under specified environmental and operating conditions. Maintainability is the probability that maintenance of the system will retain the system in, or restore it to, a specified condition within a given time period. Availability is the probability that the system is operating satisfactorily at any time, and it depends on the reliability and the maintainability. Hence the study of probability theory is essential for understanding the reliability, maintainability, and availability of the system. *See* PROBABILITY.

Reliability is basically a design parameter and must be incorporated into the system at the design stage. It is an inherent characteristic of the system, just as is capacity, power rating, or performance. A great deal of emphasis is placed on quality of products and services, and reliability is a time-oriented quality characteristic. There is a relationship between quality or customer satisfaction and measures of system effectiveness, including reliability and maintainability (**Fig. 1**). Customers are concerned with the performance of the product over time.

General Principles

The reliability level must be established at the design phase, and subsequent testing and production will not raise the reliability without a basic design change. With increasing system complexity, reliabil-
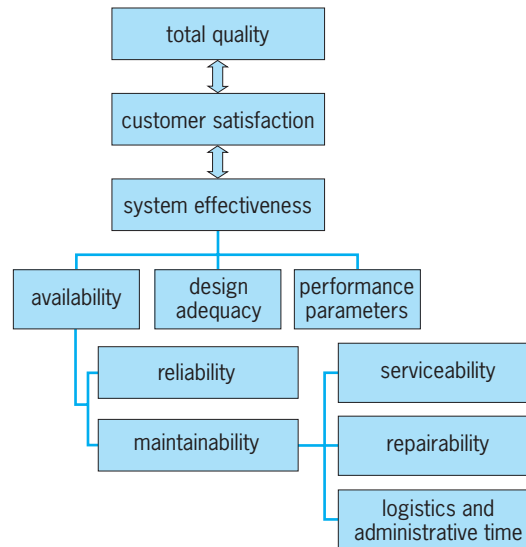
Fig. 1. Relationship between reliability, maintainability, and customer satisfaction.

ity becomes an elusive and difficult design parameter to define and achieve. It also becomes more difficult to control, demonstrate, and ensure as an operational characteristic under the projected conditions of use by the customer. However, history has demonstrated that where reliability is recognized as a necessary program development component, with the practice of various reliability engineering methods throughout the evolutionary life cycle of the system, reliability can be quantified during the specification of design requirements, predicted by testing, controlled during production, and sustained in the field.

System reliability. To analyze and measure the reliability and maintainability characteristics of a system, there must be a mathematical model of the system that shows the functional relationships among all the components, the subsystems, and the overall system. The reliability of the system is a function of the reliabilities of its components. A system reliability model consists of some combination of a reliability block diagram or a cause-consequence chart, a definition of all equipment failure and repair distributions, and a statement of spare and repair strategies. All reliability analyses and optimizations are made on these conceptual mathematical models of the system.

Reliability block diagram. A reliability block diagram is obtained from careful analysis of the manner in which the system operates. An analysis has to be done of the effects on overall system performance of failure of the various components; the support environment and constraints, including such factors as the number and assignment of spare parts and repair persons; and the mission for the system.

The engineering analysis, which has to be done on the system in order to develop a reliability model, consists of four steps: (1) development of a functional block diagram of the system based on physical principles governing the operations of the system; (2) development of the logical and topological
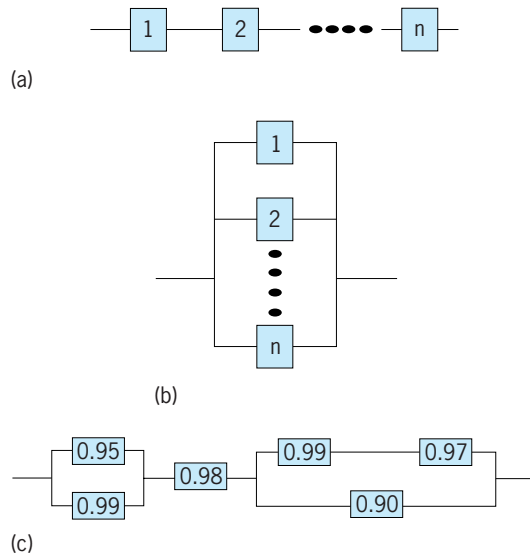
(a)

(b)

(c)

Fig. 2. System reliability block diagrams. (*a*) Series system. (*b*) Parallel system. (*c*) Combination system.

relationships between functional elements of the system; (3) use of performance evaluation studies to determine the extent to which the system can operate in a degraded state; and (4) definition of the spares and repairs strategies (for maintained systems).

Based on the engineering analysis, a reliability block diagram is developed that is used for calculating various measures of reliability and maintainability. The reliability block diagram depicts all the success or failure combinations for the system. Two of the guidelines for drawing these diagrams are: A group of components that are essential for the performance of the system and its mission are drawn in series (**Fig. 2***a*); and components that can substitute for other components are drawn in parallel (Fig. 2*b*). The failure behavior of all the redundant components must be specified. Some of the common types of redundancies are (1) active redundancy or hot standby, where the component has the same failure rate as if it were operating in the system; (2) passive redundancy, spare or cold standby, where the standby component cannot fail—generally assumed of spare or shelf items; (3) warm standby, where the standby component has a lower failure rate than the operating component—usually a realistic assumption.

Structure. The structure is called a series structure when the system functions if and only if all the *n* components of the system function (Fig. 2*a*). The components are assumed to fail or function independently of one another.

The structure is said to be parallel when the system functions if at least one of the *n* components of the system functions (Fig. 2*b*).

The reliability of a system that consists of both series and parallel subassemblies can also be computed (Fig. 2*c*). The reliability values (probability of success) for each component is given in each block.

Reliability measures. Various measures are used to quantify and evaluate the reliability of a system. The reliability function is used to evaluate reliability over

time *t*. Reliability is concerned with the successful performance of a system, and it is evaluated by the time-to-failure random variable, *T*. Then reliability (*R*) at time *t*, *R*(*t*), is the probability that the system will not fail by time *t*. This definition can be expressed mathematically by Eq. (1). Here $P[T > t]$ is the prob-

$$
\begin{aligned}
R(t) &= P[T > t] \\
&= 1 - P[T \le t] \\
&= 1 - F(t) \\
&= 1 - \int_0^t f(\tau)\,d\tau
\end{aligned}
\tag{1}
$$

ability that the time-to-failure *T* is greater than *t*, and *f*(*t*) and *F*(*t*) are the probability density function and cumulative distribution function for *T*, respectively. For example, if the time to failure is exponentially distributed, then these functions are given by Eqs. (2) and (3), and reliability at time *t* is given by Eq. (4). If

$$
f(t) = \lambda e^{-\lambda t} \qquad\qquad t \ge 0, \lambda > 0 \quad (2)
$$

$$
F(t) = \int_0^t \lambda e^{-\lambda t} d\tau = 1 - e^{-\lambda t} \qquad t \ge 0 \tag{3}
$$

$$
R(t) = e^{-\lambda t} \qquad\qquad t \ge 0 \tag{4}
$$

there is some specified time or mission duration, t^*, for which the sytem must perform its intended function, then the mission reliability is given by $R(t^*)$. For many products, it is convenient to express reliability in terms of the mean life or expected life, given by Eq. (5). The quantity $E[T]$ is also called the mean

$$
E[T] = \int_0^\infty t f(t)\,dt = \int_0^\infty R(t)\,dt \tag{5}
$$

time to failure (MTTF) or mean time between failures (MTBF) for repairable systems. For the exponential distribution, it is given by Eq. (6).

$$
E[T] = \frac{1}{\lambda} = \text{MTBF} \tag{6}
$$

Another common measure of reliability is B_α life, which is the α percentile for the life of the product. Thus, $\alpha/100 = F(B_\alpha)$. B_{10} is the 10th percentile for the life of the product, and many times is used to specify the life requirements for the product.

If there is a large population of the items whose reliability is of interest, then for replacement and maintenance purposes it becomes important to determine the rate at which the items in the population, which have survived at any point in time, will fail. This is called the failure rate or hazard rate *h*(*t*), and is given by $h(t) = f(t)/R(t)$.

The failure rate for most components follows the life characteristic curve (**Fig. 3***a*). There are three types of failures, namely quality failures, stress-related failures, and wear-out failures (Fig. 3*b*). The sum total of these failures gives the overall failure rate (Fig. 3*a*). The failure rate curve, also known as the bathtub curve (Fig. 3*a*), has three distinct periods.
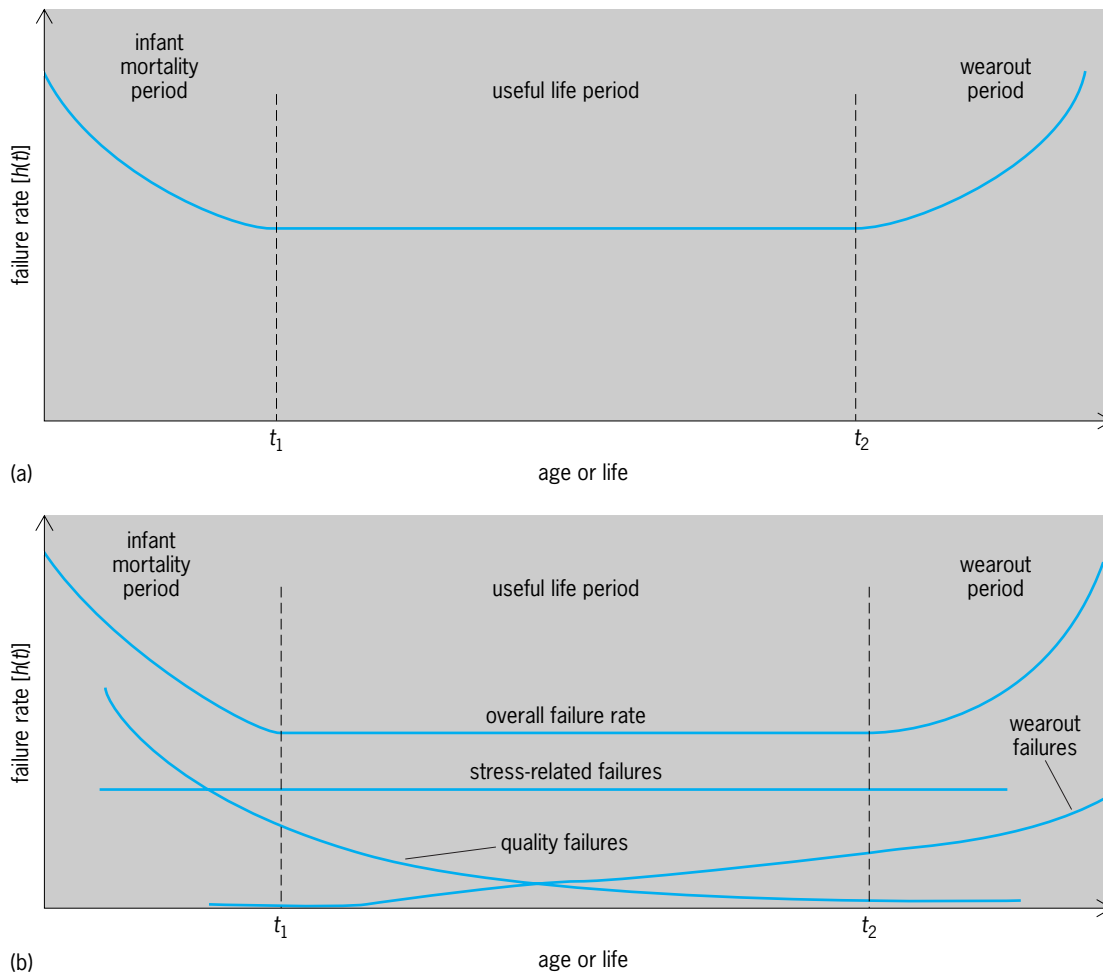
Fig. 3. Failure rate curves. (*a*) Failure rate–life characteristic curve (bathtub curve). (*b*) Failure rate curves based on components of failure.

The initial decreasing failure rate is termed infant mortality and is due to the early failure of substandard products. Latent material defects, poor assembly methods, and poor quality control can contribute to a high initial failure rate. A short period of in-plant product testing, termed burn-in, is used by manufacturers to eliminate these early failures from the consumer market. The flat, middle portion of the failure-rate curve represents the design failure rate for the specified product as used by the consumer market. During the useful-life portion, the failure rate is relatively constant. It might be decreased by redesign or by using a higher safety margin. Finally, as products age they reach a wear-out phase characterized by an increasing failure rate.

Durability. Durability is a special case of reliability. Durability is the probability that an item will successfully survive its projected service life, overhaul point, or rebuild point (whichever is the most appropriate durability measure for the item) without a durability failure. A durability failure is considered to be a malfunction that precludes further operation of the item and is great enough in cost, safety, or time to restore that the item must be replaced or rebuilt.

Maintainability. Maintainability is a measure of the ease and rapidity with which a system or equip-ment can be restored to operational status following a failure. It is a characteristic of equipment design and installation, personnel availability in the required skill levels, adequacy of maintenance procedures and test equipment, and the physical environment under which maintenance is performed. Maintainability is expressed as the probability that an item will be retained in or restored to a specific condition within a given period of time, when the maintenance is performed in accordance with prescribed procedures and resources.

One measure of maintainability is the mean time to repair (MTTR), which is defined as the mean of the distribution of system repair time. If a system has N subsystems, the MTTR for the system is given by Eq. (7), where λ_i is the failure rate of the ith re-

$$\text{MTTR} = \frac{\sum_{i=1}^{N} \lambda_i t_i}{\sum_{i=1}^{N} \lambda_i} \qquad (7)$$

pairable subassembly in the system, and t_i is the expected time to repair the system when its ith subassembly fails. The MTTR is also given by the total corrective maintenance time (the time required for actions performed, as a result of failure, to restore the equipment or system to operable conditions)

divided by the total number of corrective mainte-
nance actions (actions resulting from the nonoper-
ablity of the equipment or system) during a given
period of time.

Several other measures for maintainability are in
use. The mean time between unscheduled mainte-
nance actions (MTBUMA) is the average number of
time units for which a system functions successfully
between successive unscheduled maintenance ac-
tions. Active maintenance time is the time during
which preventive and corrective maintenance work
is actually being done on the item. Different types
of maintenance are distinguished: Periodic mainte-
nance is maintenance performed on equipment on
the basis of hours of operation or calendar time
elapsed since the last inspection; and preventive
maintenance is the systematic care, servicing, and
inspection of equipment and facilities for the pur-
pose of maintaining them in serviceable condition
and detecting and correcting incipient failures. Fi-
nally, the maintenance ratio (MR) is given by Eq. (8),

$$MR = \frac{MPH}{OH} \qquad (8)$$

where MPH is the number of maintenance person-
hours and OH is the number of system operating
hours.

Availability. Availability is defined as the probabil-
ity that a system is operating satisfactorily at any point
in time. This definition considers only the operat-
ing time and the downtime excluding idle time. The
steady-state availability (A) of a system with known
MTBF and MTTR is given by Eq. (9).

$$A = \frac{MTBF}{MTBF + MTTR} \qquad (9)$$

Design reliability methodology. The best way to
improve reliability is at the concept and design
stage by using various failure prevention techniques.
Two common techniques are failure mode, effects,
and criticality analysis (FMECA), which represents
a bottom-up approach; and fault-tree analysis (FTA),
which represents a top-down approach. Both repre-
sent qualitative as well as quantitative approaches for
assessing the consequences of failure and its preven-
tion.

Failure mode, effects, and criticality analysis. This is an it-
erative documented process of a systematic nature
performed to identify basic faults at the part level
and to determine their effects at high levels of as-
sembly (**Fig. 4**). Failure mode, effects, and criticality
analysis can be performed by utilizing either actual
failure modes from field data or hypothesized fail-
ure modes derived from design analysis, reliability
prediction activities, and experiences relative to the
manner in which parts fail. In their most complete
form, failure modes are identified at the part level,
which is usually the lowest level of direct concern
to the designer of the product or process. In addi-
tion to providing insight into failure cause-and-effect
relationships, the failure mode and effects analysis
provides a disciplined method for proceeding part
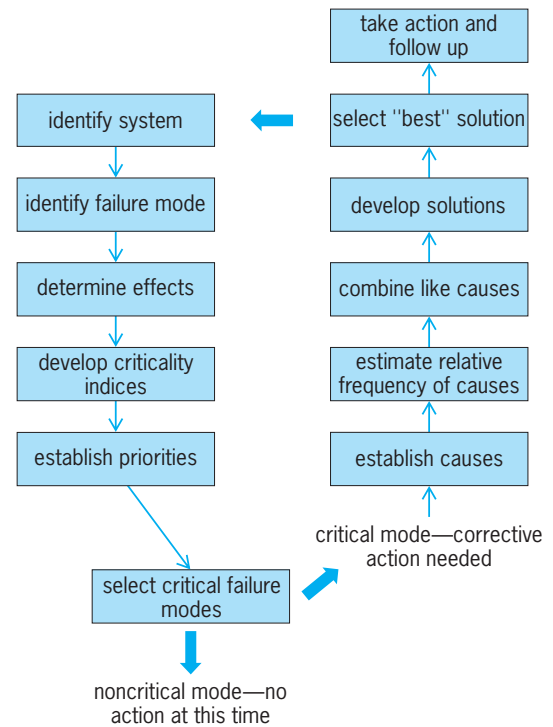
Fig. 4. Diagram of methodology for failure mode, effects,
and criticality analysis.

by part through the system to assess failure conse-
quences.

Failure modes are analytically induced into each
component, and failure effects are evaluated and
noted, including severity and frequency (or probabil-
ity) of occurrence. As the first mode is listed, the cor-
responding effect on performance at the next higher
level of assembly is determined. The resulting failure
effect becomes, in essence, the failure mode that im-
pacts the next higher level. Iteration of this process
results in establishing the ultimate effect at the sys-
tem level. Once the analysis has been performed for
all failure modes, it is usually the case that each effect
or symptom at the system level is caused by several
different failure modes at the lowest level. This re-
lationship to the end effect provides the basis for
grouping the lower-level failure modes.

By using this approach, probabilities for the occur-
rence of the system effect can be calculated, based on
the probability occurrence of the lower-level failure
modes. Based on these probabilities, and a severity
factor assigned to the various system effects, a criti-
cality number can be calculated. Criticality numerics
provide a method of ranking the system-level effects
derived previously. Criticality numerics also provide
the basis for corrective action priorities, engineering
changes, and resolution of problems in the field.

Fault-tree analysis. This is a tool that lends itself well
to analyzing failure modes found during design, fac-
tory test, or field data returns. The fault-tree analysis
procedure can be characterized as an iterative docu-
mented process of a systematic nature performed to
identify basic faults, determine their causes and ef-
fects, and establish their probabilities of occurrence.

The approach involves several steps, among which is the structuring of a highly detailed logic diagram that depicts basic faults and events that can lead to system failure and safety hazards. The next steps are collecting basic fault data and failure probabilities for use in computation; and using computational techniques to analyze the basic faults, determine failure mode probabilities, and establish criticalities. The final step involves formulating corrective suggestions that, when implemented, would eliminate (or minimize) those faults considered critical.

Reliability growth, demonstration, and testing. As the product goes through the various life cycles, reliability should be estimated and predicted. These values, when plotted at selected points in the life cycle, result in a growth curve that reflects comparative reliability levels. The improvement in reliability is dependent upon the nature of the product. Desirable reliability growth results from planning, designing, testing, producing, and using the product according to specified rules.

The purpose of reliability demonstration and testing is to determine the current reliability levels of the product. Tests must be designed in such a manner that maximum information can be obtained from a minimum amount of testing. A major problem in the design of adequate tests is that of simulating the real-world environment. During its lifetime the product is subjected to many environmental factors, such as temperature, vibrations and shock, and rough handling. These stresses may be encountered singly, simultaneously, or sequentially. In addition, there are many other random factors. *See* OPTIMIZATION; RISK ASSESSMENT AND MANAGEMENT.

Software Reliability

Software reliability is the probability that software will not cause the failure of a product for a specified time under specified conditions. This probability is a function of the inputs to and use of the product as well as a function of the existence of faults in the software. The inputs to the product determine whether existing faults, if any, are encountered. Software reliability is also sometimes defined as the ability of a program to perform a required function under stated conditions for a stated period of time.

As computers spread through society, the size and complexity of software systems continually increase. Many institutions rely heavily on computers to provide vital services. International banking, police investigations, and airline reservations are dependent on computers. Computer software systems support safety-critical functions in the control of nuclear reactors, defense systems, air traffic, space exploration, and many consumer products including medical devices. The complexity of software systems, the trust placed in them, and the potentially catastrophic consequences of software failure demand quantitative guarantees of software quality.

Software faults and software failure. A fault is the manifestation of an error in software. A fault, if encountered, may cause a failure and sometimes is synonymous with a "bug." Failure is the inability of a product or component to perform a required function within specified limits; failures will be observed during testing or actual use of the software.

Whereas hardware failures are frequently due to physical wearout, software failures are due to design faults (bugs). These faults are defects, which may be due to incorrect or incomplete specifications in the design phase of the software, or due to simple programmer error. When the software is executed, it may encounter an input for which the software was not designed or tested. Under some of these inputs, a fault will induce failure.

While the number of faults remaining in a program affects its performance, a more informative measure of the reliability of software is its failure rate (failure intensity), the number of failures occurring per unit time. Indeed, studies have found that systems with many faults can still be highly reliable, if each fault occurs very infrequently.

Design for software reliability. The source of a software fault is a design fault, and hence when a software fault is removed correctly it will never recur. Thus, if all the faults were eventually removed, the software would be fault-free and would execute without failure forever. This is in contrast to hardware reliability where, due to physical causes and degradation, all systems will eventually fail. The techniques to achieve software reliability can be classified under three categories: fault avoidance, fault removal, and fault tolerance.

Fault-avoidance techniques are engineering and management tools to prevent the creation of faults. In addition to correct software requirements and specifications, some of the approaches are control of complexity, performance of top-down functional decomposition, modularization, hierarchical design, and adherence to structured approaches. *See* SOFTWARE ENGINEERING.

Fault-removal techniques attempt to discover and correct faults before software is released. Discovery is done through software testing for quality, reliability, safety, maintainability, and overall qualification. *See* SOFTWARE TESTING AND INSPECTION.

Since it is virtually impossible to guarantee fault-free software, fault-tolerance techniques are used to detect and recover from faults when the software is in use. N-version programming is one such approach. Several independently written programs are run in parallel; if their outputs are in conflict, a predetermined scheme such as majority voting determines the output to be used. *See* FAULT-TOLERANT SYSTEMS.

Software reliability engineering (SRE) activities to achieve high reliability should be applied throughout the life cycle of the product starting from concept, through delivery, to the end of the product's useful life (see **table**).

Software reliability modeling. Models are a major statistical tool for assessing software reliability. Typically, a model combines a probabilistic description of the software failure process with a statistical inference procedure to make predictions of software failure based on past performance data. The data are

Software reliability engineering activities	
Stage of product life cycle	Activities
Concept, feasibility, and requirements	Determine functional profile
	Define and classify failures
	Identify customer reliability needs
	Conduct trade-off studies
	Set reliability objectives
Design and implementation	Allocate reliability among components
	Engineer to meet reliability objectives
	Focus resources based on functional profile
	Manage fault introduction and propagation
	Measure reliability of acquired software
System test and field trial	Determine operational profile
	Conduct reliability growth testing
	Track testing progress
	Project additional testing needed
	Obtain certification for reliability objectives
Delivery, upgrades, and maintenance	Project postrelease staff needs
	Monitor filed reliability versus objectives
	Track customer satisfaction with reliability
	Time introduction of new features
	Guide product and process improvement

obtained from testing and are usually a sequence of execution times between successive failures (interfailure times).

The earliest model for measurement of software reliability was proposed in 1972 independently by Z. Jelinsky and P. Moranda and by M. Shooman. The model makes the following assumptions: (1) failures occur randomly, (2) the program's failure rate is proportional to the number of remaining bugs it contains, and (3) at each failure a fault is fixed. Using the statistical method of maximum likelihood, the interfailure time data are used to estimate the total number of bugs and the constant of proportionality.

More advanced software reliability models and customer-centered tools have been developed. These tools allow software reliability to be measured and improved. One of the first software reliability measurement tools was developed in 1977 and has been commercially available since 1990. It was followed by SMERFS (Statistical Modeling and Estimation of Reliability Functions of Software), developed at the U.S. Naval Surface Warfare Center. Some other tools are SRMP (Software Reliability Modeling Program) and CASRE (Compiler Aided Software Reliability Estimation).

Unlike earlier models, which assume that the random discovery of faults (that is, the input to the program) is the only source of uncertainty in the failure process, more sophisticated models view the software failure process to be the result of two sources of uncertainty: the selection of input to the program, and the effect of debugging. The uncertainty in debugging may be due both to the size (rate and occurrence) of the fault removed and to the fact that debugging may not be perfect. Advanced models also attempt to estimate the probability of software failure when testing reveals no errors. These methods are based on bayesian estimation, and incorporate the testing results, information about the input distribution during use, and prior assumptions about the probability of failure of the software. Software reliability measurement is a useful tool in assessing when or whether software is ready for release, and for estimating software reliability for future maintenance costs. *See* BAYESIAN STATISTICS; QUALITY CONTROL.

Kailash C. Kapur

Bibliography. S. Bittanti (ed.), *Software Reliability Modelling and Identification*, 1989; E. A. Elsayed, *Reliability Engineering*, Addison Wesley, 1996; W. G. Ireson, C. F. Coombs, and R. V. Moss (eds.), *Handbook of Reliability Engineering and Management*, 2d ed., 1996; K. C. Kapur and L. R. Lamberson, *Reliability in Engineering Design*, 1977; D. Kececioglu, *Reliability Engineering Handbook*, vols. 1 and 2, 1991; K. B. Klaasen and J. C. L. van Peppen, *System Reliability: Concepts and Applications*, 1989; M. R. Lyu (ed.), *Handbook of Software Reliability Engineering*, 1996.

Reluctance

A property of a magnetic circuit analogous to resistance in an electric circuit.

Every line of magnetic flux is a closed path. Whenever the flux is largely confined to a well-defined closed path, there is a magnetic circuit. That part of the flux that departs from the path is called flux leakage.

For any closed path of length l in a magnetic field H, the line integral of $H \cos \alpha \, dl$ around the path is the magnetomotive force (mmf) of the path, as in
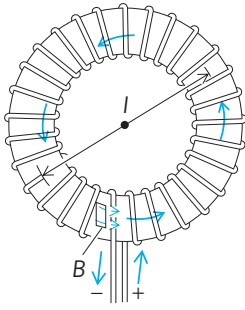
A toroidal coil, illustrating the concept of reluctance.

Eq. (1), where α is the angle between H and the path.

$$\text{mmf} = \oint H \cos \alpha \; dl \qquad (1)$$

If the path encloses N conductors, each with current I, Eq. (2) holds.

$$\text{mmf} = \oint H \cos \alpha \; dl = NI \qquad (2)$$

See MAGNETISM.

Consider the closely wound toroid shown in the **illustration**. For this arrangement of currents, the magnetic field is almost entirely within the toroidal coil, and there the flux density or magnetic induction B is given by Eq. (3), where l is the mean circumference of the toroid and μ is the permeability.

$$B = \mu \frac{NI}{l} \qquad (3)$$

The flux ϕ within the toroid of cross-sectional area A is given by either form of Eqs. (4), which is similar in form to the equation for the electric circuit,

$$\Phi = BA = \frac{\mu A}{l} NI$$
$$\Phi = \frac{NI}{l/\mu A} = \frac{\text{mmf}}{l/\mu A} = \frac{\text{mmf}}{\mathcal{R}} \qquad (4)$$

although nothing actually flows in the magnetic circuit. The factor $l/\mu A$ is called the reluctance $\mathcal{R}$ of the magnetic circuit. The reluctance is not constant because the permeability μ varies with changing flux density. From the defining equation for reluctance, it is seen that when the mmf is in ampere-turns and the flux is in webers, the unit of reluctance is the ampere-turn/weber.

Reluctances in series. For the simple toroid, all parts of the magnetic circuit have the same μ and the same A. More complicated circuits may include parts that differ in permeability, in cross section, or in both. Suppose a small gap were cut in the core of the toroid. The flux would fringe out at the gap, but as a rough approximation, the area of the gap may be considered the same as that of the core.

The magnetic path then has two parts, the core of length l_1 and reluctance $l_1/\mu_1 A$, and the air gap of length l_2 and reluctance $l_2/\mu_2 A$. Since the same flux is in both core and gap, this is considered a series cir-

cuit and Eq. (5) holds. Since the relative permeability

$$\mathcal{R} = \mathcal{R}_1 + \mathcal{R}_2 = \frac{l_1}{\mu_1 A} + \frac{l_2}{\mu_2 A} \qquad (5)$$

of the ferromagnetic core is several hundred or even several thousand times that of air, the reluctance of the short gap may be much greater than that of the much longer core. For any combination of paths in series, $\mathcal{R} = \Sigma l/\mu A$. Then Eq. (6) holds.

$$\Phi = \frac{\text{mmf}}{\Sigma \mathcal{R}} = \frac{\text{mmf}}{\Sigma l/\mu A} \qquad (6)$$

Reluctances in parallel. If the flux divides in part of the circuit, there is a parallel magnetic circuit and the reluctance of the circuit has the same relation to the reluctances of the parts as has the analogous electric resistance. For the parallel circuit Eq. (7) is

$$\frac{1}{\mathcal{R}} = \frac{1}{\mathcal{R}_1} + \frac{1}{\mathcal{R}_2} + \cdots \qquad (7)$$

valid. *See* RELUCTANCE MOTOR.　　Kenneth V. Manning

Reluctance motor

An alternating current motor with a stator winding like that of an induction motor, and a rotor that has projecting or salient poles of ferromagnetic material. When connected to an alternating-current source, the stator winding produces a rotating magnetic field, with a speed of $4\pi f/p$ radians per second ($120f/p$ revolutions per minute), where f is the frequency of the source and p the number of magnetic poles produced by the winding. When the rotor is running at the same speed as the stator field, its iron poles tend to align themselves with the poles of that field, producing torque. If a mechanical load is applied to the shaft of the motor, the rotor poles lag farther behind the stator-field poles, and increased torque is developed to match that of the mechanical load. This torque is given by the equation below,

$$\tau = \frac{p}{2} \phi^2 \frac{dR}{d\delta}$$

where ϕ is the flux per pole, determined largely by the applied voltage. The quantity $dR/d\delta$ is the rate of change of magnetic reluctance per pole with respect to δ, the angle of lag in mechanical radians. This quantity typically varies as $\sin p\delta$. Here, $p\delta = 2\delta_e$, where δ_e is the lag angle in electrical radians. Therefore, at constant torque load the rotor runs in synchronism with the stator field, with the rotor poles lagging the field poles by a constant angle. *See* ELECTRICAL DEGREE.

This motor develops torque only at synchronous speed, and thus no starting torque is produced. For that reason, induction-motor rotor bars are usually built into the pole faces, and the motor starts as an induction motor. When the rotor speed approaches that of the magnetic field, the pole pieces lock in step with the magnetic poles of the field, and the rotor runs at synchronous speed.

Single-phase reluctance motors may be started by the methods used for single-phase induction motors, such as capacitor, split-phase, or shaded-pole starting. *See* ALTERNATING-CURRENT MOTOR; ELECTRIC ROTATING MACHINERY; INDUCTION MOTOR; MOTOR; SYNCHRONOUS MOTOR.   George McPherson, Jr.

Bibliography. T. J. Miller, *Brushless Permanent-Magnet and Reluctance Motor Drives*, 1989; T. J. Miller, *Switched Reluctance Motors and Their Control*, 1993; J. Rosenblat and M. H. Friedman, *Direct and Alternating Current Machinery*, 2d ed., 1998; C. G. Veinott and J. E. Martin, *Fractional and Subfractional Horsepower Electric Motors*, 4th ed., 1986.

Remipedia

A unique class of the Crustacea whose members, at first glance, resemble polychaete worms in overall body form. This impression comes from their multiple serial segments and lack of obvious body divisions (tagmosis; see **illus.**). The lack of body organization suggests that remipeds are the most primitive crustaceans yet recognized. When compared with the other presumably primitive crustacean class, the Cephalocarida, remipeds are relatively large, with mature individuals ranging from 0.6 to 1.7 in. (15 to 42 mm) in body length. Although the trunk appendages are biramous and paddle-shaped in both classes, endites and epipods that are present in cephalocarids are lacking in remipeds. *See* CEPHALOCARIDA.

Discovery. The first remiped species was discovered by divers in 1981 in an achialine cave on Grand Bahama Island. In subsequent years, five more Recent species have been described. Additionally, one previously described fossil species has been assigned to this class. As now constituted, the Remipedia comprise two orders, three families, and four genera; however, two or three potentially new species have already been collected, and as more marine caves are explored additional species will be discovered.

Morphology. Recent remipeds (order Nectiopoda) are characterized by the lack of a carapace but with a distinct cephalic shield covering the head. A pair of frontal processes of undetermined function is located just in front of the biramous, multiarticulated antennules. The antennae are also biramous and paddlelike in form. A lacinia mobilis separates the incisor and molar processes of the mandibles, which are at least partially enclosed in the atrium oris formed by the large labrum, or upper lip. Three pair of postoral food-handling appendages (maxillule, maxilla, and maxilliped) are raptorial in structure, and distinctive among the four genera. The trunk consists of approximately 29–32 somites in mature adults, with no discernible division into thorax and abdomen. All of the serially similar trunk limbs are biramous, paddlelike, and directed ventrolaterally. Eyes are lacking. The anal segment is provided with a pair of caudal rami. Remipeds are hermaphroditic, with female gonopores present on the seventh trunk segment
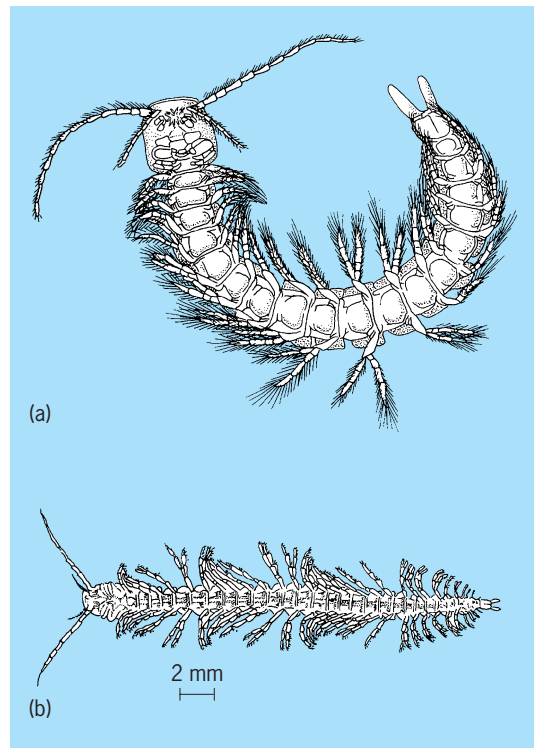
Representative Remipedia. (*a*) *Speleonectes lucayensis* (*after F. R. Schram, Crustacea, Oxford University Press, 1986*). (*b*) *Lasionectes entrichoma* (*after J. Yager and F. R. Schram, Lasionectes entrichoma, new genus, new species* (*Crustacea: Remipedia*) *from anchialine caves in the Turks and Caicos, British West Indies, Proc. Biol. Soc. Wash., vol. 99, 1986*).

and male gonopores on the fourteenth. Excretion is by means of a pair of maxillary glands located in the head. The digestive system consists of a distensible anterior and tubular posterior foregut, a long midgut with paired diverticula in each segment, and a short hindgut.

Ecology. Only limited ecological information on remipeds is available. All species known to date have been found in caves where oxygen levels of the water are low. It is not uncommon for several species to coexist in a single cave. Little is known about feeding habits of remipeds; however, comparisons of various adaptations of the raptorial feeding appendages suggest that species competition is avoided by prey selectivity. At present the distribution of remipeds appears restricted to the subtropical and tropical Atlantic Ocean islands of the Bahamian Archipelago and Canary Islands.

Phylogeny. The evolutionary position of the Remipedia and the relationship to other crustacean taxa are speculative at this time. Clearly the lack of trunk specialization is a primitive condition not seen in other crustacean taxa. Serial homology of trunk appendages is also generally accepted as the primitive state. Similarly, biramous appendages, such as those found in remipeds, have long been postulated as the precursor of the more specialized stenopodus walking leg. However, in view of incorporation of the anterior-most thoracic appendage (maxilliped) into the cephalon (head), the specialized feeding

appendages, and restricted habitat, it seems premature to postulate the Remipedia as being at the base of the crustacean phylogenetic tree. *See* CRUSTACEA.

Patsy A. McLaughlin

Bibliography. F. R. Schram, J. Yager, and M. J. Emerson, *Remipedia,* pt. I: *Systematics*, San Diego Soc. Nat. Hist. Mem. 15, 1986; J. Yager, *Cryptocorynetes haptodiscus*, new genus, new species, and *Speleonectes benjamini*, new species, of remipede crustaceans from anachialine caves in the Bahamas, with remarks on distribution and ecology, *Proc. Biol. Soc. Wash.,* vol. 100, 1987; J. Yager, Remipedia, a new class of Crustacea form a marine cave in the Bahamas, *J. Crust. Biol.*, vol. 1, 1981.

Remote-control system

A control system in which the issuing of the control command and its execution are separated by a relatively significant distance. The system normally includes a command device where the control command is entered, and an actuator that executes it. These are connected by a transmission medium that transmits the command, usually in a coded format.

The transmission medium may be a mechanical link, where the command is transmitted as force; a pneumatic or hydraulic line, where pressure represents the command; an electrical line with a voltage or current signal; or radio or infrared waves that are modulated according to the command. *See* MODULATION.

The simplest remote-control systems are limited to switching-type functions. These systems operate basically in an open loop, that is, without relying on feedback (although an acknowledgment of the command's execution may be provided). Some typical examples are a ceiling lamp turned on and off by a light switch via an electrical wire; the on/off function of a television receiver and videocassette recorder with an infrared remote controller; power switches in distribution networks operated from a control center; and railway switches operated from a remote-control room.

The most characteristic remote-control systems involve feedback that is provided by the human operator. The person issuing the control command senses (usually sees) the result of the control action and guides the system accordingly. This kind of operation can be found, for example, in remote control of the aspects of a television picture, remote control of toy cars and airplanes by wire or radio, remote operation of large construction cranes, and cockpit control of an airplane's engines and control surfaces, usually by means of a hydraulic system. *See* FLIGHT CONTROLS.

Teleoperation represents an important class within remote-control systems with human feedback. In contrast to the systems noted above where the control action is limited to precisely defined functions, teleoperators (or remote manipulators) act as extensions of the human hand. They are employed in situations where access is difficult or impossible

or where the environment is hazardous for humans, such as in underwater and space operations, or in the presence of radiation, chemical, or biological contamination. *See* REMOTE MANIPULATORS.

Strictly speaking, many automatic control systems (that is, feedback control systems with no human operator in the loop) may also be considered as remote controllers. This is the case whenever the sensing of the controlled variable and the automatic formation of the control command are removed from the actuator. A typical example is the heating and air-conditioning system of a building, where room thermostats operate remotely located furnaces, compressors, and fans. Similar systems are frequently found in the control of large industrial plants. *See* CONTROL SYSTEMS.

Janos Gertler

Remote manipulators

Mechanical, electromechanical, or hydromechanical devices which enable a person to perform manual operations while separated from the site of the work. Remote manipulators are designed for situations where direct contact would be dangerous to the human (working with radioactive material), where direct human contact is ill-advised or impossible (certain medical procedures), and where human force-producing capabilities are absent (the disabled) or need to be amplified to complete some task (industrial assembly or construction).

Basic defining elements are common to almost all remote manipulators (**Fig. 1**). An input device or control handle allows the operator to command the remote manipulator. The movement of the input device is received by a control station that translates
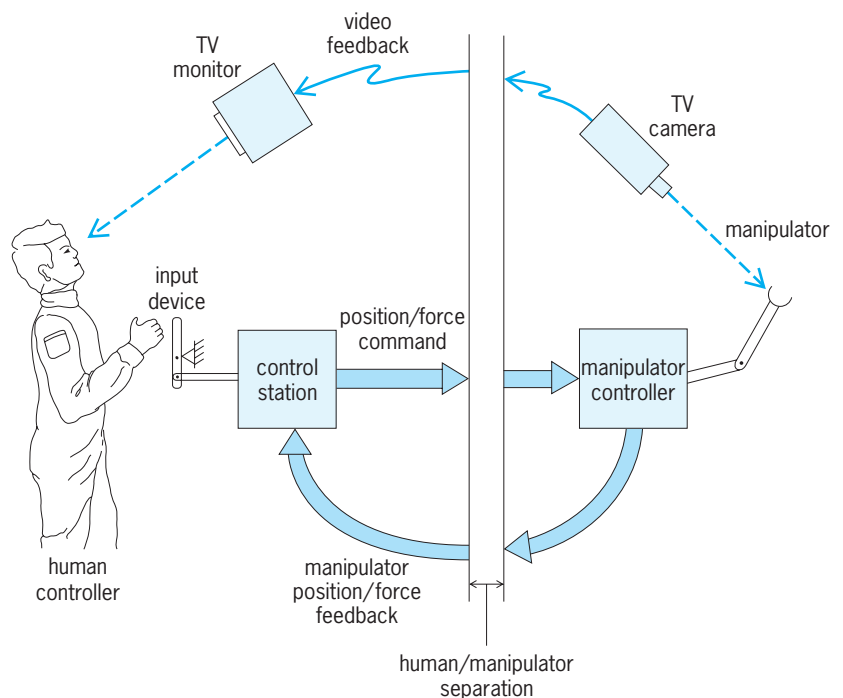

Fig. 1. Elements of a typical remote manipulator system.

the inputs into a form that can be transmitted over the distance separating the human and remote manipulator. This translation can be mechanical, using cables and linkages, or electrical/electromagnetic, using the movement of the input device to generate an electrical or electromagnetic signal that is easily transmitted to the remote manipulator. Since vision is an important cue that humans use in direct manipulation, visual feedback of a remote manipulator's actions typically must be provided. In some remote manipulation systems, tactile feedback to the human operator is provided; that is, forces proportional to those being exerted by the remote manipulator on the object are fed back to the human through the input device. Such force feedback is important in certain tasks where the possibility of damage to the manipulator or object can occur.

Historical development. Perhaps the earliest example of remote manipulation occurred in the mid-1940s, when master-slave devices were designed and constructed to aid in the handling of radioactive materials. The term "master-slave" refers to the parallel construction and operation of the human input device (master) and the remote manipulator (slave). Typically, the remote manipulator was a mechanical copy of the input device used by the human,

although the remote manipulator could be scaled in size. Communication between these early devices was by direct mechanical linkages. Soon, however, electronic communication was used. Later, digital computers were used to develop sophisticated interfaces between operator and remote manipulator. *See* RADIOCHEMICAL LABORATORY.

The master-slave implementation for remote manipulation implies a nearly one-to-one correspondence between human controller input and remote manipulator motion. In some applications this is impractical, such as where the manipulator needs to be moved a considerable distance when performing a task. In such applications, the remote manipulator has been designed to provide a rate of change of position or velocity in response to human controller position changes.

Applications. The growth in the number and variety of remote manipulator applications has been aided by enabling technologies such as digital computers, lightweight materials, and video communication links. Space applications include the space shuttle remote manipulator arm, which has been used to retrieve and launch large satellites (**Fig. 2**). The arm is operated by astronauts in the shuttle orbiter cabin, and employs graphic displays of the forces and
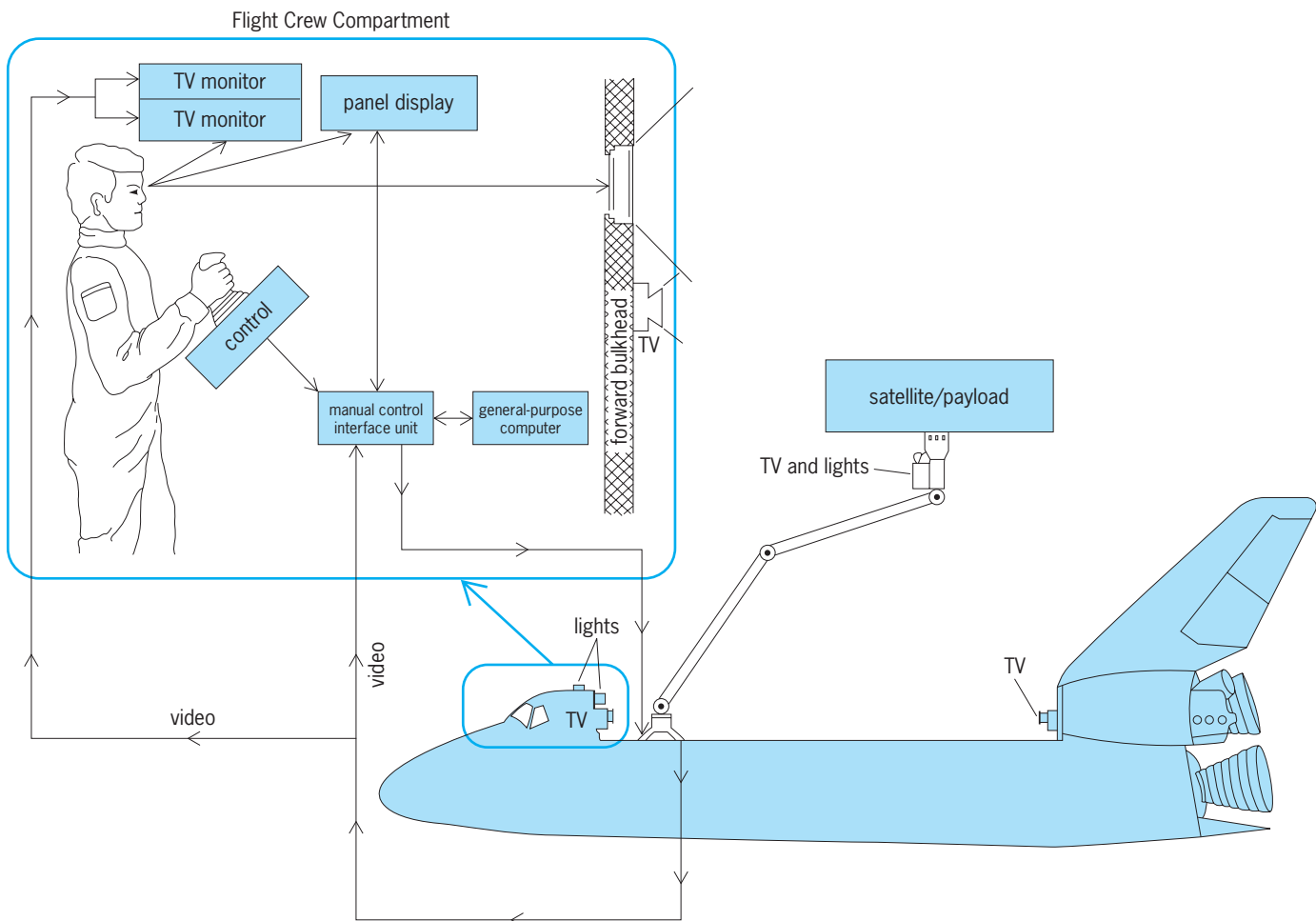
Fig. 2. Space shuttle orbiter with a remote manipulator arm.
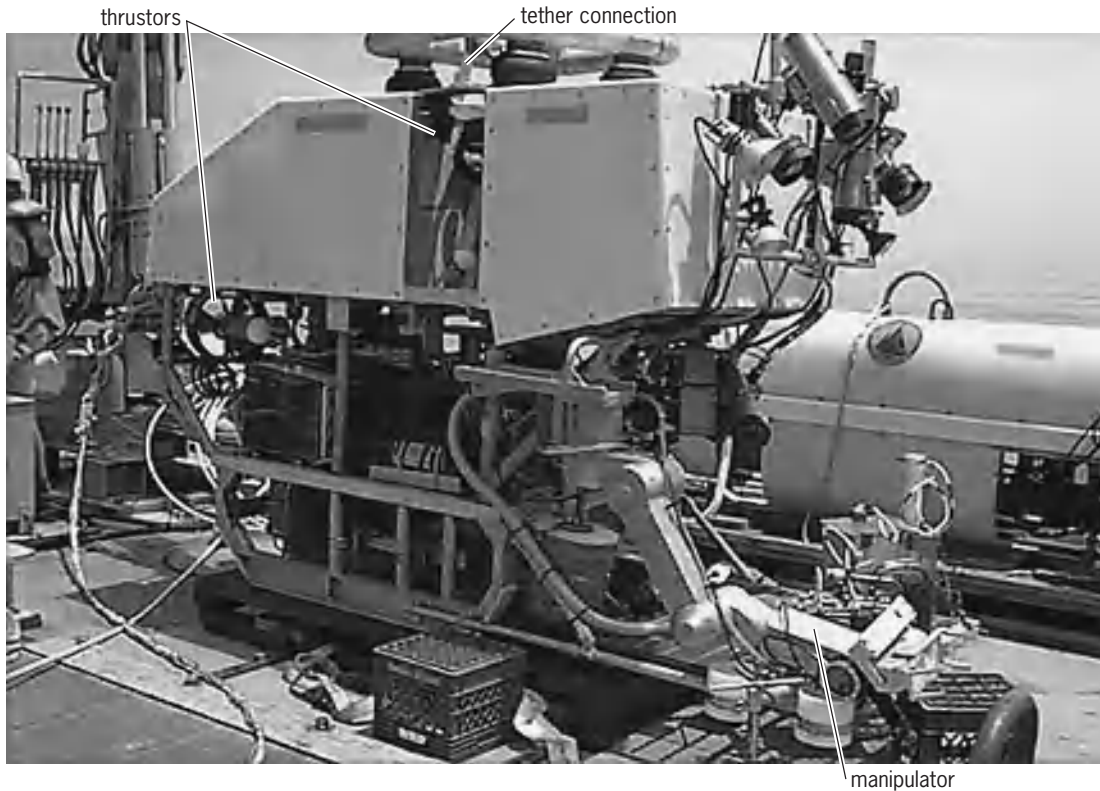
Fig. 3. *JASON* **remotely operated underwater vehicle. (***Woods Hole Oceanographic Institution***)**

torques being applied by the manipulator arm to the satellite. *See* SPACE SHUTTLE.

Crewless, remotely operated underwater vehicles are one of the fastest-growing applications of remote manipulators. These vehicles are used in oceanographic research and in the petroleum industry for inspecting and maintaining underwater well heads. The remotely operated underwater research vessel *JASON*, which is used in oceanographic research (**Fig. 3**), can operate at depths of 6000 m (20,000 ft). This vehicle has a manipulator that is remotely operated from a control station on a surface ship. The vehicle is connected to the surface ship by a tether that carries both command and feedback signals. *See* UNDERWATER VEHICLE.

Biomedical applications represent another rapidly growing area for remote manipulators. For example, abdominal surgery using minimally invasive (laparoscopic) procedures is increasingly common. A simplified representation of the procedure is shown in **Fig. 4**. The parallels between the general description of a remote manipulator (Fig. 1) and the surgical device are evident. Another biomedical application involving arthroscopic surgery is for human joint disease and injury, especially in professional sports. In this area, an endoscope, a device combining visual feedback, control, and remote manipulation, is used. This device typically consists of a fiber-optic bundle with additional tubes for moving fluids to or from the end point, and includes additional wires operating an end-point manipulation device to grip, incise, or cauterize diseased or injured tissue. Such endoscopic devices can perform surgery without large-scale tissue

damage. *See* FIBER-OPTICS IMAGING; MEDICAL IMAGING; OPTICAL FIBERS; SURGERY.

Aiding the disabled is an important use of remote manipulators. One example is devices that
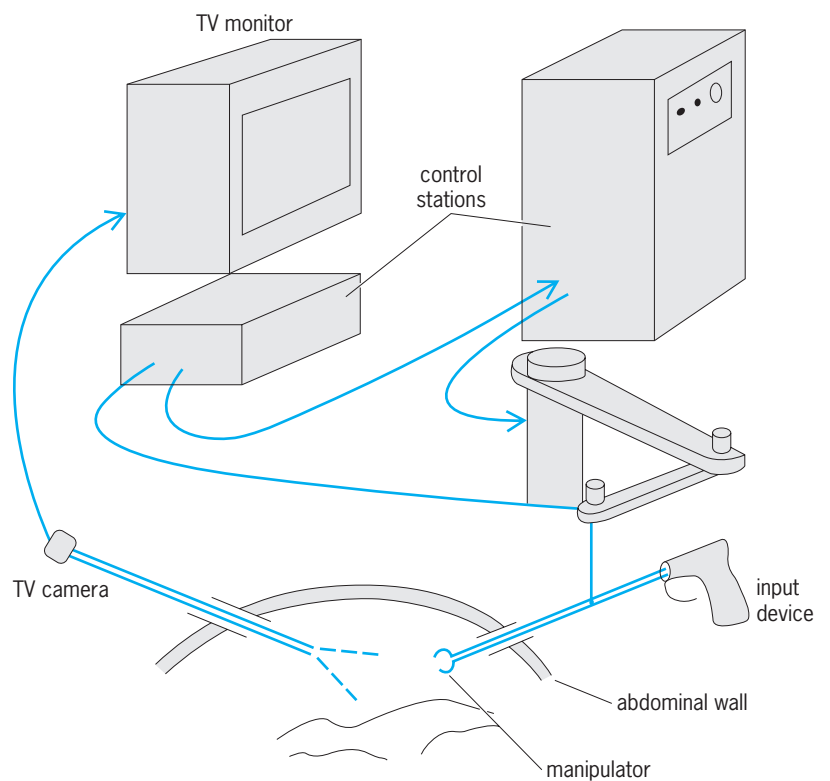
Fig. 4. Remote manipulation used in laparoscopic surgery.

allow individuals with little or no control of their upper extremities to feed themselves. Such devices have been referred to as teletheses (alluding to the extrasensory perception of distant objects). These devices are closely coupled to the user, acting as extensions of the human body. For example, an upper-body telethesis uses movements of the leg and foot as a control input. The manipulator itself is mechanical in nature, with an end effector (working component of a robot) consisting of a spoonlike utensil. Head-controlled feeding mechanisms have also been proposed for quadriplegics whereby head and neck movements will provide command inputs to the feeding manipulator. *See* ROBOTICS.

Devices that augment the strength of the human have been proposed for industrial applications. For example, a load-sharing manipulator has been envisioned in which human arm-manipulator coordination effectively allows the human to work with a "partner" that has considerably more strength. Another concept involves a "force extender." While not "remote" in the usual sense of the word, the manipulator is directed by the human and magnifies the forces supplied by the human operator. The human is thus able to move and manipulate masses considerably larger than those possible using muscle force alone.

Recreation of sensory input. The most crucial aspect of remote manipulation is the ability to recreate, with the appropriate realism, sensations that a human would perceive at a remote site—that is, to create a virtual reality so that the remote nature of the manipulating task is completely transparent to the operator. Significant progress is being made in this area. Appropriate re-creation of the physical cues (visual, aural, and tactile sensations) that the human would sense in the task is a major hurdle. If significant distance separates the remote manipulator from the human operator, delays can occur between the time that human actions are sent to the manipulator and the resulting manipulated-object changes are sensed and returned to the human. Delays of more than a few tenths of a second have been shown to destabilize a remotely manipulated task involving continuous human control inputs. Accommodating these delays is a difficult issue. *See* VIRTUAL REALITY.

Supervisory control. The necessity of continuous human inputs in the control of remote manipulators often results in a high degree of operator workload. Partial automation of some of these remote tasks is possible through the use of supervisory control by the human. Taking the remotely operated underwater vehicle as an example, supervisory control might be implemented by the human commanding the vehicle to move to a specified position near the ocean floor (perhaps at the head of an oil well). The vehicle's navigation to this position would be done without direct human intervention. More precise tasks, such as positioning the vehicle's manipulator, would then be performed by continuous human control. In such supervisory remote manipulation, the control

station must be more sophisticated than in the case of continuous manipulation, and must receive and operate upon feedback information from the remote location. *See* CONTROL SYSTEMS; HUMAN-MACHINE SYSTEMS.   Ronald A. Hess

Bibliography. N. Moray, W. R. Ferrell, and W. B. Rouse (eds.), *Robotics Control and Society*, 1990; *NASA/Air Force Annual Conferences on Manual Control*, 1964–1986; T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*, 1992; L. I. Slutski, *Remote Manipulation Systems*, 1998.

Remote sensing

The gathering and recording of information about terrain and ocean surfaces without actual contact with the object or area being investigated. Remote terrain sensing is part of the large subject of remote sensing, which deals with the gathering and recording of information on many types of natural phenomena from a distance (see **table**). Remote sensing uses the visual, infrared, and microwave portions of the electromagnetic spectrum (**Fig. 1**).

Humans have always used remote sensing in a primitive form. In ancient times, they climbed a tree or stood on a hill and looked and listened, or sniffed for odors borne on the wind. Although it is possible to sense the environment without the use of instruments, it is not possible to record the sensed information without the aid of instruments referred to as remote sensors. These instruments are a modern refinement of the art of reconnaissance, an early example being the first aerial photograph taken in 1858 from a balloon floating over Paris.

The eye is sensitive only to visible light, a very small portion of the electromagnetic spectrum (Fig. 1). Cameras and electrooptical sensors, operating like the eye, can sense and record in a slightly larger portion of the spectrum. For gathering invisible data, instruments operating in other regions of the spectrum are employed. Remote sensors include devices that are sensitive to force fields, such as gravity-gradient systems, and devices such as antennas that record the reflection or emission of electromagnetic energy. Both passive electromagnetic sensors (those that rely on natural sources of illumination, such as the Sun) and active ones (those that utilize an artificial source of illumination such as radar) are considered to be remote sensors. Several remote-sensing instruments and their applications are listed in the table. *See* COLOR VISION.

Terrain Characteristics

Each type of surface material (for example, soils, rocks, vegetation, and ocean waves) absorbs and reflects solar energy in a characteristic manner depending upon its atomic and molecular structure (Fig. 1c). In addition, a certain amount of internal energy is

Some areas of application for remote sensor instruments

Technique	Agriculture and forestry	Geology and planetology	Hydrology	Oceanography	Geography
Visual photography	Soil types Plant vigor Disease	Surface structure Surface features	Drainage patterns	Sea state Erosion Turbidity Hydrography	Cartography Land use Transportation Terrain and vegetation characteristics Thematic mapping
Multispectral imagery	Same as above	Lithological units Formation boundaries	Soil moisture	Sea color Biological productivity	Same as above
Infrared imagery and spectroscopy	Vegetation extent Soil moisture Plant conditions	Thermal anomalies Faults	Areas of cooling Soil moisture	Ocean currents Sea-ice type and extent Sea-surface temperature	Thermal activity in urban areas Land use
Radar: imagery, scatterometry, altimetry	Soil characteristics Plant conditions	Surface roughness Structural framework	Soil moisture Runoff slopes	Sea state Tsunami warning	Land-ice type and extent Cartography Geodesy
Passive microwave radiometry and imagery	Thermal state of terrain Soil moisture		Soil moisture Snow and ice extent	Sea state Sea-surface temperature Sea-ice extent	Snow and ice extent

emitted which is partially independent of the solar flux.

The absorbed, reflected, and emitted energy can be detected by remote sensing instruments in terms of characteristic spectral signatures and images. These signatures can usually be correlated with known rock, soil, crop, terrain, or ocean surface features. Chemical composition, surface irregularity, degree of consolidation, and moisture content are among the parameters known to affect the records obtained by electromagnetic remote sensing devices. Selection of the specific parts of the electromagnetic spectrum to be utilized in terrain or ocean sensing is governed largely by the photon energy, frequency, and atmospheric transmission characteristics of the spectrum (Fig. 1).

Types of Sensing

Because remote sensing is a composite term which includes many types of sensing, its meaning can be best understood by describing several of the types. Remote sensing is generally conducted by means of remote sensors installed in aircraft and satellites, and much of the following discussion refers to sensing from such platforms. *See* SCIENTIFIC SATELLITES.

Photography. Photography is probably the most useful remote sensing system because it has the greatest number of known applications, it has been developed to a high degree, and a great number of people are experienced in analyzing terrain photographs. Much of the experience gained over the years from photographs of the terrain taken from
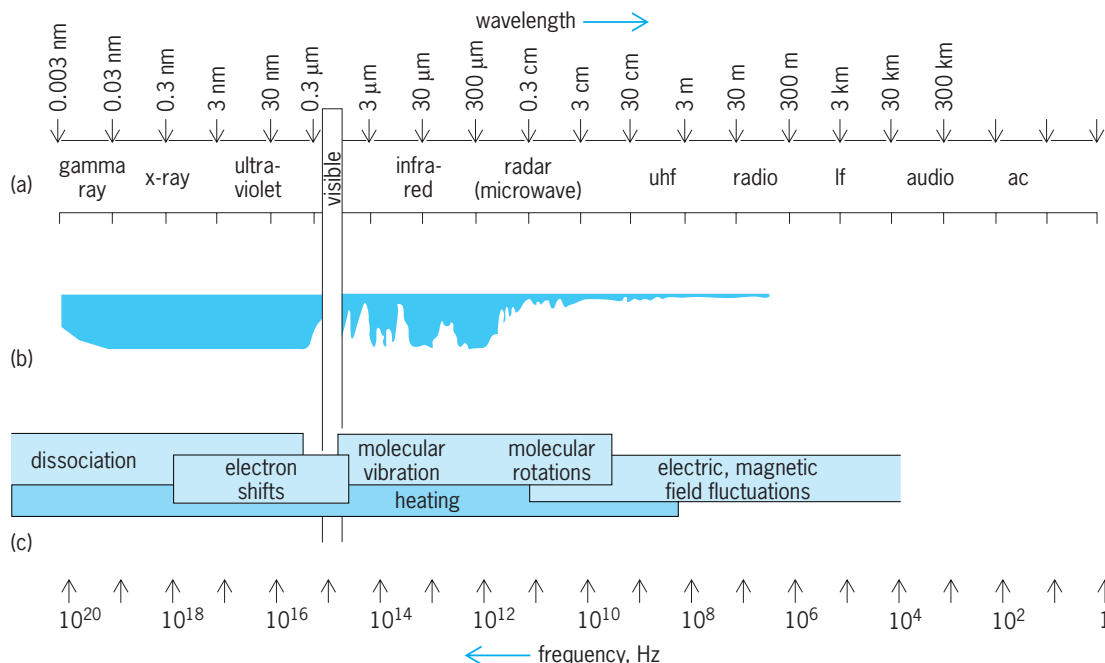
Fig. 1. Characteristics of the electromagnetic energy spectrum which are of significance in remote sensing. (*a*) Regions of electromagnetic spectrum. (*b*) Atmospheric transmission. The colored areas are regions through which electromagnetic energy is not transmitted. (*c*) Phenomena detected.

aircraft is being drawn upon for use in space. *See* PHOTOGRAPHY.

Conventional. Results of conventional photography experiments carried out in the short-duration crewed spacecraft such as a *Gemini*, *Apollo*, and *Skylab*, and in the long-duration uncrewed spacecraft such as *ERTS*, *Landsat*, *Nimbus*, *NOAA*, and *DMSP*, have vividly shown the applicability of these systems in space. These results indicate that space photography provides valuable data for delineating and identifying various terrain features such as those shown in **Fig. 2**.

The photomap in Fig. 2 was made by rectifying and fitting the photographs to the cultural plate of the existing line map of the area. The result combines all the original photographic detail with the cultural line data important for geographic orienta-
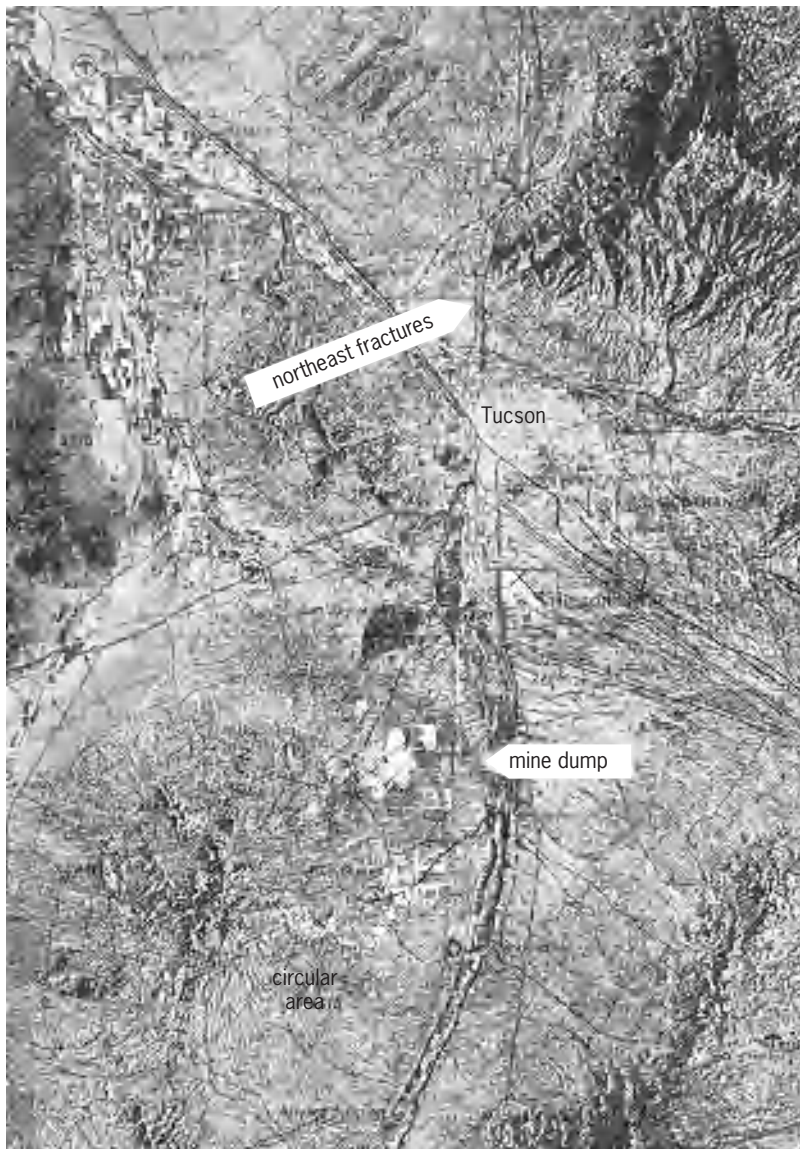

Fig. 2. Space photomap of the Tucson, Arizona, area, using 80-mm-focal-length Hasselblad cameras on the *Apollo 9* flight. The strong northeast-trending fractures in the mountains east of Tucson are apparent. The white patches on the northeast flank of the circular area are open pits and mine dumps, associated with already discovered mineral deposits.

tion and location. This combination provides considerably more terrain data than conventional topographic or raised relief maps. A number of major copper deposits occur in this region, and their location is in part controlled by the presence of intense fracturing, as shown on the photomap. Also important is the presence of intrusive rocks, which sometimes show up as circular areas. Fault intersections, intrusive centers, and alteration halos are important clues in the search for new mineral deposits of this copper type. Detailed study of such high-quality spacecraft imagery is very useful for unraveling the structural framework of potential mineral districts.

Multispectral. Multispectral photography can be defined as the isolation of the reflected energy from a surface in a number of given wavelength bands and the recording of each spectral band separately on film. This technique allows the scientist to select the significant bandwidths in which a given area of terrain displays maximum tonal contrast and, hence, increases the effective spectral resolution of the system over conventional black-and-white or color systems.

Because of its spectral selectivity capabilities, the multispectral approach provides a means of collecting a great amount of specific information. In addition, it has less sensitivity to temperature, humidity, and reproduction variables than conventional color photography and retains the high resolution associated with broadband black-and-white mapping film.

Multispectral imagery. An advance in multispectral sensing is the use of multispectral scanning systems which record the spectral reflectance by photoelectric means (rather than by photochemical means as in multispectral photography) simultaneously in several individual wavelengths within the visual and near-infrared portions of the electromagnetic spectrum (0.47–1.11 micrometers). Such instruments have been used in aircraft since the early 1960s and were used in the *Earth Resources Technology Satellite* (*ERTS 1*) and in *Landsat B*.

In the satellite cases, optical energy is sensed by an array of detectors simultaneously in four spectral bands from 0.47 to 1.1 μm. As the optical sensors for the various frequency bands sweep across the underlying terrain in a plane perpendicular to the flight direction of the satellite, they record energy from individual areas on the ground. The size of these individual areas (instantaneous fields of view) is determined by the resolution capability (spot size) of the optical scanner. The smallest individual area distinguished by the scanner is called a picture element or pixel, and a separate spectral reflectance is recorded in analog or digital form for each pixel. A pixel covers about 1 acre (approximately 0.4 hectare, or 4047 m^2) of the Earth's surface in the case of the *Landsat*, with approximately 7.5×10^6 pixels composing each *Landsat* image (115 × 115 mi or 185 × 185 km in area).

Such multispectral scanning systems have a number of advantages over conventional photography: the spectral reflectance values for each pixel can be transmitted electronically to ground receiving stations in near-real time, or stored on magnetic tape in

the satellite until it is over a receiving station. When the signal intensities are received on the ground, they can be reconstructed almost instantaneously into the virtual equivalent of conventional aerial photographs (provided the resolution size of the individual pixels is fine enough). In the case of conventional photography, the entire film cassette generally has to be returned physically to the ground, before the film can be developed and studied. Such physical return of film cassettes is not practical for long-duration satellites such as *ERTS* which are continuously recording vast amounts of data. *See* FACSIMILE.

Computer enhancement of imagery. There are a number of other advantages of multispectral scanning systems. Since they record reflected radiation energy in a number of discrete wavelengths, these radiance values can be used singly or combined by digital computer processing to provide a response optimized for particular terrain features.
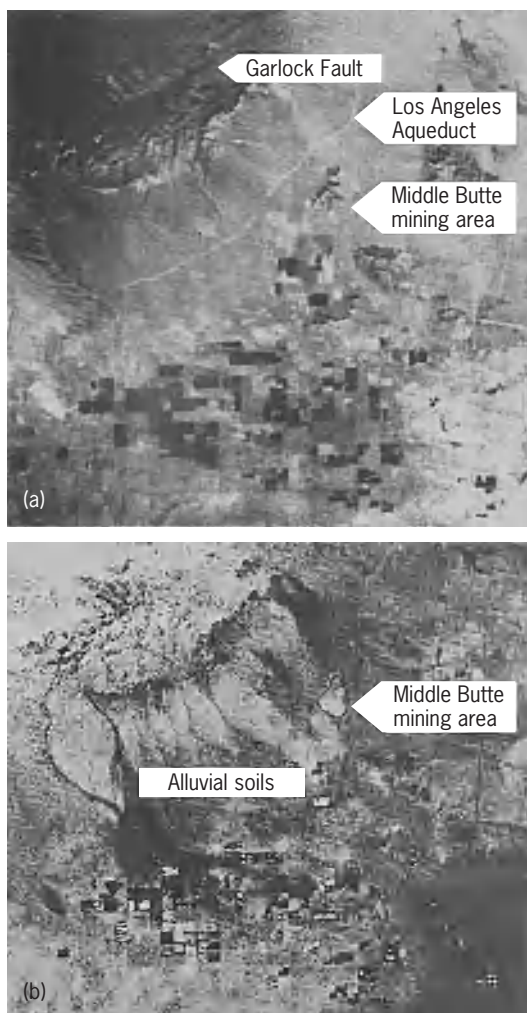
Fig. 3. Images taken by the earth resources technology satellite over the Mojave Desert of California on October 21, 1972: (*a*) band 4 (0. 5–0.6 μm); (*b*) a ratio of band 5 (0.6–0.7 μm) to band 4. The alluvial soils and Middle Butte mining area appear as bright areas on image *b*, while they are difficult to distinguish on image *a*. (*From P. M. Merifeld et al., Enhancement of Geological Features near Mojave, California by Spectral Band Ratioing of ERTS MSS Data, report prepared under U.S. Geological Survey Contract 14-08-0001-13911, 1974*)
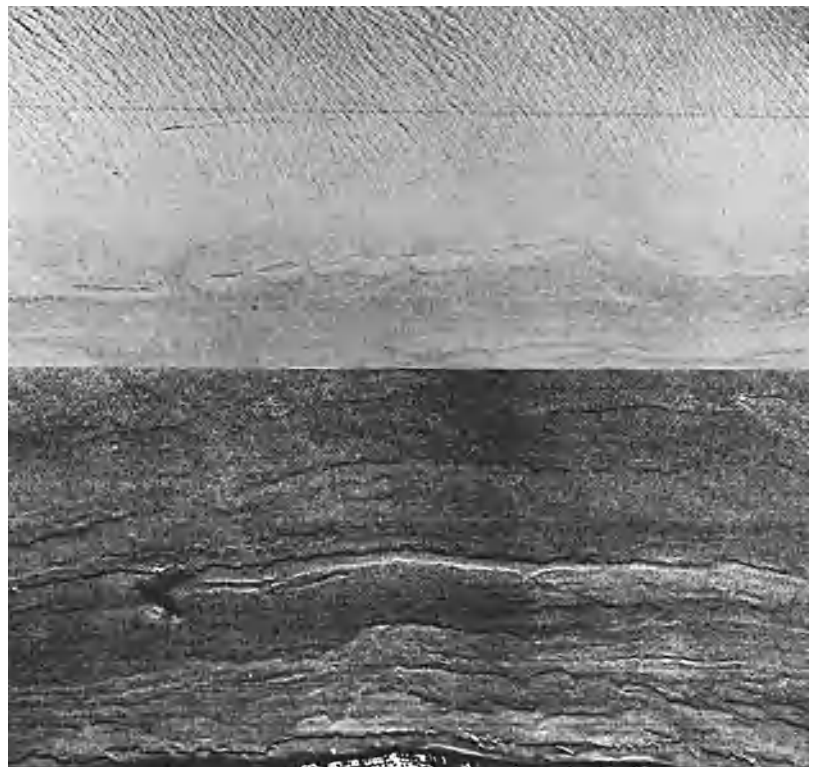
Fig. 4. Ratio image (0.50–0.54 μm)/(0.48–0.52 μm) from the Florida coast, acquired by an airborne multispectral scanner operating in the visible wavelengths. The reflection from the sea surface has been filtered out in the lower image, and the absorption effects of the water column have been removed by means of an algorithm. The residual reflections are from the sea floor and are indicative of surface roughness. (*Environmental Research Institute of Michigan, under contract to the U.S. Navy*)

One type of such computer enhancement which has proved to be very useful involves the preparation of ratio images. The individual spectral responses are ratioed, picture element by picture element using a computer, and they can also be contrast-stretched to further enhance the spectral differences. Such ratio images are very useful in detecting such items as subtle soil changes (**Fig. 3**) and hydrothermally altered areas (important for mineral exploration), as well as in discriminating most major rock types, and in detecting the nature of the sea floor (**Fig. 4**).

Another advantage of the multispectral scanner is its ability to extend the spectral coverage (beyond the 0.9-μm cutoff available with conventional photographic films) into the infrared wavelengths of the electromagnetic spectrum. *See* IMAGE PROCESSING.

Infrared. Infrared is electromagnetic radiation having wavelengths of 0.7 to about 1400 μm. All materials continuously emit infrared radiation as long as they have a temperature above absolute zero in the Kelvin scale. This radiation involves molecular vibrations as modified by crystal lattice motions of the material. The total amount and the wavelength distribution of the infrared radiation are dependent on two factors: the temperature of the material and its radiating efficiency (called emissivity). *See* EMISSIVITY; INFRARED RADIATION.

Thermal infrared radiation is mapped by means of infrared scanners similar to the multispectral
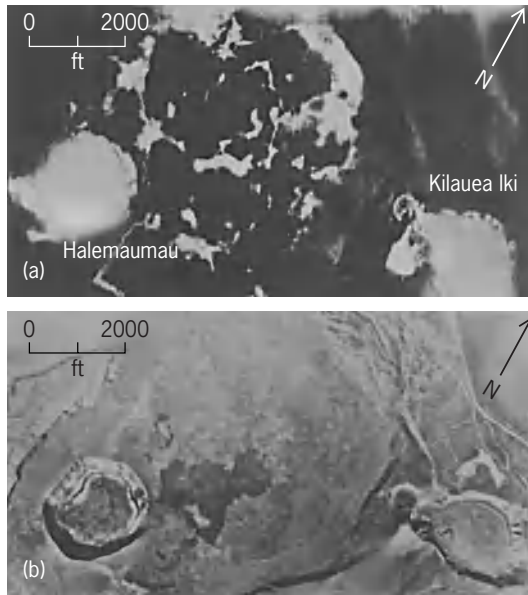
Fig. 5. Kilauea Volcano, Hawaii. (*a*) Infrared image. Tonal variation shows distribution of radiant heat; the brighter the tone, the warmer the surface. (*b*) Photograph by AT-11 mapping camera. Two vents left and right are within caldera. (*U.S. Geological Survey*)

scanners described previously, but in this case radiated energy is recorded generally in the 8–14-μm portion of the electromagnetic spectrum.

The imagery provided by an infrared scanning system gives information that is not available from ordinary photography or from multispectral scanners operating in the visual portion of the electromagnetic spectrum. The brightness with which an object appears on an infrared image depends on its radiant temperature. The hotter the object, the brighter it will appear on a positive image (**Fig. 5**). Radiant temperature is largely dependent upon chemistry, grain size, surface roughness, and thermal properties of the material. The time of day or night and the season of the year at which the infrared energy is recorded are also important factors, particularly when water or moisture exists in the near-surface terrain materials, since the water has a different thermal inertia relative to rock and soil materials and can therefore appear as a thermal anomaly. Because moisture frequently collects in geologic faults and fractures, infrared images are useful in detecting such features (**Fig. 6**).
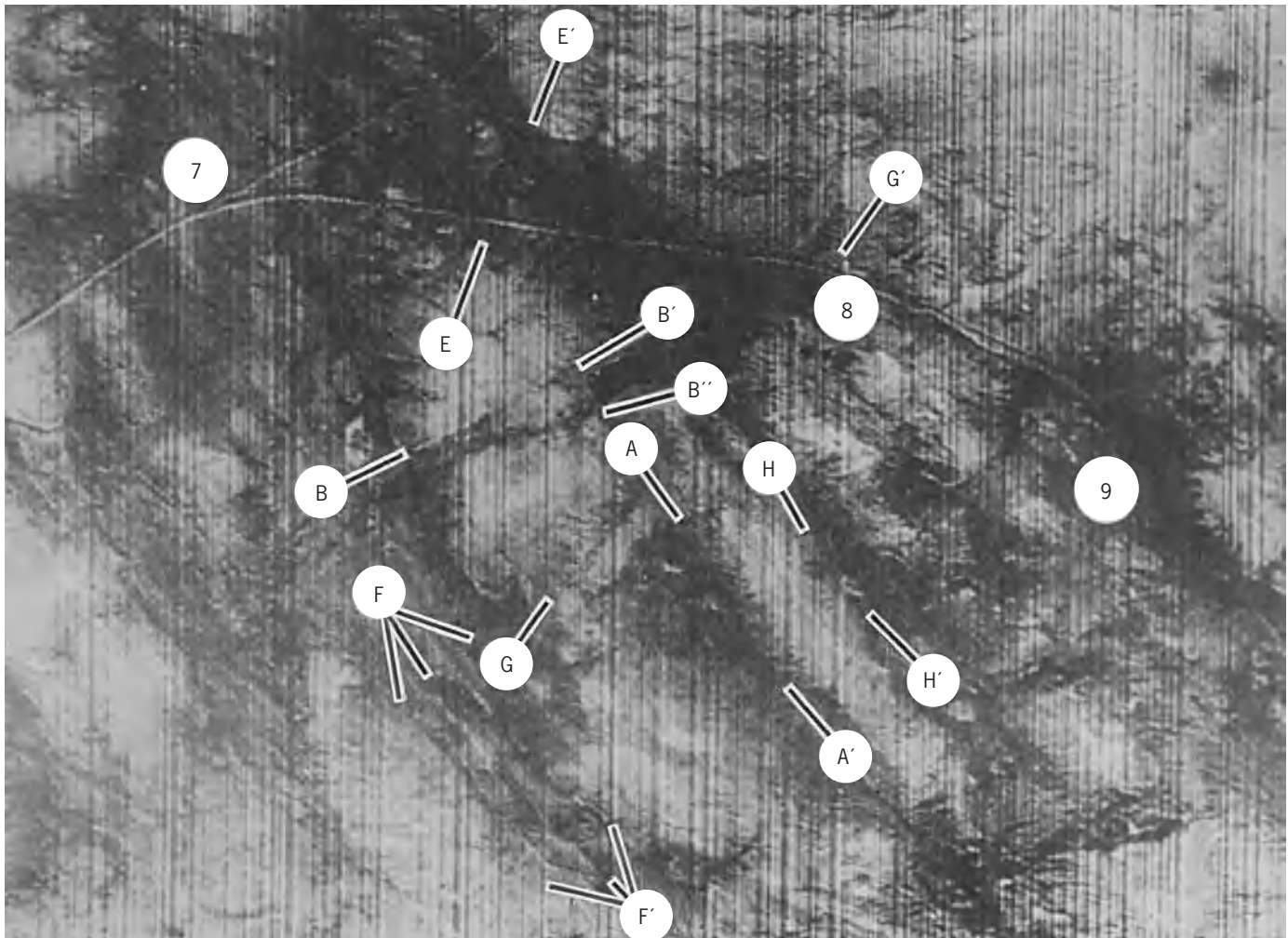
Fig. 6. Thermal infrared image of part of the Goldfields, Nevada, mineral district. Moisture is greater in the fracture zones of this district, which show up as dark bands mainly trending NW-SE (labeled FF′, AA′, HH′). A few less well-developed fractures trend ENE (labeled BB′ and BB″). In this district, commercial mineralization is best developed in fracture zones and in areas where hydrothermal alteration is most intense.

Although it is possible to obtain relatively high resolutions (spot sizes of a few feet) from infrared imagers in low-flying aircraft (Fig. 6), it has been more difficult to achieve such fine resolutions from satellite altitudes, partially because of the lower sensing-element sensitivity of infrared scanners (whose sensing elements must be cooled, in contrast to visible spectrum sensors), and partially because there is always a trade-off between field of view (generally large for satellite imagery as compared with aircraft imagery) and resolution. The larger the field of view, the more difficult it is to obtain fine resolution. Despite these restrictions, routine thermal mapping from long-duration satellites is now on the order of 2300–3000 ft (700–900 m) per picture element on the ground (*Nimbus 5* and *NOAA 2* satellites). The multispectral scanner on the short-duration *Skylab* space flight had an 260-ft (80-m) instantaneous field of view and recorded temperature differences of 0.40 K.

In the past, thermal infrared images were generally recorded on photographic film. Videotape records are replacing film as the primary recording medium and permit better imagery to be produced and greater versatility in interpretation of data.

Thermal mapping from satellite altitudes is proving to be useful for a number of purposes, one of which is the mapping of thermal currents in the ocean (**Fig. 7**). Such currents change their position quite frequently; in many cases these changes occur daily or even more frequently. The positions of such thermal boundaries are important for a variety of uses (fishery, naval, and so forth).

Thermal infrared mapping (thermography) from aircraft and satellite altitudes has many other uses also, including the mapping of volcanic activity (Fig. 5) and geothermal sites, location of groundwater discharge into surface and marine waters, and regional pollution monitoring.

Microwave radar. This type of remote sensing utilizes both active and passive sensors. The active sensors such as radar supply their own illumination and record the reflected energy. The passive microwave sensors record the natural radiation. A variety of sensor types are involved. These include imaging radars, radar scatterometers and altimeters, and over-the-horizon radar using large ground-based antenna arrays, as well as passive microwave radiometers and images. *See* MICROWAVE; RADAR.

One of the most significant advantages of these instruments is their all-weather capability, both day and night. The active radars also possess a certain amount of foliage-penetration ability which is valuable in jungle terrains (**Fig. 8**). This imagery illustrates the unique capability of side-looking radar airborne to penetrate both cloud cover and dense jungle vegetation and to map the structural fabric of the underlying terrain with considerable detail. It has been demonstrated that radar return amplitude is affected by the composition of the illuminated area, its moisture content, vegetation extent and type, surface roughness, and even temperature in certain circumstances.
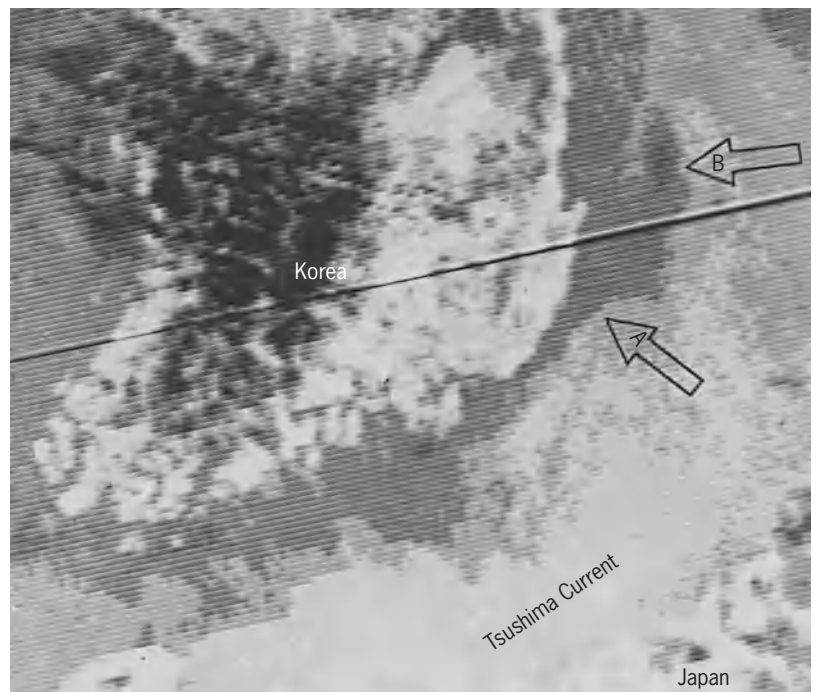
Radar returns are recorded in various forms to

Fig. 7. Thermal infrared image of the Korea Straits, acquired by the Defense Meteorological Satellite Program (DMSP) on October 13, 1972. DMSP infrared-sensor ground resolution is 2 nautical miles (4 km). The light-colored waters (warm) are part of the Tsushima Current, while the darker gray areas (cool) are the coastal shelf waters. An oceanic "front" forms the boundary between the continental shelf waters and the Tsushima Current. The front turns seaward at A, and a cyclonic eddy is located at B.

aid in their analysis. The forms having primary geoscience interest can be placed in one of the following categories: (1) scattering coefficients, which are a powerful tool in studying the nature of radar return and which have been correlated with different terrain types and directly applied to oceanic surface studies; (2) altimetry data; (3) penetration measurements, which have been utilized for mineral exploration, for example, detection of faults through moisture associated with them; and (4) radar images.

Of these techniques, imagery generally presents the optimum information content for terrestrial geologic purposes, partly because well-developed photographic interpretation techniques are applicable to radar image analysis. These images are especially valuable for delineating various structural phenomena, as shown in Fig. 8.

The image record of the terrain return is affected by the frequency, angle of incidence, and polarization of the radar signal. For example, if the terrain being imaged is covered by vegetation, a K-band (35 GHz) signal records the vegetation, whereas a P-band (0.4 GHz) signal is likely to penetrate the vegetation and thus record a combination of vegetation and soil surface. In general, each frequency band represents a potential source of unique data. The angle of incidence of the incident wave affects the image because of radar shadowing on the back side of protruding objects. The angle of incidence can also show differences because of changes in orientation of the many facets on a surface which affect the return strength. Although information is lost in the shadow region, the extent of the shadow indicates the height of the object and therefore has been

Fig. 8. Side-looking radar (SLAR) imagery of a portion of the jungle-covered Guayana Precambrian Shield of Venezuela. This imagery was acquired by a Goodyear synthetic-aperture X-band (3.12-cm) radar system mounted on board an Aero Service Caravelle aircraft. The flight altitude was 46,000 ft (14,030 m). The east-west dimension of the image mosaic is approximately 104 mi (168 km).

useful in emphasizing linear features such as faults and lineations reflecting joint systems.

Although imaging radars have been flown on aircraft for a number of years, it was not until 1978 that the first space-borne radar imagery became available with the launching of *Seasat* on June 28, 1978. This research satellite carried a radar altimeter, a radar scatterometer, a visible and infrared radiometer, and a scanning multichannel microwave radiometer, in addition to a synthetic-aperture radar (SAR) imaging in the L-band (1.275 GHz) with a swath width of 62 mi (100 km). *Seasat* failed in orbit on October 10, 1978, as a result of a massive short circuit in the elec- trical system. Fortunately, a considerable amount of data was recorded from several experimental test areas. **Figure** 9 is a mosaic of *Seasat* radar imagery taken over the Cape Cod, Massachusetts, area. Internal waves are readily apparent. These are waves with lengths of several hundred meters or more between crests, and they originate below the surface of the ocean due to internal slippage between water layers of different density. The amplitude of the waves is greatest at the density discontinuity. They are generally well developed in coastal waters. Such waves have been noted on a number of the *Seasat* SAR images. The brighter areas on the bottom of Fig. 9 are believed to be due to higher winds

in these areas, although no detailed analysis of this image supported by accompanying sea surface truth data has been published. Additional airborne and satellite flights are planned with SAR over controlled test sites because the optimum flight trajectories (azimuths), depression angles, wavelengths (X-, L-band, and so on), and polarizations (HH, HV, and so on) needed for extracting the full data content from the signals received are not fully understood. Although the analysis of oceanographic data returns from SAR are still in their infancy, it is already clear that SAR will yield very valuable information because of its high resolution and all-weather (cloud-penetrating) capability.

High-frequency (hf) radar. Such radars utilize frequencies in the 3-30-MHz portion of the electromagnetic spectrum (median wavelength of about 20 m) and are thus not within the microwave part of the spectrum. The energy is transmitted by ground-based antennas in either a sky-wave or surface-wave mode. In the sky-wave mode, the energy is refracted by the various ionospheric layers back down to the Earth's surface some 500–1800 mi (800–3000 km) (on a single-hop basis) away from the hf radar antenna site. The incident waves are reflected from such surface features as sea waves, and the reflections are enhanced when the wavelength of the

Fig. 9. Mosaic view of part of the Cape Cod, Massachusetts, area acquired by synthetic-aperture radar on board _Seasat_, 1978. The top of the figure is north. Nantucket Island is in the lower left portion of the figure. Internal waves are visible east of Cape Cod. The bright area in the lower part of the figure is believed to be roughened sea due to stronger wind action in this area. (_Johns Hopkins University Applied Physics Laboratory, under contract to the U.S. Navy_)

sea waves is twice the wavelength of the incident radar waves (Bragg reflection effect). In addition, the Doppler principle is applied to determine the component of sea-wave velocities moving toward or away from the radar. Spot sizes observed on the ocean surface by this sky-wave method are now as fine as 9 mi (15 km) in azimuth by 2 mi (3 km) in range, and it is expected that these resolution cells can be reduced by further development work.

In addition to this experimental use of hf radars for remote sensing of sea state, they are being used as

telemetry (communication) and direction and range-finding links in conjunction with disposable buoys or drogues. The buoys are used to detect the sea state, surface and subsurface temperature, salinity, and so forth. The information is relayed back to the interrogating site by either sky-wave or surface-wave modes. The surface waves are that portion of the hf wavefront that follows closely along the Earth's surface. Ranges of 60–180 mi (100–300) km or so are possible with this mode, and the buoys can be interrogated and located by low-cost direction-finding

loops and commercial battery-powered receivers. The big advantage of these ground-based hf radars is that they can maintain continuous monitoring of an ocean area, while satellite-borne radars can monitor only the areas that they happen to be flying over at the time. In the case of the *Landsat* system, the satellite comes over (revisits) the same area about every 18 days, while in the case of the *Defense Meteorological Satellite Program (DMSP)* craft the frequency of revisits can be up to several times daily, depending upon how many satellites are in orbit at one time. Neither the *Landsat* nor *DMSP* carry all-weather radar systems, however. *See* METEOROLOGICAL SATELLITES.   Peter C. Badgley

Bibliography. American Society of Photogrammetry, and Remote Sensing Staff and American Congress on Surveying and Mapping, *Remote Sensing*, vol. 1, 1987; E. A. Beaumont and N. H. Foster (eds.), *Remote Sensing*, 1992; A. Chedin (ed.), *Microwave Remote Sensing of the Earth System*, 1989; A. P. Cracknell and L. W. Hayes, *Introduction to Remote Sensing*, 1990; P. Curran, K. Kondratyev, and V. Kozogerov, *Remote Sensing of Soils and Vegetation*, 1990; G. P. De Loor (ed.), *Radar Remote Sensing of the Atmospheres and the Oceans*, 1984; S. A. Drury, *A Guide to Remote Sensing: Interpreting Images of the Earth*, 2d ed., 1998; G. M. Foody and P. J. Curran (eds.), *Environmental Remote Sensing Data Systems and Networks*, 1995; M. T. Halbouty, Geologic significance of Landsat data for 15 giant oil and gas fields, *Amer. Ass. Petrol. Geol.*, 64(1):8–36, 1980; P. M. Mather (ed.), *TERRA-2, Understanding the Terrestrial Environment: Remote Sensing Data Systems and Networks,*, 1990; R. M. Measures, *Laser Remote Sensing: Fundamentals and Applications*, 1984, reprint 1992; P. Pamploni, *Microwave Radiometry and Remote Sensing Applications*, 1989: J. Vernberg (ed.), *Processes in Marine Remote Sensing*, 1981.

Renewable resources

Agricultural materials used as feedstocks for industrial processes. For many centuries agricultural products were the main sources of raw material for the manufacturing of soap, paint, ink, lubricants, grease, paper, cloth, drugs, and a host of other nonfood products. During the early 1900s, the advances in organic synthesis in western Europe and the United States led to the use of coal as an alternative resource; in the 1940s, oil and natural gas were added as starting materials as a result of great advances in catalysis and polymer sciences. Since then the petrochemical industry has grown rapidly as the result of the abundance and low price of the starting materials as well as the development of new products, such as nylon and other polymers, water-based paints, and detergents. However, with the rapidly increasing economies of the nations of the world, these developments did not ever result in reduction in the utilization of agricultural products as industrial materials. Although specific nonfood crops, such as linseed or cotton, may have suffered from these developments, animal fats, vegetable oils, starch, cellulose, and other renewable agricultural resources still are used extensively in industry, and many new uses have been developed since the 1950s.

Price increases for oil and natural gas, as well as the recognition that many major industrial nations depend on imported agricultural raw materials, have resulted in an upsurge in research and development. Private companies have been disposed to reconsider domestic agricultural resources as alternative materials for manufacturing processes.

Fats and oils. Animal fats, marine and vegetable oils, and their fatty acid derivatives have always played a major role in the manufacturing of many industrial products. Some of these commodities are produced solely for industrial end uses; examples are linseed, tung, castor (not counting minor amounts used for medicinal purposes), and sperm whale oils. Others, such as tallow and soybean oil, are used for both edible and industrial products.

Most fats and oils are triglycerides; that is, they consist of a glycerine molecule esterified with three identical or different fatty acids. Some are used in their original state, with minor chemical modification or formulation. Examples are drying oils in paints (linseed, tung) and castor and rapeseed oils as lubricants. Hydrogenation of fats and oils leads to solid products for candles and waxes. Soybean and linseed oils are also epoxidized for use as stabilizers in poly(vinyl chloride), or made into alkyds for paints.

A large proportion of fats and oils are split to glycerine and fatty acids or converted into fatty methyl esters. They, in turn, are used directly or converted to fatty amines or fatty alcohols. Monounsaturated fatty acids, such as oleic or erucic acid, are split further by oxidation of the double bond to produce a mixture of mono- and dicarboxylic acids. Oleic acid is thus converted to pelargonic and azeleic acids, and erucic acid to pelargonic and brassylic acids. All of these products serve as starting materials for manufacture of soaps, detergents, lubricants, and plasticizers. The energy crises of the 1970s led to serious consideration of vegetable oils and their simple esters as emergency fuels to replace diesel fuel, especially for farming operations. *See* DETERGENT; FAT AND OIL; LUBRICANT; SOAP.

Carbohydrates. Starch, cellulosics, and gums also have been used for many centuries as industrial materials, whereas sugar crops, such as sugarcane and sugarbeet, have mainly satisfied world food needs.

Starch. Starches from potato, maize, and many other major crops have been used as thickening agents and adhesives for a variety of industrial applications and as substrates in fermentation and enzyme-mediated processes for the manufacturing of many products. Unmodified as well as derivatized starches play a major role in the production

of paper and textiles. The most prominent fermentation product from starch is industrial ethanol, used both as solvent and as fuel extender in gasoline (gasohol). Economic and technical developments have led to fermentation alcohol becoming cheaper than ethanol from petrochemical ethylene. Research since the 1940s has created new industrial products, including the so-called Super Slurper and related water-absorbing polymers (graft copolymers of starch and acrylonitrile), insoluble starch xanthate (used to clean up wastewater from the metal-plating industry), water-soluble laundry bags, biodegradable plastics, and encapsulating agents for pesticides. *See* ETHYL ALCOHOL; STARCH.

Cellulose. The main use of cellulose has been in fiber form for manufacturing paper and textile products, and the principal sources are hard- and softwoods. Other prominent sources are cotton and flax, as well as bagasse and straw from agricultural residues. Although petrochemical research has led to the production of many types and large quantities of synthetic fibers, manufacture of cellulose-based products has not decreased appreciably worldwide. Cellulose has also been modified chemically for manufacturing plastics (celluloid), fibers (rayon viscose), and thickening agents (carboxy methyl cellulose and so on). However, the development of synthetic plastics and fibers has drastically cut the use of cellulose for these purposes. *See* MANUFACTURED FIBER; PAPER; TEXTILE.

Since cellulose is so abundant in nature, much effort has been expended to explore the potential of this material for production of glucose as the starting material for fermentation processes. Although this is easy to accomplish with starch, so far no economical route for preparing glucose from cellulose has been developed. In addition to the cellulose-to-glucose conversion being more difficult, cellulose normally is closely associated with lignin, which in itself is a rather inaccessible polymer. However, once achieved, such conversion would revolutionize utilization of cellulose as well as of hemicellulose, a polymer of pentose sugars, and also a major constituent of plant materials. *See* CELLULOSE; HEMICELLULOSE; LIGNIN.

Gum. Plant gums, including guar, tragacanth, and gum arabic, have served as thickening agents for a long time and for many purposes. Most industrial nations depend on imports for these materials. They can now be supplemented or replaced by microbial polysaccharides, such as xanthan gum, which are made by fermentative conversion of starch. One prominent use of gums is in drilling fluids for secondary oil recovery. *See* GUM.

Hydrocarbons. Natural rubber and turpentine are excellent examples of plant-derived hydrocarbons. The development of synthetic rubbers during and after World War II has never threatened the demand for natural rubber; there is generally a world shortage. Turpentine is a product of the wood and paper pulp industry and is used as a solvent and thinner in paints and varnishes. *See* PINE TERPENE; RUBBER.

Novel crops. The threat that industrial nations might be separated from part or all of their traditional sources of raw materials through political and economic upheavals or natural calamities has resulted in a renewed effort to develop additional crops for local agriculture. In the United States, research has provided a number of candidate species that either are now in commercial development or are ready for the time when circumstances warrant such development. Examples are jojoba (liquid wax ester to replace sperm whale oil), guayule (alternate source of natural rubber), kenaf (paper fiber with annual yields much higher than available from trees), and crambe and meadowfoam (long-chain fatty acids, since erucic acid is no longer available from rapeseed oil). There is also active research involving *Cuphea* species (alternate source of lauric and other medium-chain fatty acids, to augment coconut oil), *Vernonia* (source of epoxy oil), and several other promising plants. For example, the Chinese tallow tree has the potential of producing 2.2 tons per acre (5 metric tons per hectare) of seed oil that could be used for manufacturing fuel and other chemicals.   L. H. Princen

Bibliography. K. R. Anthony, J. Meadley, and G. Robbelen (eds.), *New Crops for Temperate Regions*, 1993; K. T. Bird and P. H. Benson, *Seaweed Cultivation for Renewable Resources*, 1987; F. D. Gunstone, *The Chemistry Of Oils And Fats: Sources, Composition, Properties, and Uses*, 2004; D. L. Klass, *Biomass for Renewable Energy, Fuels, and Chemicals*, 1998; G. Pahl, *Biodiesel: Growing a New Energy Economy*, 2005; J. Tickell, K. Tickell, and K. Roman (eds.), *From the Fryer to the Fuel Tank: The Complete Guide to Using Vegetable Oil as an Alternative Fuel*, 2000.

Renner-Teller effect

The splitting, into two, of the potential function along the bending coordinate in degenerate electronic states of linear triatomic or polyatomic molecules. Most of the areas and methods of molecular physics and spectroscopy assume the validity of the Born-Oppenheimer approximation. The nuclei generally move much more slowly than the electrons, the frequencies associated with electronic transitions are much higher than vibrational frequencies, and one can consider separately the three types of molecular motion: electronic, vibrational, and rotational. These statements are no longer necessarily valid for electronic states which are degenerate or at least close to degeneracy, and the Born-Oppenheimer approximation breaks down.

Degenerate electronic states usually occur in molecules having a high degree of symmetry. The symmetric equilibrium geometry which causes the electronic degeneracy is, in general, lowered in the course of molecular vibrations, and this may lead

to splitting of the potential. The molecular potential is usually expressed in terms of a polynomial expansion in displacements r, and, in nonlinear molecules, the linear terms may lead to coupling of the electronic and vibrational degrees of freedom. The resulting breakdown of the Born-Oppenheimer approximation is in this case known as the Jahn-Teller effect. In linear molecules the symmetry is lowered during bending vibrations. In the bending potential the linear (and other odd) terms are zero by symmetry. The first nonvanishing terms which can couple the degenerate electronic states are quadratic in the bending coordinate. The results of this coupling in linear molecules are referred to as the Renner-Teller effect, or simply the Renner effect. *See* JAHN-TELLER EFFECT.

Splitting of electronic potential. In a linear tri-atomic molecule in a Π electronic state, the twofold degeneracy will be lifted whenever, in the course of the bending vibration, the nuclei assume a nonlinear configuration. As was first pointed out by G. Herzberg and E. Teller and shown in detail by R. Renner, the electronic potential function will split into two (see **illus.**). As noted above, the potential energy expansion in terms of displacement r along the bending coordinate can, for reasons of symmetry, contain only even terms, as in Eq. (1). In the

$$V = ar^2 + br^4 + \cdots \qquad (1)$$

presence of the vibronic coupling of the electronic states, the potential will be split into components usually denoted by V^+ and V^-, as in Eq. (2), where

$$V^{\pm} = \left(a \pm \frac{\alpha}{2}\right)r^2 + \left(b \pm \frac{\beta}{2}\right)r^4 \qquad (2)$$
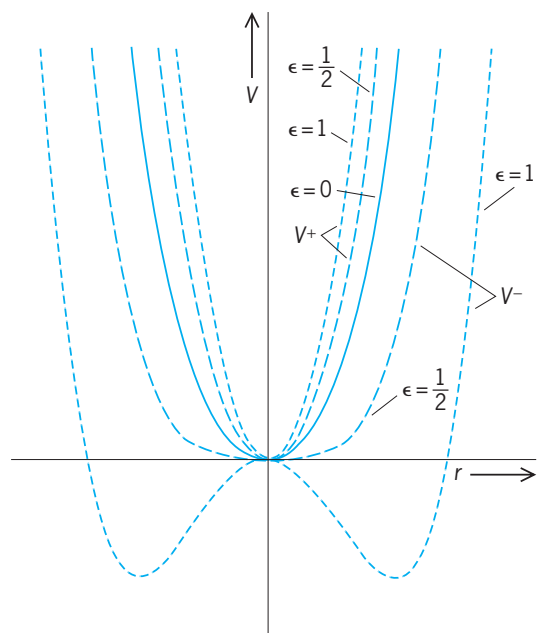
Bending potential function V for a degenerate electronic state of a linear molecule as a function of displacement r, and for three values (0, 0.5, 1.0) of the Renner parameter ϵ. In the absence of Renner-Teller splitting ($\epsilon = 0$) there is a single curve. When $\epsilon \neq 0$ the degenerate state splits into two components, V^+ and V^-.

α and β are the coupling terms. Often it is sufficient to consider only the quadratic coupling and neglect higher terms in the expansion.

Splitting of vibronic levels. The splitting of the electronic potential will, of course, affect the molecular vibrations and will result in splitting of the individual vibrational levels. Renner obtained approximate formulas for calculating the vibronic term values for a given level with vibrational quantum number v_2 and vibronic angular momentum K when the vibronic interaction is relatively small and the quartic terms negligible, Eqs. (3)–(5). Here ω_2 is the bending fre-

$$G^{\pm}(v_2, K) = \omega_2\sqrt{1 \pm \epsilon}(v_1 + 1) \qquad K = 0 \qquad (3)$$

$$G^{\pm}(v_2, K) = \omega_2[(v_2 + 1) - \tfrac{1}{8}\epsilon^2 K(K + 1)]$$
$$v_2 = K - 1, K \neq 0 \quad (4)$$

$$G^{\pm}(v_2, K) = \omega_2(1 - \tfrac{1}{8}\epsilon^2)(v_2 + 1)$$
$$\pm \tfrac{1}{2}\epsilon\sqrt{(v_2 + 1)^2 - K^2} \qquad v_2 > K - 1 \quad (5)$$

quency and the Renner parameter, $\epsilon = \alpha/2a$, is a measure of the strength of the vibronic interaction.

A well-studied example of Renner-Teller effect is the excited $A^1\Pi_u$ state of the C_3 radical. The observed vibronic levels conform relatively well to those calculated from Eqs. (3)–(5), even though the interaction is rather large in this case. When the vibronic interactions are very large, the vibronic levels will not be well described by Eqs. (3)–(5). In this case the energy levels of the molecule are best determined by setting up and diagonalizing the appropriate hamiltonian matrix.

Bent equilibrium configuration. It is easily seen that for $\epsilon > 0.5$ the lower potential V^- will possess potential minima at $r \neq 0$, corresponding to a bent configuration. (This is shown in the illustration for $\epsilon = 1$.) In extreme cases these minima in the lower potential may be so deep that a number of vibronic levels will lie below the central, linear local maximum. Under these circumstances it may be expedient to treat the lower component in terms of a nondegenerate electronic state of a bent molecule.

States of higher multiplicity. More complicated cases occur in Π and Δ states of higher multiplicity. In the presence of nonzero electron spin the effects of both vibronic and spin-orbit interactions must be considered. Particularly when both interactions are of comparable magnitude, their effects are not simply additive, but have to be treated simultaneously. Approximate formulas for treating these cases have also been developed. This situation is quite common, particularly in free radicals and molecular ions, and numerous examples of the Renner-Teller effect in $^2\Pi$ and $^3\Pi$ states have been studied. *See* MOLECULAR STRUCTURE AND SPECTRA.

Vladimir E. Bondybey; Terry A. Miller

Bibliography. G. Duxburg and R. N. Dixon, The Renner effect in a bent triatomic molecule using the adiabatic approach, *Mol. Phys.*, 43:225–274, 1981; G. Herzberg, *Electronic Spectra of Polyatomic Molecules*, 1966; G. Herzberg and E. Teller,

Schwingungstruktur der Elektronenübergänge bei mehr atomigen Molekülen, *Z. Phys. Chem.*, B21:410–466, 1933; Ch. Jungen and A. J. Merer, Orbital angular momentum in triatomic molecules, *Mol. Phys.*, 40:1–23, 1980; R. Renner, Zur Theorie der Wechselwirkung Zwischen Elektronen- und Kernbewegung bei drei atomigen, strabförmigen Molekülen, *Z. Phys.*, 92:172–193, 1934.

Rennin

The common name for chymosin, a proteolytic enzyme that is used to coagulate milk in cheesemaking. Rennin participates in the cheese-ripening process through its proteolytic activity.

Preparation. The traditional source of rennin is the fourth stomach (abomasum) of calves that have been fed only milk. The stomachs are dried or salted, cut into small pieces, and soaked in 10% salt brine to extract enzyme components from the stomach lining. The enzyme extract, commonly known as rennet, contains a mixture of around 80–95% rennin and 5–20% bovine pepsin, another proteolytic enzyme with milk-coagulating activity. Rennin also is produced commercially by using genetically engineered microorganisms. Recombinant deoxyribonucleic acid (DNA) techniques are used to clone and express the animal gene for rennin in the bacterium *Escherichia coli* and the yeast *Saccharomyces cerevisiae.*

Rennet may be used as a coagulant for cheesemaking or may be purified further by ion-exchange chromatography to yield pure rennin. Rennet or rennin is preserved with sodium benzoate and then standardized for milk-clotting activity.

Milk coagulation. Casein, the major milk protein, consists of three protein subunits, α-, β-, and k-caseins, arranged in a macromolecular structure known as a micelle. Casein micelles remain dispersed in milk because of the stabilizing action of k-casein. Milk coagulation occurs when casein micelles are destabilized by rennin. The process takes place in two distinct phases. Rennin catalyzes the first, or enzymatic, phase by cleaving one specific bond in k-casein. This results in the liberation of a glycomacropeptide and the loss of micelle stability. The second, or nonenzymatic, phase occurs after approximately 80% of total micellar k-casein has been cleaved by rennin. Casein micelles then agglomerate in the presence of ionic calcium, entrapping milk fat globules in the process. A three-dimensional spongelike matrix is formed that becomes the structural backbone of the resulting cheese curd. *See* CASEIN; MICELLE; MILK.

Cheese ripening. Approximately 6% of the rennin used to coagulate milk is retained in active form in cheese curd. During cheese ripening, rennin modifies the curd protein structure through its proteolytic action on α-casein, leading to textural changes described as a loss of curdiness. Casein peptides resulting from rennin action become precursors for flavor compounds in some cheeses such as Cheddar.

See CHEESE; ENZYME.   Paul Kindstedt

Bibliography. P. F. Fox (ed.), *Cheese: Chemistry, Physics and Microbiology*, vol. 1: *General Aspects*, 1987; P. Walstra and R. Jenness, *Dairy Chemistry and Physics*, 1984; N. P. Wong (ed.), *Fundamentals of Dairy Chemistry*, 3d ed., 1988.

Renormalization

A program in quantum field theory consisting of a set of rules for calculating S-matrix amplitudes which are free of ultraviolet (or short-distance) divergences, order by order in perturbative calculations in an expansion with respect to coupling constants. *See* SCATTERING MATRIX.

Divergences in quantum field theory. To describe the nature of the problem, it is useful to consider the simple example of a ϕ^4 theory defined by the lagrangian density $L(x)$ in Eq. (1) in four-dimensional

$$L(x) = \frac{1}{2}\partial_\mu\phi(x)\partial^\mu\phi(x)$$
$$- \frac{1}{2}m^2\phi(x)^2 - \frac{1}{4!}\lambda\phi(x)^4 \quad (1)$$

Minkowski space-time. Here, $\phi(x)$ is a quantum field operator depending on the four-vector x, ∂_μ and ∂^μ represent differentiation with respect to space-time coordinates, and m and λ are parameters. *See* LAGRANGIAN FUNCTION; RELATIVITY.

If one attempts to calculate any physical process in a perturbative expansion with respect to the coupling constant λ, in terms of Feynman diagrams, one has to confront divergent integrals. An example is the one-loop contribution to two-body scattering which is of order λ^2 (the second diagram in the **illustration**), and is proportional to the logarithmically divergent momentum integral given by expression (2),

$$\lambda^2 \int d^4k[(p_1 + k)^2 - m^2 + i\epsilon]^{-1}$$
$$\cdot [(p_2 - k)^2 - m^2 + i\epsilon]^{-1} \quad (2)$$

where p_1 and p_2 are the initial momenta of the two particles, and k is the loop momentum. *See* FEYNMAN DIAGRAM.

These ultraviolet (large-momentum) divergences have their origin in short-distance singularities occurring in the product of the quantum field operators (or their matrix elements) such as $\phi(x_1)\phi(x_2)$ as x_1 approaches x_2 [see expression (5)]. Therefore,
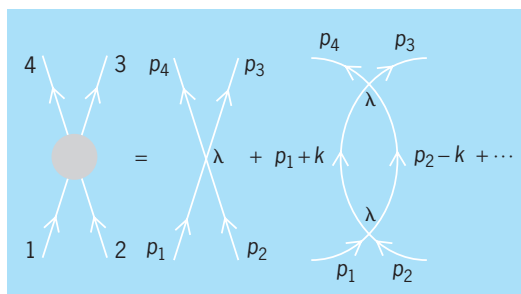
Two-body scattering in lowest-order perturbative expansion.

the quantum theory is not well defined since the lagrangian (or hamiltonian) contains these fields multiplied at the same point $x_1 = x_2 = x$. Renormalization is the procedure for constructing a well-defined finite quantum field theory, and deals with the proper definition of such singular terms in the lagrangian.

Regularization procedure. First the nature and kind of singularities must be identified, and then they should be removed to define the physical theory. The theory is first rendered finite by introducing a regularization parameter so that as it approaches a limiting value the divergences appear as definite singularities in this parameter. This process is called regularization. There are many methods of regularization. For example, in the cutoff method the integral in Eq. (2) will be rendered finite if it is cut off at $k^2 = \Lambda^2$, where Λ is finite. The result of integration, which is a function of Λ, will be seen to diverge as $\log \Lambda$ when $\Lambda \to \infty$. Another, more modern method is dimensional regularization: The integral is first performed in n dimensions rather than four. It will converge for $n < 4$ and will have a definite dependence on n. An example of this dependence is given by Eq. (3), where $\Gamma(\alpha)$ is the Euler

$$\int d^n k (-k^2 + c)^{-\alpha} = i\pi^{n/2} c^{(n/2)-\alpha} \frac{\Gamma[\alpha - (n/2)]}{\Gamma(\alpha)} \quad (3)$$

gamma function. All integrals, including the one in expression (2), $\Gamma(\alpha)$ is the Euler gamma function. All integrals, can be put in the form of Eq. (3) by using Feynman parameters to combine denominators. The answer is then analytically continued to complex values of n and the four-dimensional theory is defined as the limit of n approaching 4. In this way, as long as $n \neq 4$ the gamma functions in Eq. (3) are finite, and only finite well-defined quantities are manipulated. In this method the singularities reappear as poles of the form $(n - 4)^{-k}$ as n approaches 4. The behavior of the integral of expression (2) is given by expression (4). The remainder of this article is restricted to

$$\frac{2i\pi^2}{n-4} + (\text{finite}) + 0(n-4) \quad (4)$$

dimensional regularization. *See* GAMMA FUNCTION.

Renormalization of parameters. In a regularized quantum field theory, one can, in principle, calculate finite expressions for all the Feynman diagrams that contribute to any physical process, to any desired order in perturbation theory. The answer will be a well-defined function of the momenta of the external legs, the parameters of the theory, such as m and λ in Eq. (1), and the number of dimensions n, which is taken as a complex number. As n approaches 4, the singularities can be studied and the method for removing them can be given. The degree of divergence of a graph is determined by Weinberg's theorem. In a field theory such as ϕ^4 of Eq. (1), one finds only two types of Feynman graphs that diverge as $n \to 4$. Those are the graphs with two external legs (propagator) and four external legs (scattering amplitude) and any other graph that contains these two types of graphs as subgraphs. Therefore, to further

study the singularities it is sufficient to concentrate only on the two- and four-point functions of expression (5), whose Fourier transforms give the scatter-

$$\langle 0|T(\phi(x_1)\phi(x_2))|0\rangle$$

$$\langle 0|T(\phi(x_1)\phi(x_2)\phi(x_3)\phi(x_4))|0\rangle \quad (5)$$

ing amplitude. The singularities occur at short distances in the product of the field operators as points approach each other. This suggests that they can be removed from all physical processes by properly defining the local product of operators that appear in the lagrangian of Eq. (1). It turns out that indeed it is possible to absorb and remove all the infinities by a simple renormalization of the finite number of constants that define the theory. A theory is called nonrenormalizable when this procedure fails, and it can be rendered finite only by introducing an infinite number of renormalization parameters. *See* FOURIER SERIES AND TRANSFORMS.

Nonphysical nature of parameters. The possibility of renormalization by a redefinition of parameters such as masses and coupling constants that appear in the lagrangian hinges on the fact that these are not the physical quantities that would be observed as the prediction of the theory. For example, the physical finite coupling constant is defined as the strength of the interaction that an experimentalist will observe in a two-body scattering process with initial four momenta p_1 and p_2, and final four-momenta p_3 and p_4 (see illus.). The strength of the interaction is defined by the measured value of the physical probability amplitude, $A(p_1,p_2,p_4,\lambda,m)$, at an agreed value of the external momenta. This clearly is a complicated function of the parameter λ when the perturbation series is summed. The strength of the interaction is equal to λ only in lowest order and in general differs from it in the full theory. Similar remarks apply to the parameter m. The true mass of the particle in the theory differs from m. It is in general a complicated function of m and λ and is defined as the location of the pole in the full propagator. Similarly, the field $\phi(x)$ does not create the correctly normalized interacting particle when applied on the vacuum. It differs from the correctly normalized field ϕ_R by the multiplicative wave function renormalization constant $\sqrt{Z}$. Renormalization theory is the process of rewriting systematically the field theory in terms of the physical, finite, renormalized coupling constant λ_R, mass m_R, and field ϕ_R. If the theory is renormalizable and rewritten in terms of renormalized quantities, then all infinities cancel and all physical quantities, such as S-matrix elements, are finite and meaningful, as the regularization parameter is removed.

Power series in the renormalized parameters. An outline of this procedure can be given as follows. Since it has been agreed that λ, m and the renormalization of the field are not the observed physical quantities, they need not be finite. They are reexpressed as power series in terms of finite renormalized parameters λ_R and m_R and the regularization parameter. In the dimensional regularization scheme they take the form of

Eqs. (6), where the renormalized field ϕ_R is defined

$$\lambda = \mu^{4-n}\left(\lambda_R + \frac{a_1(\lambda_R)}{n-4} + \frac{a_2(\lambda_R)}{(n-4)^2} + \cdots\right)$$

$$m^2 = m_R^2\left(1 + \frac{b_1(\lambda_R)}{n-4} + \frac{b_2(\lambda_R)}{(n-4)^2} + \cdots\right) \quad (6)$$

$$Z = 1 + \frac{c_1(\lambda_R)}{n-4} + \frac{c_2(\lambda_R)}{(n-4)^2} + \cdots$$

as $\phi = \sqrt{Z}\phi_R$. Here $a_i(\lambda_R)$, $b_i(\lambda_R)$, $c_i(\lambda_R)$ are power series in the finite parameter λ_R; that is, $a_i(\lambda_R) = \Sigma a_{ik}\lambda_R^k$, and so forth, where the coefficients a_{ik}, b_{ik}, c_{ik} are to be determined so as to render the theory finite as described below. Here μ is a mass parameter called the renormalization point, whose role is better understood by studying the renormalization group (discussed below). All Green's functions or probability amplitudes such as the one in the illustration will be renormalized by setting the momenta at $p^2_i = \mu^2$. Furthermore, in Eq. (6), λ_R and m_R implicitly depend on μ, while λ and m are independent of it. *See* GREEN'S FUNCTION.

The perturbation series is now arranged as an expansion in the finite parameter λ_R and not in λ. This can be done by rewriting the lagrangian in terms of renormalized field ϕ_R and the parameters λ_R and m_R by substituting directly from Eqs. (6) into Eq. (1). Thus, the lagrangian takes the form of Eq. (7), where

$$L = \frac{1}{2}\partial_\mu\phi_R\partial^\mu\phi_R - \frac{1}{2}m_R^2\phi_R^2$$

$$-\frac{1}{4!}\mu^{4-n}\lambda_R\phi_R^4 + \Delta L_{\text{counterterm}} \quad (7)$$

$\Delta L_{\text{counterterm}}$, the "counterterm" piece, has the same functional dependence on ϕ_R as the explicitly written terms in Eq. (7) but with coefficients that correspond to the pole terms in Eq. (6). This piece would diverge as $n \to 4$ and is a power series in λ_R involving the coefficients a_{ik}, b_{ik}, c_{ik} mentioned above.

Minimal subtraction. These coefficients are determined in order to cancel all infinities by the following procedure: The amplitude $A(p_i,\lambda_R,m_R,\mu,n)$ and the propagator $\Delta_F(p_i,\lambda_R,m_R,\mu,n)$ are calculated to any desired order in λ_R in terms of Feynman graphs [these are related to the Green's functions in expression (5)]. To these there are contributions from the explicit part of the lagrangian in Eq. (7) and also from the counterterms which involve the coefficients a_{ik}, b_{ik}, and c_{ik} up to the desired order in λ_R. Then, order by order, one requires that the poles $(n-4)^{-i}$ must completely cancel in the three measurable quantities given in Eqs. (8). This is called the minimal subtrac-

$$A(p_i, \lambda_R, m_R, \mu, n)|_{p_2=\mu_2} = \text{finite as } n \to 4$$

$$\Delta_F^{-1}(p, \lambda_R, m_R, \mu, n)|_{p_2=\mu_2} = \text{finite as } n \to 4$$

$$\frac{\partial^2}{\partial p^2}\Delta_F^{-1}(p, \lambda_R, m_R, \mu, n)|_{p_2=\mu_2} = \text{finite as } n \to 4 \quad (8)$$

tion scheme, and it completely fixes the coefficients a_{ik}, b_{ik}, and c_{ik}. There are other subtraction schemes

which require the above quantities to be equal to a fixed observed value rather than being simply finite as $n \to 4$. The observable quantities such as the S-matrix elements are independent of the particular subtraction scheme adopted.

Finite S-matrix. As mentioned above, the only "primitive" infinities in the ϕ^4 theory appear in the Green's functions of expression (5) containing two or four external legs and in those graphs containing them as subgraphs. With the above renormalization procedure (choosing a_{ik}, b_{ik}, c_{ik}) all such infinities in these two- and four-point functions have been arranged to cancel. It remains to be shown that as $n \to 4$ (removal of cutoff) the S-matrix is finite. This is done by mathematical induction: Assume that it is true to order m in perturbation theory, then prove it to order $m + 1$. This is carried out successfully if the theory is renormalizable. Therefore, all infinities can be removed by a proper definition of the products of the field operators appearing in the lagrangian. This amounts to properly identifying the observable finite quantities which led to a renormalization of the field, the coupling constant and the mass as in Eqs. (6). Any other physical quantity can now be calculated and will be a function of only the renormalized finite parameters λ_R and m_R, in addition to momenta.

Examples of renormalizable fields. So far the only field theories known to be renormalizable in four dimensions are those which include spin-0, spin-$^1/_2$ and spin-1 fields such that no term in the lagrangian exceeds operator dimension 4. The operator dimension of any term is calculated by assigning dimension 1 to bosons and derivatives ∂_μ, and dimension $^3/_2$ to fermions. Spin-1 fields are allowed only if they correspond to the massless gauge potentials of a locally gauge-invariant Yang-Mills-type theory associated with any compact Lie group. The gauge invariance can remain exact or can be allowed to break via spontaneous breakdown without spoiling the renormalizability of the theory. In the latter case the spin-1 field develops a mass. The successful quantum chromodynamics theory describing the strong forces and the SU(2) × U(1) Weinberg-Salam-Glashow gauge model of unified electroweak particle interactions are such renormalizable gauge models containing spin 0, $^1/_2$, and 1 fields. The renormalization procedure in a gauge theory is much more complicated than in the simple ϕ^4 theory because of gauge fixing and lack of either manifest unitarity or Lorentz invariance, but gauge theories have been shown to be renormalizable. *See* ELECTROWEAK INTERACTION; FUNDAMENTAL INTERACTIONS; GAUGE THEORY; LIE GROUP; QUANTUM CHROMODYNAMICS; QUANTUM ELECTRODYNAMICS; WEAK NUCLEAR INTERACTIONS.

Renormalization group. An important topic in renormalization theory is the renormalization group. This is the study of the dependence of the theory on the renormalization point μ that appeared in the subtraction procedure of Eqs. (6) and (8). Observable quantities such as the S-matrix elements, the measured strength of the interaction in the illustration, and so forth, do not depend on μ, but the finite

expansion parameter λ_R or the mass parameter m_R do. All computations depend on the lagrangian in Eq. (1) which has no information about μ. Changing the value of the renormalization point in Eq. (6) induces new values of λ_R and m_R in such a way as to keep the measurable S-matrix elements unchanged. Since changing the mass parameter μ can be viewed as a change of scales, the renormalization group is intimately connected to scale transformations in the theory. It makes it possible to study the high- or low-energy behavior of the field theory by applying scale transformations to the momenta of the particles involved in a particular reaction. It is then found that the value of the effective expansion parameter $\lambda_R(\mu)$ of the theory depends on the energy scales of the reaction under study. It becomes convenient to choose the value of μ such that $\lambda_R(\mu)$ is small enough for a valid perturbative expansion. For certain gauge theories the effective coupling constant, which is a measure of the interaction, decreases as the energy scale increases. This behavior is called asymptotic freedom, and it explains why quarks act as free particles (small effective coupling constant) when they collide at very high energies and come to within very short distances of each other. *See* QUARKS; SYMMETRY LAWS (PHYSICS). Itzhak Bars

Effective field theory. Effective field theory is a general and powerful method for analyzing quantum field theories over a wide range of length scales. Together with a closely related idea, the Wilson renormalization group, it places renormalization theory on a more general, physical, and rigorous basis. This method is most naturally developed in the Feynman path integral formulation of quantum field theory, where amplitudes are given by an integral over all histories. Each history is weighted by a phase equal to the classical action divided by Planck's constant. *See* ACTION.

An effective field theory is a quantum field theory which describes physics only at distances greater than some small length ℓ. The integral over histories is cut off, wavelengths less than ℓ being omitted. The physics on scales shorter than ℓ is parametrized by taking a completely general effective action. Factors of ℓ are inserted into the action so as to make the constants in the lagrangian dimensionless (in units where the speed of light and Planck's constant are equal to 1). The action has a series of terms that contain higher and higher powers of ℓ.

There are two evident differences between an effective field theory and the older idea of a renormalizable field theory: The integral over histories is cut off at a length ℓ, and the action contains an infinite number of terms. Because of the cutoff, ultraviolet divergences are no longer present, and the renormalization theorem takes a more physical form. That is, amplitudes at distances much greater than ℓ depend on the unknown physics at distances below ℓ only through a finite number of parameters. This is very plausible from dimensional analysis, because all but the first few terms in the lagrangian contain positive powers of the small length ℓ. These few terms are precisely the ones retained in renormalizable field

theory. More generally, there may be measurable effects proportional to ℓ or ℓ^2, or even higher powers. These are again parametrized by a finite number of constants. In applications to statistical mechanics, a similar expression for the free energy (instead of the action) is used. *See* FEYNMAN INTEGRAL.

Wilson renormalization group. It is useful to organize the path integral by scales, integrating first wavelengths between the cutoff ℓ and a slightly greater scale ℓ^1. The result is an effective field theory with the new cutoff ℓ^1 and a new action, which now includes the effects of physics on scales between ℓ and ℓ^1. The action is thus a function of the cutoff scale, and the process of integrating out a narrow band of wavelengths gives rise to a flow, a differential equation for the couplings. This is the Wilson renormalization group equation, which has broad applications in particle physics, condensed-matter physics, and statistical mechanics.

A point where all the parameters in the equation vanish is a fixed point of the flow; if the flow at this point is inward, it is termed an infrared fixed point. The physics at very long distances will be governed by an infrared fixed point, so such points define the possible phases of the theory. This is the basis of K. Wilson's theory of critical phenomena. *See* CRITICAL PHENOMENA; QUANTUM FIELD THEORY.
Joseph Polchinski

Bibliography. J. C. Collins, *Renormalization*, 1986; R. J. Creswick, H. A. Farach, and C. M. Poole, *Introduction to Renormalization Group Methods in Physics*, 1991; C. Itzykson and J. B. Zuber, *Quantum Field Theory*, 1980; W. Marciano and H. Pagels, *Phys. Rep.*, C36:137, 1978; K. Wilson, Problems in physics with many scales of length, *Sci. Amer.* 241:(2)158–179, August 1979; K. Wilson, The renormalization group and critical phenomena, *Rev. Mod. Phys.*, 55:583–600, 1983.

Reproductive behavior

Behavior related to the production of offspring; it includes such patterns as the establishment of mating systems, courtship, sexual behavior, parturition, and the care of young. Successful reproductive efforts require the establishment of a situation favorable for reproduction, often require behavior leading to the union of male and female gametes, and often require behavior that facilitates or ensures the survival and development of the young; the mere union of gametes is not generally sufficient for successful reproduction. For each species, there is a complex set of behavioral adaptations that coordinate the timing and patterning of reproductive activity. Typically, this entails integration of both overt behavioral and internal physiological events in both male and female, all of which are intricately enmeshed in manners adapted to the environment in which the animals live. The behavioral patterns related to reproduction tend to be relatively stereotyped within a species, but diverse among different species—especially

distantly related species. The end products of cycles of reproductive activity are viable, fertile offspring which, in turn, will reproduce and thus perpetuate the species.

Underlying the integrated and cooperative behavior leading to the perpetuation of the species are the interests of individual animals. Natural selection is most often seen as working at the level of the individual; those individuals that will be favored by natural selection are those that are able to survive and reproduce effectively. If there is genetic variation within a population, it is the genes that relate to successful reproduction by individuals that will be perpetuated and come to be best represented in future generations. The concept of fitness relates to the relative contribution to the genetic structure of future populations by individual organisms. Thus, what appears to drive the evolution of integrated and cooperative reproductive systems is the action of natural selection favoring those individuals behaving in ways that maximize their reproductive success. *See* ORGANIC EVOLUTION.

The emphasis in studies of reproductive behavior has been upon sexual reproduction—that form of reproduction that involves substantial genetic recombination. It should be remembered, however, that many species reproduce asexually, via autogamy, budding, parthenogenesis, or similar processes. The essence of asexual reproduction is that the offspring generally bear genotypes that are close, if not exact, copies of those of their parents. Reproductive behavior may nevertheless be quite complex. For example, in the parthenogenically reproducing lizards of the genus *Cnemidophorus*, females court, mount, and ride each other in a manner remarkably similar to that of closely related species that reproduce sexually. *See* REPRODUCTION (ANIMAL).

Each successful episode of reproductive activity takes place in a given context, begins with some preliminary events, includes the actual mating, has important sequelae, and leads to parental behavior. The events encompassing each aspect of the sequence must be described before they can be analyzed with respect to underlying mechanisms, development in the individual, evolutionary history, and adaptive significance.

Context

The context of reproductive behavior includes such factors as the season of breeding, the mating system, and social factors such as territoriality and dominance.

Seasonality. Seasonal breeding is the rule in many species—especially those living in temperate zones. The reason appears related to the seasonal availability of resources for the developing young. Species tend to bear their young at the time that will be most favorable, often in the late spring or early summer. Because gestation periods are so variable, the time of breeding tends to vary greatly among species. For example, many species of deer and elk have a fall rut, coyotes and wolves often breed in winter, and many seals and sea lions breed in late spring.

Along with reproductive behavior, a whole complex of behavioral and physiological changes occurs seasonally. Such changes are most dramatic in migratory species, such as the short-tailed shearwater, a bird that arrives on islands near Tasmania and lays most of its eggs within a period of a few days in November. Some annual cycles, such as those of ground squirrels, are driven by endogenous processes and thus show a cycle of about a year even when the animal is deprived of any external cues relating to season.

Cycles often are tied to external stimuli. Annual breeding cycles can be tied to such stimuli as rainfall, temperature, plant growth, or day length. Because it tends to be the most reliable cue, daylength is used by a large number of species. Syrian golden hamsters, for example, show a regression of their gonads in the fall when there are 12.5 h or less of light per day. However, they recrudesce about 4 or 5 months later, regardless of the photoperiod.

The roles of the pineal gland, melatonin, and the hypothalamic-pituitary-gonadal axis in the control of annual cycles in hamsters have been delineated. In canaries, parts of the brain related to singing are sexually dimorphic (larger in males than in females), but these too vary seasonally and are considerably smaller in the fall than in the spring.

Field studies of hormones in sparrows have shown that seasonal changes in male territorial behavior are tightly linked to changes in testosterone levels induced by social and environmental stimuli. Changes in male copulatory behavior are less testosterone-dependent, being triggered by female solicitation displays.

In sum, as the result of the action of natural selection, largely on the timing of births, whole complexes of behavioral, physiological, and morphological adaptations are integrated into seasonal patterns.

Mating systems. The relationships between individual males and females and the degree of exclusivity in mating are part of the mating system of a species or population. There are three basic mating system types: monogamy, polygamy, and promiscuity.

Monogamy. In this mating system the reproductive unit is generally a single male and a single female, the partners copulate only with each other, there may be shared parental care, and there is some kind of prolonged pair bond. Monogamy is very common among birds (it has been estimated that 90% of avian species are monogamous). There are many species, however, that fall in the gray areas between a monogamous system and some other form. In some species, there is a prolonged pair bond and shared parental care, but occasional matings outside the pair. In others, there may be mating exclusivity but no prolonged bond. Animal behaviorists disagree as to which of these characteristics should be defining attributes of a monogamous system.

Polygamy. In polygamous mating systems, there is again a prolonged association, but more than two individuals are involved in the relationship. When there is prolonged association and an essentially

Fig. 1. Sage grouse lek. Males stake out and defend territories on a traditional breeding ground. Females breed with males in selected territories. (*Courtesy of R. Haven Wiley*)

exclusive mating relationship between one male and two or more females, the system is termed polygynous. For example, bats of the species *Phyllostomus hastatus* roost in cave colonies, in clusters that consist of a single adult male and a harem of about 18 females. The harem male actively defends the cluster from other males, many of which live in bachelor groups, and he sires all or most of the offspring born to the harem females.

If there is a prolonged association but it is between a single female and two or more males, the system is termed polyandrous. In the American jaçana, a pond-dwelling bird, a female defends a large territory that encompasses the territories of several males with which she associates. Females are larger and more aggressive than males and lay clutches of eggs in nests in each of the males' territories. Males incubate the eggs.

Promiscuity. In this mating system, there is no prolonged bond formed, and there are multiple matings by members of at least one of the sexes. Some species of grouse, bears, and wildebeest appear to be promiscuous.

Territoriality and dominance. Territoriality or dominance occurs in many kinds of mating systems. A territory is an area that is defended against conspecific animals (those of the same species). It may be occupied by a single individual, a bonded male-female pair, or a larger group. If an intruder and a resident engage in an aggressive interaction in a territory, it is the resident that generally wins. Interestingly, if the same two animals are in conflict in the other individual's territory, the tables are turned, and the one that is now within its own territory generally wins even though it lost to the first animal when off its own territory. An example is provided by male speckled wood butterflies, which compete for spots of sunlight on the ground that serve as mating territories. Resident males almost always win encounters in their territories. *See* TERRITORIALITY.

The resident of a territory generally has privileged access to the resources on that territory. Where the territory is relatively large, as in many diurnal songbirds, it may include sufficient resources to support a bonded pair and their offspring. By contrast, in many colonially nesting marine birds the territories may encompass little more than a nest site, while food and other resources are collected at a distance. In a special form of territoriality, a lek, males of some species, such as sage grouse, defend small territories that are used only for breeding (**Fig. 1**). After male territories have been staked out, females approach the lek and breed with some of the males. The females then depart; males play no further role in reproduction but continue contests for territories and matings with additional females.

Whereas in a territorial system the outcome of a contest is generally predictable given only the location of the encounter, in a dominance relationship an individual wins regardless of location. In troops of various species of primates, for example, there may be a single dominant male and a hierarchy of males ranking below him with each dominant to males of lower rank. Such relationships can be quite stable over time. Often the dominant males have privileged access to resources, such as food and mating partners, but the evidence for such effects needs expansion. There are, of course, many varieties of dominance relationships, with male hierarchies, female hierarchies, mixed-sex hierarchies, and triangular relationships that are departures from linearity. Dominance-related contests may occur seasonally, generally peaking in intensity during the breeding season, and can be of great importance in determining which individuals reproduce.

The dividing line between dominance and territoriality is not always clear, and both can occur within a species. In some species of lizards, for example, several males may defend territories against each other, but with one dominant male that can move about the entire area. With a system based on dominions, there are zones of increasing and decreasing dominance. There is not overt defense of borders, but individuals nearest their core area generally secure access to resources.

Not all mating systems are based on dominance or territoriality. For example, scramble competition polygyny is common in the thirteen-lined ground squirrels (*Spermophilus tridecemlineatus*). Males compete in locating and mating with females but do not display dominance or territoriality; competitive mate searching is favored.

Preliminary Events

Much of reproductive behavior takes place prior to mating. Mate choice and then courtship are both essential to successful reproduction in animals.

Mate choice. Pairs often appear to result from an active choice of partners. In *The Descent of Man and Selection in Relation to Sex*, Charles Darwin differentiated sexual selection from natural selection. Sexual selection is related to the differential ability of individuals to secure mating partners, and is now generally agreed to be a subcategory of natural selection. Darwin noted that males generally compete among themselves for access to mating partners. He interpreted such structures as the elaborate plumage of some birds and mammalian horns and

antlers as functioning in such a context. Females, however, were viewed as rarely lacking mating partners. Rather, the problem for females was taken to be selection of a partner or partners from among several males. Thus, there developed a notion of the active and ardent male coupled with the coy and choosy female. Such differences are often related to the relative availability of potential mating partners for males and females and to the amount of reproductive investment characteristic of each sex. As will be discussed below, investments by females are generally regarded as larger than those by males, and hence females may be a limiting resource for the reproductive activity of males. There are some species, however, such as the American jaçana, where males appear to invest more and the relationship is reversed.

Female choice. The importance of female choice remains controversial. Ultimately, females can benefit by mating with males that are exceptional either in their ability to accrue resources or in the possession of "good genes," or both. The animals' basis for mate choice may be quite straightforward and correlated with these ultimate benefits but there is no implication that the animals understand these relationships. The female that mates preferentially with large males gains benefits without understanding the forces of natural selection that have shaped her preference.

A female could potentially increase her reproductive success by pairing with a male that provides more resources than do other males. In some avian species, for example, males arrive at the breeding ground first, establish territories, and display to females. The female that pairs with a male in a prime territory containing a protected nest site and an ample supply of food may fare better than one selecting a male on a marginal territory. In the black-tipped hangingfly, as in some other insects, courting males present a prey item to females. The female choosing a male with a large prey item gains a larger resource and may have greater reproductive success even though there is no prolonged male-female bond. In addition to the material benefits in these examples, however, the female's offspring will possess genes that could lead to fitness levels superior to those of alternatives. Benefits gained through genes and resources in such cases may be very difficult to separate.

In many species, however, males may contribute no obvious resources beyond their ejaculates. Mating is brief, there is no prey presentation, and pair bonds are nonexistent. One might suppose that selection would favor females mating with males carrying genes relating to traits that are adaptive in their habitat. For example, if large size might allow her offspring to sequester a differentially large number of mating partners or resources and if size has a genetic component, females would be better off mating with large males than with smaller ones. There are many studies demonstrating such choice in a variety of situations.

Kinship. Kinship may also play a role in mate choice. The offspring of matings between close kin are often of lower viability or fertility than those of animals more distantly related. There appear to be many adaptations that function to minimize incestuous breeding and the decrements in fitness that may follow. On the other hand, interspecific matings may be infertile and populations may become adapted to local conditions, making matings with individuals from other populations disadvantageous. Females thus should not mate with individuals too much different from themselves. There is some indication that females may prefer mates with an optimal level of discrepancy from themselves—similar but not too similar. In Japanese quail, for example, individuals reared with their siblings mate preferentially with first cousins. The conditions of rearing often play a major role in determining which individuals are treated as kin; although the ultimate benefit may be avoidance of inbreeding, the proximate cue may be familiarity.

The major histocompatibility complex (MHC) genes of house mice code for individuality of the immune systems in the body. Mice can discriminate individuals of different MHC genotypes, apparently on the basis of odor differences. They have been shown to display mate choice related to MHC genes, generally preferring mates different from themselves. This may provide the offspring with the capacity to respond to a wider range of antigens than would otherwise be the case. *See* HISTOCOMPATIBILITY.

Male choice. Males forming prolonged pair bonds invest much, and it is not surprising that they may exercise mate choice as well. Even males of promiscuous species, however, may show some evidence of mate choice. The male's ejaculate represents a nontrivial investment; it may take days for a male to replenish his supply of sperm or the secretions of his reproductive accessory glands. The spermatophores of some cricket species may constitute up to 30% of body weight. With a limited number of ejaculates to deliver, some exercise of male choice is to be expected and has been observed. In a variety of species of invertebrates, for example, males prefer large females to small ones. In both *Drosophila* and prairie voles, males prefer unmated females to recently mated females. Male red-spotted newts, with limited sperm-producing ability, prefer large females as\break mates.

Courtship. Courtship entails a sequence of behavioral patterns that eventually may lead to the completed mating. Patterns of courtship are quite diverse among different species but generally entail reciprocal signaling between male and female.

A favorite behavioral pattern of classical ethologists, the courtship behavior of the three-spined stickleback, provides a case in point (**Fig. 2**). At the onset of the breeding season, males move away from schools, adopt coloration typical of the breeding season (including a red belly), and become territorial. They build a nest within the territory. When the female enters the territory, the male initiates a zigzag dance, in which he darts back and forth toward and away from the female and the nest. Females signal their acceptance by adopting a head-up posture. The male leads the female to the nest

and she follows. He hovers above the nest, pointing at its entrance; the female enters the nest. The male then nudges the base of the female's tail with his snout which stimulates the female to spawn. The male fertilizes the eggs, the female leaves, and the sequence begins again with a new female until several females have spawned in the nest. The male cares for the eggs and the young. The courtship sequence can be seen to involve coordination and reciprocal stimulation between male and female as the encounter progresses stepwise to its completion with mating.

Mate choice is an important function of courtship. Many bouts of courtship break off without going to completed matings, often as a result of choice on the part of one or both partners. An important basis for choice is the species of the potential mate. Because of the disadvantages of interspecific matings, a whole array of adaptations, called reproductive isolating mechanisms, function to prevent such matings. Courtship isolation is one of the more important of these. A courtship sequence may be initiated but broken off when the partner fails to signal, or to respond to a signal in the manner characteristic of the species.

There is often much assessment of potential mates beyond that of species identity during courtship. It is during courtship that a male hangingfly presents his prey that may be accepted or declined by the female

Fig. 3. Pair of copulating scorpionflies (*Hylobittacus apicalis*, also called black-tipped hangingflies). The female on the right is feeding on a nuptial gift (blowfly). (*Courtesy of Randy Thornhill*)

(**Fig. 3**). In common terns it is the male that provides most of the food for the hatchlings. During courtship the male presents fish to the female. It has been found that the amount of food a male feeds a female during courtship is correlated with the amount of food he will later provide for the young, and with the total weight of the clutch. The female appears to select a male on the basis of the food presented and, in so doing, is apparently able to assess his effectiveness as a resource provider. There is no implication that it is done consciously.

Another function of courtship relates to synchronization. The gametes must be shed at a time when sperm are viable, eggs are ripe, male and female are in the appropriate state of readiness, and the environment is supportive of reproductive effort. The progressive interactive sequence of the courtship episode allows for coordinated events to occur at times appropriate for successful reproduction.

Arthropods. Courtship patterns are diverse among arthropods. Initial detection can occur in a number of ways. Silk moths, for example, are extremely sensitive to a sex attractant, called bombykol, produced by the female. A single molecule of bombykol can initiate a nerve impulse in a receptor cell; males can orient to females even in response to very low concentrations of the substance. More familiar to humans are the auditory signals of species such as crickets and katydids, which function to attract females and, in addition, sometimes to repel other males. They may thus aid in defense of territories. Signal patterns are species-characteristic. Females prefer the calls of males of their own species, while hybrid females may even prefer the calls of hybrid males. *See* CHEMICAL ECOLOGY.

Many species of arthropods use visual signals, the most spectacular of which are perhaps the courtship signals of fireflies. In many species the males fly above the ground and emit flashes in a species-typical pattern. The females are generally in the grass below and respond to a male's signal with a characteristic flash delivered with a specific latency after a male's signal. The male approaches the female, and pair members continue to signal as they approach and eventually mate. A skilled field investigator can

male   female

                                  appears
zigzag
dances
                                  courts

leads
                                  follows

shows nest
entrance
                                  enters
                                  the nest
trembles
                                  spawns
fertilizes

(c)
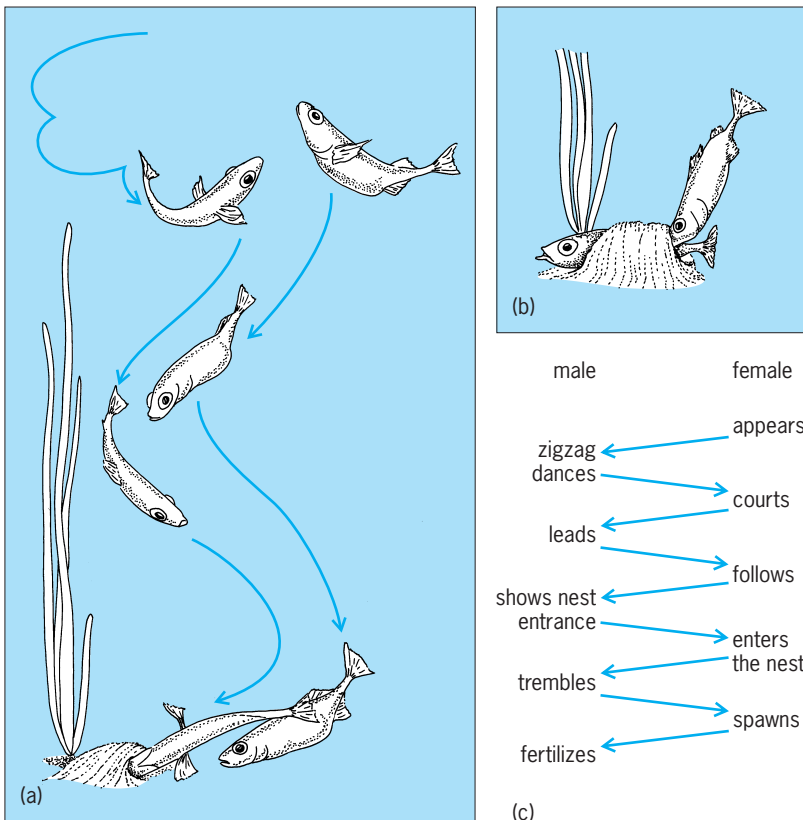
Fig. 2. Mating behavior of sticklebacks. (*a*) Zigzag dance of a male leading a female to the nest. (*b*) The male stimulating the female in the nest to lay eggs. (*c*) Schematic portrayal of the sequential events in a courtship sequence. (*After N. Tinbergen, The Study of Instinct, Oxford University Press, 1951*)

attract a male by mimicking the female's response pattern. Females of predatory firefly species have adopted the same trick: by mimicking the pattern of the females of the male's species, the predatory female attracts the male and devours him. This is but one instance of a predator cuing in on a courtship signal as a means of obtaining prey. Signaling male crickets, to take another example, also attract predators as well as females. *See* ANIMAL COMMUNICATION.

Male-male competition occurs during firefly courtship as well. In species in which the male signal involves a series of flashes, another male of his species may "inject" a flash into a signal. Such injected flashes may disrupt the courtship interaction. Patterns of such flashes provide further examples of the ways in which behavior is adapted to serve the interests of individuals, and cooperation in the perpetuation of the species occurs only so long as it is in the interests of the individuals to act in such a manner.

Fishes. The courtship pattern of sticklebacks, described above, is but one of the many diverse courtship patterns of fishes. Those of the tropical fishes kept by aquarium enthusiasts are perhaps the most familiar. Male guppies are among the most persistent courters. They often adopt a sigmoid (S-shaped) posture with the tail deflected, and a quivering motion. The display is usually given in the presence of a female, and during the display the male may move toward her. The male's gonopodium is a modified anal fin and may be swung laterally and forward, and thrust toward the female during courtship.

The beautiful patterns of the anabantoids, such as Siamese fighting fishes and gouramis, can be observed in a home aquarium. After the male builds a bubble nest on the surface of the water, there are sequences of patterns such as tail beating, leading, lateral displays, sigmoid displays, and circling prior to spawning.

Amphibians and reptiles. The calls of bullfrogs and other anurans are familiar to many. Males typically arrive at the breeding area first and establish a territory. Their species-typical mating calls generally attract females, stimulate female reproduction, and function in reproductive isolation. They may also attract predators. Some males, called satellites, may escape the risk of predator attraction by not calling at all. Rather, they may adopt a post near the territory of a calling male and attempt to mate with females approaching the calls. Such males often meet with some reproductive success. Such alternative mating strategies within a single stable breeding population are not uncommon.

Courtship in many species of anoles and other lizards involves characteristic patterns of head bobbing, together with other bodily movements. In the American "chameleon," *Anolis carolinensis*, there is a bright red fold of skin, the dewlap, that is exposed during such displays (**Fig. 4**).

A female red-sided garter snake emerging from hibernation typically encounters a mating ball of as many as 100 males competing to inseminate her. The

Fig. 4. The extended dewlap of a male anole lizard (*Anolis carolinensis*) is part of the striking display pattern in this and related species. (*Courtesy of Thomas A. Jenssen*)

male displays patterns of tongue flicking and chin rubbing prior to coupling.

Birds. The courtship displays of ducks and geese have been the subject of intensive analysis. Patterns are similar in closely related species and, indeed, courtship patterns can be used to clarify taxonomic relationships that may be ambiguous when only morphological characters are used in classification. Such patterns as bill shaking, the tail shake, the grunt-whistle, and the head-up tail-down display can be seen in many duckponds.

Males of various species of Australian bowerbirds build structures (bowers) that attract females. In the golden bowerbird, males build a maypole bower that often is 8 ft (2.5 m) high. The bowers are decorated with mosses, flowers, and fruits, and the males display and vocalize in them. Satin bowerbirds paint and decorate their bowers elaborately, especially with items that are blue.

Bird songs may function in courtship, territorial advertisement, and other contexts.

Mammals. Courtship in mammals is generally less spectacular than in other taxa. Olfactory cues, which are important in many taxa, predominate in mammals. Thus, there is often much mutual grooming, sniffing, and licking, especially of the partner's genital area. Males of some species display a Flehmen response, in which the upper lip is curled and the neck is extended. The posture probably facilitates exposure of the odor to the vomeronasal organ via a duct in the front of the mouth. Animals of many mammalian species can discriminate an animal's species, sex, reproductive condition, and sometimes other characteristics as well, solely on the basis of olfactory stimuli. A variety of marking patterns make the odors particularly accessible.

Mating

Mating represents the consummation of sexual activity wherein the sperm are transferred from male to female so that the fertilization of ova can occur. Fertilization may occur internally (within the female's body) or externally (outside the body of the female). *See* FERTILIZATION.

Patterns of mating behavior. Different taxa display not only different modes of fertilization but also different patterns of mating behavior.
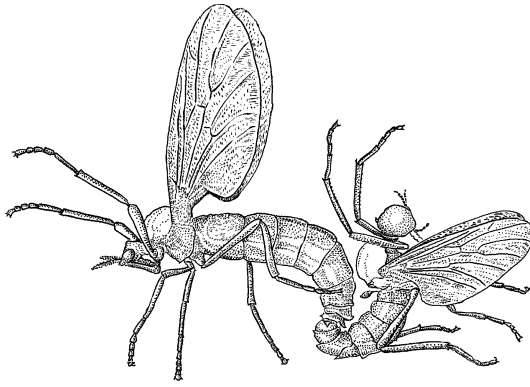
Fig. 5. Pair of lovebugs (*Plecia nearctica*) copulating. The female is on the left. These animals often remain in copulation for 2 days or more. (*After R. Thornhill and J. Alcock, The Evolution of Insect Mating Systems, Harvard University Press, 1983*)

Arthropods. In many species of arthropods the sperm are transferred in a spermatophore, a saclike structure in which they are enclosed. The spermatophore may also serve to plug the vaginal orifice and may have nutritional value. In some species of mites, pseudoscorpions, and millipedes, the male simply deposits the spermatophore on the substrate and the female comes along and picks it up; the pair may never meet. External fertilization occurs in a number of aquatic species, such as horseshoe crabs.

Copulation may be quite prolonged in species with internal fertilization; the mean duration of copulation in Florida "lovebugs" (*Plecia nearctica*) in the laboratory was found to be 56 h (**Fig. 5**). In many species the male clasps and locks the female during copulation. These clasps are maintained by a remarkable array of adaptations of mandibles, genital claspers, antennae, and modified legs. In the beetle *Lytta nuttalli* the penis contains dorsal and ventral penile spines that catch onto folds in the vaginal wall to secure a lock. Copulation in the southern green stink bug may last anywhere from 5 min to 14 days.

Male giant water bugs brood the fertilized eggs on their backs. Females deposit eggs a few at a time, and males insist on repeated copulations before each oviposition. In one case a pair coupled over 100 times in 36 h during the transfer of 144 eggs. In contrast, copulation in damselflies lasts just a little more than a minute. Sperm are transferred only during the last part of the copulation. Prior to that, males engage in undulating movements that, because of the special structure of the damselfly penis, serve to displace any sperm from males that mated previously with the female.

Fishes. In various species of the anabantoid bubble-nest builders mentioned above, as in many other species, fertilization is external. Sperm and eggs are released simultaneously, and fertilized eggs float up to the nest. Fertilization is also external in the Agnatha (jawless fishes). In the sea lamprey the female attaches herself to a rock with her sucker mouth; the male attaches himself similarly to the female and wraps his body around hers. Both vibrate their hindquarters and release gametes simultaneously.

Fertilization among sharks and other cartilaginous fishes is internal with sperm transfer effected via the male's clasper, a specialized intromittent organ derived from the basal portion of the pelvic fins. Sperm are propelled down a groove in the clasper. *See* COPULATORY ORGAN.

Fertilization is internal in many species of bony fishes as well. Copulations in guppies last just a few seconds. In the platyfish, copulations last 1 s or more; during longer copulations, pairs appear locked together as they swim but then they pull apart violently.

Oral fertilization has been found in the cichlid fish, *Pseudocrenilabrus multicolor*. The females brood eggs in their mouths, and the males have conspicuous anal-fin markings that resemble eggs. When females snap at the egglike markings as if to retrieve them, they suck up sperm that can fertilize the eggs within their mouths. External fertilization may also occur.

Amphibians and reptiles. In most anurans the basic mating pattern is the mating clasp, or amplexus, in which the male secures a firm hold on the female with his forelegs; fertilization is external (**Fig. 6**). Tailed frogs (*Ascaphus truei*) have an intromittent organ; during amplexus it is inserted into the female's cloaca, an opening near the base of the tail. In African toads there is also internal fertilization, effected by direct contact between the cloacae of the male and female.

In salamanders, fertilization is generally accomplished with a spermatophore. Typically the interaction begins with courtship, and the male deposits a spermatophore on the substrate. The female walks around and picks up the spermatophore between the lips of her cloaca. In some species, such as the red-spotted newt, the male captures the female in amplexus prior to depositing the spermatophore—especially if the female is unresponsive.

Many species of reptiles have internal fertilization, which is effected via one of the male's paired hemipenes. In anoles, for example, the male gains a neck grip on the female, twists his tail around her body, and inserts a hemipenis.

Birds. Although the courtship patterns of birds are remarkably diverse, there is little variability at

Fig. 6. Amplexus. In frogs and toads the male secures a firm clasp on the female. (*Photograph by D. Kim Sawrey*)
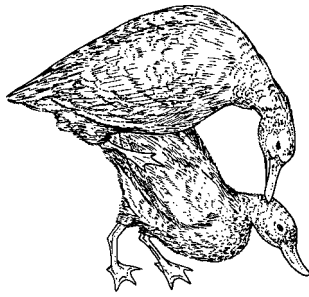
Fig. 7. Copulatory stance with the "cloacal kiss" in whistling ducks. (*After E. G. Bolen, Copulatory behavior in Dendrocygna, Southwest. Natural., 18:341–357, 1973***)**

mating. Males of most species lack a penis; fertilization is effected via a "cloacal kiss," wherein the male and female cloacal glands are brought together and sperm are passed from male to female. Typically, the male stands on the female's back, sometimes with a neck grip, and the cloacae are brought together (**Fig. 7**).

Mammals. Mammalian copulatory patterns are quite diverse, although fertilization is always internal. Patterns can be usefully examined in relation to four characteristics: locks, intravaginal thrusting, multiple intromissions, and multiple ejaculations. In locking species, there is a mechanical tie, resistant to separation, between the genitalia of the male and female. Locking is found in some marsupials, short-tailed shrews, some rodents (such as golden mice, grasshopper mice, and some woodrats), and many carnivores (such as dogs and wolves).

In some species, pelvic thrusting ceases when the male gains insertion; this is true of many rodents, lagomorphs, and ungulates. In other species, however, as in virtually all primate species, there is repetitive intravaginal thrusting.

The third characteristic is the necessity of multiple intromissions as a prerequisite to ejaculation. In many species of rodents and primates the male repeatedly mounts the female and gains vaginal insertion only to dismount without ejaculating. Such multiple intromissions are prerequisite to ejaculation. In other species, ejaculation can occur the first time the male mounts and gains vaginal insertion.

The occurrence of ejaculation appears to terminate the copulatory interaction of a particular male and female in some species, such as certain shrews, squirrels, primates, pinnipeds, and ungulates. In other species, the pair may mate for several ejaculations.

Because each of these characteristics can be either present or absent, there are 16 theoretically possible patterns of copulatory behavior, if that behavior is described solely by the four characteristics mentioned here. The copulatory behavior of laboratory rats has been particularly well studied. The male mounts the female from behind. The female then adopts a stereotyped posture, lordosis, with a concave flexion of the spine, elevation of the head and tail region, and deflection of the tail (**Fig. 8**). During intromissions, the male mounts, gains penetration, and dismounts after an insertion of just 250–300 ms. He may engage in other behavior for a minute or so before remounting. About 10 such intromissions may be spaced a minute or so apart before the male displays a longer insertion that signals ejaculation. After a refractory period of approximately 5 min, the sequence begins anew—this time generally with fewer intromissions that are spaced more closely together. After about seven such ejaculations, refractory periods become so long that males can be regarded as satiated. This pattern can be seen as one with no locking, no intravaginal thrusting, multiple intromissions prerequisite to ejaculation, and multiple ejaculations. It is also characteristic of several species of gerbils, deer mice, and harvest mice.

The copulatory pattern of dogs contrasts with that of laboratory rats. Although males may mount (without insertion) several times prior to completed matings, they are capable of ejaculation the first time they mount. On gaining intromission, the male makes distinctive postural adjustments and begins a pattern of stepping with the rear legs. The bulbus glandis of the penis becomes engorged with blood and locks the male and female together. After an intense ejaculatory reaction the male typically dismounts while still locked to the female. In a study of beagles, the mean duration of locks was 14 min. Dogs display locking, intravaginal thrusting, ejaculation on a single insertion, and multiple ejaculations.

Japanese macaques display a pattern with no lock, intravaginal thrusting, multiple intromissions, and multiple ejaculations. By contrast, in black-tailed deer there is no lock, no thrusting, ejaculation on a single insertion, and just one ejaculation.

In many species the female is far from passive. In rodents, for example, the female often solicits copulations by approaching the male and then darting away. Female rats will learn an arbitrary task when the reward is the presentation of a sexually active male that will copulate with her. In primates the female uses a "present" posture to invite copulations; the same posture is used in other contexts, such as in indicating social subordination.

Mechanisms underlying mating behavior. Sensory control of mammalian copulatory behavior is generally complex, with no single sensation critical to the normal occurrence of behavior. Visual cues, such as the primate present posture, are important in some species. Olfactory cues play an especially important

Fig. 8. Lordosis in a female rodent. When mounted, the females of many species adopt the lordotic posture rendering copulation possible. Note the concave arching of the spine. (*After L. C. Drickamer and S. H. Vessey, Animal Behavior: Concepts, Processes, and Methods, Willard Grant, 1982***)**

role in a wide range of species. As might be expected, tactile sensitivity is critical to normal behavior; interference with an animal's sense of touch, especially that from the genital region, disrupts behavior. Auditory cues are important in many species. In various species of rodents, for example, males emit ultrasonic calls, inaudible to human observers, which are important in the interaction.

Hormones are critical to the display of copulatory behavior in most species. In general, hormonal effects in adults are reversible. If an animal's gonads are surgically removed, mating behavior will often disappear. Such behavior reappears, however, after treatment with appropriate hormone regimens—generally androgen injections for males and injections of estrogen followed by progesterone for females. In most species, female sexual behavior appears more closely tied to hormones than does male behavior. Males may continue to copulate for weeks after castration, whereas females cease to mate almost immediately after a surgery. This may be termed an associated reproductive pattern, as mating coincides with both gamete production and maximum levels of steroid hormones.

Species with dissociated patterns generally live in harsh environments, where there is little time for mating. Mating may occur in the absence of gamete production at a time when steroid hormone levels are low. In Canadian red-sided garter snakes (*Thamnophis sirtalis parietalis*), for example, mating occurs within about 3 weeks after emergence from hibernation. Later, testosterone levels increase and gametes are formed, to be stored until the next breeding season. Gonadal sex hormones do not play a major role in the control of mating in these species.

Many parts of the nervous system are important to the mediation of copulatory behavior. The reflexes that compose normal sexual behavior are partially organized in the spinal cord and hindbrain. A male dog with a transaction in the thoracic region of the spinal cord displays a complete set of spinal reflexes in response to tactile stimulation even though the neural connections with the brain have been completely severed. Within the brain, structures in the region of the hypothalamus are particularly important for reproductive behavior. This region lies at the base of the brain and sits atop the pituitary gland—the master control gland of the endocrine system. Various structures within the region are important in regulating the pituitary gland and, via both direct and indirect routes, controlling copulatory behavior as well. A variety of structures in the midbrain and forebrain have been found to be part of complex networks involved in the regulation of copulatory behavior. *See* NERVOUS SYSTEM (VERTEBRATE); PITUITARY GLAND.

Development of copulatory behavior. Copulatory behavior is not static, but changes, as the result of the interactions of genes and the environment, as an animal matures and ages. In vertebrates, sexual activity is delayed until a time of puberty. This is generally hormone-dependent; precocious copulatory behavior can be induced with exogenous hormone treatment. The behavior of the animal maturing naturally can be seen to unfold gradually. In the first copulatory episode of a male rhesus monkey, for example, the orientation of the mount is poor. Similarly, the first time a male rat copulates he may take longer to initiate mating and require more intromissions before ejaculating than he will later. Both age and experience can be important in effecting these changes. After a period of relative stability in adulthood, copulatory behavior often declines in old age. Often this takes the form of an increasing threshold for copulation; behavior may be quite normal once it occurs.

The early social environment is important to the development of normal copulatory behavior. In rodents, for example, deprivation of contact with the mother and siblings during development can lead to a complete lack of sexual activity or to aberrant behavior, such as mounting that is inappropriately directed. Similarly, monkeys and chimpanzees need to interact with others if adult behavior is to be normal; experience may be even more important for rhesus monkeys than for chimpanzees.

The most direct evidence for a genetic basis for copulatory behavior comes from studies related to behavioral development. Although raising an animal with an alien species may alter its mate preferences and raising it in isolation may eliminate the behavior altogether, the motor patterns of copulatory activity generally are species-typical when they occur. There is good direct evidence of genetic influences on copulatory behavior from studies of various strains and their crosses and from genetic selection experiments. Such studies have been done in house mice, laboratory rats, Japanese quail, and chickens. The importance of genetic influences has been repeatedly confirmed.

The actions of hormones during development are different from those in the adult, especially in that their actions appear less reversible. In mammals, sexual development in the absence of gonadal hormones is female; male characteristics fail to develop. If a newborn rat, for example, is deprived of its gonads, as an adult it will display the lordotic female mating posture in response to estrogen, progesterone, and mounting, but it will not show male-like behavior even if given androgen injections. By contrast, if the animal is exposed to gonadal steroids during the first few days of life, as an adult it will display normal male behavior in response to androgens, but the expression of female behavior will be blocked. Such relationships vary with taxon. In birds, for example, where males rather than females are the homogametic sex, development in the absence of hormones is male and early hormones feminize and demasculinize rather than masculinize and defeminize as in mammals.

The development of copulatory patterns is a continuous, dynamic process reflecting both social and organismic interactions. Thus, newborn rats exposed to androgens produce hormone-dependent odors, which stimulate a specific pattern of licking by the mother. That behavior, in turn, contributes to the development of masculine copulatory behavior.

The mother's licking is stimulated, in part, by her salt and water appetite. Mothers recover significant amounts of water by licking, and in turn stimulating, the young. The mother–young relationship in development is thus a reciprocal one.

Satiation. Although multiple ejaculations occur in many species, the rate of occurrence of copulatory behavior typically declines as the interaction continues. Some species, such as many species in the cat family, may continue to copulate regularly for several days. In many other species, however, a point of satiation is reached after a given number of ejaculations in a day.

Recovery from satiety also occurs at variable rates. It takes about a week or more for measures of the copulatory behavior of satiated male rats to return to the baseline levels characteristic of rested males. Recovery in male hamsters is much faster.

In many species, such as laboratory rats and various species of farm mammals, the copulatory activity of satiated males can be reactivated by replacing the original mating partner with a novel female— a phenomenon known as the Coolidge effect. Under some conditions a male that appeared satiated with one female may deliver several additional ejaculations when presented with a novel mating partner.

Sequelae to Mating

The most obvious consequence of mating activity is the fertilization of ova. With external fertilization the location of fertilization is obvious, but in species with internal fertilization, sperm and ova often come together in a specialized location, such as an oviduct or an analogous structure, and fertilization is effected. The sequelae to copulatory behavior are often somewhat more complicated than this, however.

Immediate consequences. The stimuli derived from mating can have important immediate physiological consequences. In some species, such as rabbits and cats, the ova are not released until triggered by copulation or similar vaginocervical stimulation. This is a kind of built-in "hold" and functions to broaden the range in time that a female can mate effectively. The postcoital afterreactions of female cats are correlated with intense physiological responses to copulation.

In other species, such as laboratory rats, mice, and hamsters, ovulation is spontaneous but there is an induced pseudopregnancy. A neuroendocrine reflex, operating via the nervous system, pituitary gland, and ovary, stimulates functional levels of progesterone in preparation of the uterus for the implantation of the fertilized ovum.

The proliferation of the uterine wall occurs only in response to vaginocervical stimulation. In different species of rodents the multiple intromissions preceding ejaculation and multiple ejaculations appear critical in triggering this reflex.

Sperm transport. Sperm transport is an active process in many mammalian species. One function of the repetitive multiple intromissions preceding ejaculation in rats is to facilitate sperm transport. Sperm that are delivered to the female without a full complement of preejaculatory intromissions fail to reach the uterus.

Sperm transport can be disrupted if copulation occurs too soon after a female receives an ejaculate. Female rats require about 5 min for appreciable sperm transport. If a female receives vaginal stimulation during that 5-min period, transport can be disrupted and, if she receives no more ejaculates, pregnancy may not occur. Male deer mice appear able to "cancel out" another male's ejaculation by copulating with a female within a minute or so after delivery of the ejaculate. This is presumably the result of disruption of sperm transport and possibly actual removal of sperm. The process appears analogous to that in damselflies discussed above.

Sperm competition. In many species the females copulate with more than one male during a single receptive period. Sperm competition is the competition between the ejaculates of two or more males within the reproductive tract of a single female. Because natural selection acts at the level of the individual and via differential reproduction, strong selection on effects related to sperm competition would be expected. Sperm competition was first studied extensively in insects, because of their great sperm storage capacity, but it is clear that in other species, such as sheep, deer mice, ground squirrels, and Norway rats, multimale mating may be quite common, and thus sperm competition occurs.

The pattern of sperm competition varies with the species and prevailing conditions. In the giant water bugs discussed above, for example, there is a strong last-male advantage. The last male to deposit sperm sires a disproportionately large number of young. This is a common pattern in insects. However, there are some species of insects and mammals (for example, house mice) in which the first male to ejaculate gains a disproportionate reproductive advantage. In others, such as laboratory rats and deer mice, there are no order effects. Males can gain a reproductive advantage in these species by delivering a higher percentage of the ejaculates received by a female than other males. The systems that regulate sperm competition are dynamic, and interactions are variable within species as well. The delay between males, relative numbers of ejaculations, and timing of the matings all affect which male prevails in sperm-competition interaction.

Parturition and Parental Behavior

There are various means used by animals to nurture developing young and to discharge them into the environment. In oviparous (egg-laying) species, fertilized eggs develop outside the body of the adult—even though fertilization may be internal. In viviparous (live-bearing) species, by contrast, the embryo develops within the reproductive tract of the female and derives nourishment from her. Ovoviviparous species bear live young, but they hatch from eggs that develop within the parent's body; this pattern is found in some snakes and fishes.

Patterns of parental behavior. Prior to birth or hatching, the parent may engage in behavioral patterns

that will aid the young when they arrive. This may entail preparation of a burrow or nest, provision of stored food, or acquisition of other resources. In some species the parent's aid ends with such preparations, but in others parental care may be extensive and prolonged.

Arthropods. In most arthropods, there is no parental care beyond preparation of an appropriate site and deposition of eggs in a favorable location. The most dramatic exceptions to this general rule can be found in the eusocial species—bees, wasps, ants, and termites, in which there is typically a division of labor, cooperation in the rearing of young, and an overlap of at least two generations at a time. Often there are sterile castes and care for the young is elaborate— typically in an extensive hive, nest, or burrow. Although sterile workers do not reproduce, they aid their close kin, which is to their genetic advantage even though the offspring produced are not their own. This view is compatible with the view that natural selection works at the level of the individual. However, this is only part of the story, and factors other than the close genetic relationship among females in some eusocial species appear essential in the evolution of this pattern. *See* SOCIO-BIOLOGY.

Less elaborate parental care is found in a variety of species, such as many true bugs, thrips, web-spinners, and hydrophilid beetles. Typically, mothers guard the young, shield them from predators and parasites, and may even aggressively attack predators. In some species the young may be guarded through several developmental stages.

In giant water bugs it is the males that care for the eggs, which are cemented to their backs. Males of several species of tropical assassin bugs straddle eggs and guard them from parasites and predators.

Examination and comparison of the insect species that do and do not show parental care and of the environments in which they live leads to the conclusion that parental behavior represents an adaptation to harsh living conditions of one type or another.

Fishes. The fishes may be the most variable taxon with respect to parental behavior and are ideal for tests of hypotheses regarding the evolutionary pressures creating variability in parental behavior. Eggs may be scattered on plants, over shoals, gravel, sand, or boulders, or in the open ocean, but in many species there is elaborate male care, female care, or biparental care. In seahorses and pipefishes the female deposits eggs in a special brood pouch on the male. The male stickleback aerates the nest and tends young. Active herding occurs in some species and others are mouthbreeders; they guard the young within the oral cavity during times of danger. The aquarist's *Tilapia* provide a familiar example of a mouthbreeding species.

Amphibians and reptiles. Although parental behavior in amphibians and reptiles is generally less elaborate than in other taxa, there are also many parental species. Among different species of salamanders the males or females may protect the young—often from attacks by conspecifics. Male midwife toads wrap eggs around their hindlegs after fertilization and carry them until they hatch.

Although snakes do not generate their own heat, they do absorb heat; in some species, adults coil themselves about eggs to keep them warm. Parental behavior is elaborate in many crocodilians, which guard the nest during incubation of the eggs and herd and defend a pod of young through early posthatching development.

Birds. Parental behavior is highly developed among birds. The diversity and elaborateness of bird nests are widely known. Although there are colonial species in which large numbers of birds build the nest, most commonly a single mated pair builds the nest for the eggs and defends it against intruders. In some species, "helpers," typically the young from previous years, may aid the parents in raising the young. The division of labor between male and female in nest building and incubation varies widely.

Temperatures above ambient levels are important for the development of avian eggs and are provided by body heat as one or both parents sit on the eggs. Typically, the feathers are erected so that the skin of the parent comes into direct contact with the eggs. In some species, there are specialized brood patches, areas in which blood vessels proliferate, skin is thickened, and feathers are lost during the incubation period.

The interactions of male and female in sharing parental care can be subtle. Male and female ring doves (*Streptopelia risoria*), for example, take turns sitting on the nest, with the male sitting for a period in the middle of the day and the female sitting the rest of the time. The timing of parental care is regulated by biological clocks, internal timing mechanisms that are entrained by the occurrence of light and dark during the 24-h cycle. Although the timing mechanisms differ in the male and female, the end result is an integrated pattern of shared brooding of the young. *See* BIOLOGICAL CLOCKS.

The American cowbird and various species of cuckoos are brood parasites. Rather than incubate their own eggs, they lay eggs in the nests of other species, which incubate the eggs. A variety of adaptations ensure that the parasitized host is deceived into caring for the parasites. The eggs of the parasite resemble those of the host. Typically, the eggs of parasites hatch before those of the host, thus giving the parasite an advantage in gaining parental attention. In some species the young parasite may roll the host's eggs from the nest; in others the parasite grows faster and larger than the host's young and again is able to monopolize parental resources.

The young of precocial bird species require little care; in such species as chickens and ducks the mother can be seen leading the young about, directing them to food and protecting them. In altricial species, the young are quite helpless when hatched as they are blind, unable to locomote, and unable to feed themselves. Prolonged parental care is required for normal development. It is generally believed that biparental care has evolved to meet the energetic and other demands of the developing altricial young.

Mammals. Parental care, especially maternal care, is highly developed in all species of mammals. By definition, the females of all mammalian species possess mammary glands for the nourishment of young. It is probably for this physiological reason that role reversal is less common among mammals than in other taxa.

Parturition in mammals is characterized by periods of uterine contractions, emergence of the young, delivery, and care or consumption of the placenta. Often there is appreciable squatting or straining during delivery. Prior to, during, and after delivery, there is much licking of the female's body, the newly delivered young, previously born young, and the umbilical cord. Often the female ingests the fetal membranes and the cord.

Nursing is initiated soon after birth. Typically the mother approaches and adopts the nursing posture. In cats, for example, the female lies near the kittens and arches her body about them. The young may develop specific nipple preferences. Later in development there may be mutual approach by the mother and young. The young continue to become more active in initiating nursing until finally the mother weans them.

Mammalian parental behavior also involves behavioral patterns other than nursing. Nests of varying degrees of complexity are built by various species. In rabbits, for example, the female digs or excavates a burrow and collects straw or other material with which to line it. At this stage, the female's hair begins to loosen and she plucks it and uses it to line the inside of the nest. Mothers may visit the young to nurse them only once per day, but the young are kept warm and safe. Of course, the mother must also feed, maintain social relationships, and engage in other activities necessary for survival. Her activities may be conceptualized as a kind of time sharing, the pattern of which changes as the young mature.

The manner of transporting the young varies with different species. Typically, a young rodent or canid is carried in its mother's mouth. Baboon mothers carry young on their chests, with a transition to riding on the mother's back, jockey style, that begins at about 5 weeks of age. The young of precocial mammalian species, such as most ungulates, are not transported by the mother.

Parents not only defend young from predators and conspecifics but also play an active role in tutoring them. This is most obvious in carnivorous species. A mother cat, for example, uses a method of shaping in encouraging her young to hunt. She first brings prey back to the nest and eats in front of them. Later, she leaves prey for the young to eat and still later presents them with live prey to kill for themselves. Developing carnivores accompany parents on some hunts and learn to catch prey in their company. Similarly, sea otters stay close to their mothers as they dive for food and eventually develop skills of diving and locating prey themselves. Young rodents often accompany adults on feeding trips and learn what to eat via cues from them.

The role of male mammals in caring for young has been increasingly recognized. In many species of rodents, for example, males engage in virtually all parental behaviors with the exception of nursing. The males of various species of primates and carnivores play an important and active role in the care and tuition of young.

Parental responsibilities may not end with the maturation of the young; grandparents play a role in development in some species. In vervet monkeys, for example, the presence of a maternal grandmother in a troop is associated with an increase in the survival of the infants of young adult females. Grandmothers play an active role in the development of the young, with the interaction being more prolonged with granddaughters than with grandsons.

Mechanisms of parental behavior. The occurrence and integration of cycles of parental behavior are a function of reciprocal interactions between the internal and external environments of the parents and young. Fluctuating hormone levels appear important in stimulating parental behavior and are, in turn, affected by the behavior of the partner and the young. In ring doves, a pair-forming species, the birds select a nest site, build a nest, and copulate. The male gathers most of the nesting material, and the female does most of the building. As the female's oviducts grow, she becomes progressively more attached to the nest site. When the eggs are laid, the members of the pair take turns incubating them, with the female on the eggs about three-quarters of the time. Once they are hatched, the young are fed with crop "milk," a regurgitated substance produced in the lining of the crop. The parents become progressively less interested in the young as they develop until the young fledge and a new cycle begins.

The various stages of the ring dove cycle are correlated with dramatic motivational changes. The bird that engaged in intensive nest building at one stage is equally intense about incubation at the next. These motivational changes are correlated with fluctuating hormone levels. In females, estrogen is associated with nest building and progesterone with incubation. These hormones have been shown to act on specific brain regions in females.

In mammals, too, fluctuating hormone levels are related to maternal behavior. Nest building, retrieval of pups, nursing, and grooming can be induced in laboratory rodents with various schedules of administration of estrogen, progesterone, and prolactin. Cross-transfusion of blood from a mother to a nonmaternal female stimulates maternal behavior in the recipient.

Also in mammals, various brain structures are involved in maternal behavior. With lesions of the medial preoptic area of the hypothalamus, for example, maternal rats fail to build nests, retrieve pups, or nurse. With lesions in certain portions of the limbic system of the forebrain, the behavioral patterns related to maternal behavior occur but in disrupted sequences, so that a mother may pick up and deposit pups but never complete their return to the nest. The timing of mammalian parental behavior is the

product of integration of the nervous and endocrine systems and external stimuli.

Parental investment. Parental investment entails any investment by the parent that increases the ability of the young to survive and reproduce at some cost to the parent. Much parental investment, like milk, cannot be shared; that which is given to one offspring cannot be given to another. Other kinds of parental investment, like defense against predators, is shareable. In addition, the parent making any investment is prevented from engaging in other activities, such as searching for his or her own food or seeking additional mates.

Because they contribute larger gametes and often engage in more extensive parental behavior, the females of most species display a higher level of parental investment than males. Members of the sex investing more (typically females) thus become a limiting resource for the sex investing less (typically males). It is generally agreed that this gives rise to the disparity between female and male reproductive strategies—with males more often competing for access to females and females more choosy than males.

In many nonmonogamous species, most females reproduce whereas some males may sire many offspring and others very few. Thus, variance in reproductive success is much greater among males than females. This difference in variance implies that sexual selection acts more strongly on males than females and relates to differences in parental investment and mating strategies.

Because natural selection works at the level of the individual, the interests of parent and young do not always coincide. In some situations it may be in the parent's best interest to spread resources equally among her young. Up to a point, however, it may be in the interest of a juvenile to demand a higher proportion of these resources for itself. This discrepancy of interests can generate parent-offspring conflict and the changing relationship between parent and young that leads to weaning.

Thus, the orderly integrated cycles of reproductive behavior generally require the cooperation of different individuals and result in the perpetuation of the species. Driving this activity, however, is the action of natural selection shaping individual organisms to behave in ways that will maximize their levels of fitness—contributions to the gene pools of future generations. Much of reproductive behavior can be understood in relation to these conflicting forces. The result, however, is a pattern that for each surviving species is adapted to its environment and social structure and results in the successful production of young. *See* BEHAVIORAL ECOLOGY; ETHOLOGY.
Donald A. Dewsbury

Bibliography. J. B. Becker, S. M. Breedlove, and D. Crews (eds.), *Behavioral Endocrinology*, 1992; E. M. Blass (ed.), *Developmental Psychobiology and Behavioral Ecology*, 1988; D. Crews (ed.), *Psychobiology of Reproductive Behavior: An Evolutionary Perspective*, 1987; D. A. Dewsbury (ed.), *Mammalian Sexual Behavior: Foundations for Contemporary Research*, 1981.

Reproductive system

The structures concerned with the production of sex cells (gametes) and perpetuation of the species. The comparative anatomy, human histology, embryology, physiology, endocrinology, and biochemistry of this system are treated in this article.

The reproductive function constitutes the only vertebrate physiological function that necessitates the existence of two morphologically different kinds of individuals in each animal species, the males and the females (sexual dimorphism).

The purpose of the reproductive function is fertilization, that is, the fusion of a male and a female sex cell produced by two distinct individuals. In each sex the reproductive system comprises a sex gland or gonad, which produces sex cells, or gametes, and ducts, which permit the passage of the gametes. In some animals, such as mammals, copulatory organs permit the male germ cells to be introduced into the female ducts and fertilization is internal, but in many vertebrates, such as anuran amphibians and many fishes, no copulatory organ exists and fertilization is external.

Comparative Anatomy

Egg cells, or ova, and sperm cells, or spermatozoa, are formed in the primary reproductive organs, which are collectively known as gonads. Those of the male are called testes; those of the female are ovaries. Besides giving rise to reproductive cells, both ovaries and testes give off endocrine secretions, or sex hormones, which pass into the blood or lymphatic streams and are carried to all parts of the body, where they bring about profound effects, not only on the rest of the reproductive system but on several other systems of the body as well. The gonads are paired structures, although in some forms what appears to be an unpaired gonad is the result either of fusion of paired structures or of unilateral degeneration.

The reproductive elements formed in the gonads must be transported to the outside of the body. In most vertebrates, ducts are utilized for this purpose. These ducts, together with the structures that serve to bring the gametes of both sexes together, are known as sex organs. The structures used to transport the reproductive cells in the male are known as deferent ducts and those of the female as oviducts. In a few forms no ducts are present in either sex, and eggs and sperm escape from the body cavity through genital or abdominal pores. The deferent ducts are usually the mesonephric or Wolffian ducts, which in some cases also serve to carry urinary wastes in those vertebrates in which opistonephros or mesonephros function either during embryonic or adult life. In vertebrates whose functional adult kidney is the metanephros, the Wolffian duct on each side persists as the ductus deferens. *See* KIDNEY.

In most vertebrates the reproductive ducts in both sexes open posteriorly into the cloaca. In some, modifications of the cloacal region occur and the ducts open separately to the outside or, in the male,

join the excretory ducts to emerge by a common orifice. *See* COPULATORY ORGAN; REPRODUCTION (ANIMAL).

The sex of an individual is dependent upon the chromosomes received from both parents at the time that the egg is fertilized. That the balance between maleness and femaleness is a delicate one, however, is reflected in the fact that environmental factors may assume an influential role in sexual development, and hormonal secretions may modify the extent to which various structures and even behavioral characteristics develop and are maintained.

Ovaries. A typical mammalian ovary is a solid, irregularly shaped structure indistinctly separated into an inner medulla and an outer cortex. The cortex contains numbers of ovarian follicles in various stages of development. *See* OVARY.

In certain fishes and in amphibians, snakes, and lizards, the ovaries are hollow, saccular structures. The cavity within the teleostean fish ovary is actually a closed-off portion of the coelom (body cavity) into which ripe ova are shed. This is not true of the saccular ovaries of other forms.

Cyclostomes. The adult female lamprey has a single ovary, representing a fusion of two, which runs the length of the body cavity, suspended from the mid-dorsal body wall by a single mesovarium. At the height of the breeding season it fills the greater part of the body cavity; ripe eggs are shed into the coelom and fertilization is external. The hagfish is hermaphroditic; the anterior part of the single gonad is ovarian and the posterior part testicular. Usually only one or the other region matures.

Fishes. The ovaries of most fishes are paired, although in some cases they have fused into a single organ. The large eggs of elasmobranchs are discharged from the anteriorly located ovaries directly into the body cavity. In oviparous and ovoviviparous species, following ovulation the ovarian follicles become transformed into corpora lutea, structures which presumably have an endocrine function and may play a role in the extended retention of eggs in the oviducts of oviparous species and of young in the uteri of ovoviviparous forms. Peritoneal folds form in connection with each ovary in teleosts, closing off all connection with the coelom. The anterior end of the ovarian cavity ends blindly, but in most cases continuations of the folds at the posterior end form an oviduct which opens directly to the outside. Ripe ova, sometimes numbered in the millions, are discharged into the central ovarian cavity, which is actually a part of the body cavity, and thence pass down the oviducts to the outside. The garpike *Lepisosteus* is the only ganoid fish with a saccular ovary. The ovaries are usually solid, flat elongate structures from which mature ova break out into the coelom.

Amphibians. Although the paired amphibian ovaries are saccular structures, ripe ova are liberated into the body cavity through their external walls. The shape of the ovaries varies with the shape of the body. They are long and narrow in caecilians, elongated to a lesser degree in salamanders, and short and more compact in frogs and toads. Fat bodies are associated with amphibian ovaries. They serve for the storage of nutriment and undergo profound changes during the year. A peculiar structure in the male toad, known as Bidder's organ, may under certain conditions develop into a true ovary.

Reptiles. The saccular ovaries of snakes and lizards are similar to those of amphibians, whereas turtles and crocodilians have solid ovaries. In snakes and lizards they are elongated but not symmetrically disposed. Only the yolk of reptilian eggs is formed in the ovaries, and this represents the true ovum. The size of the eggs is generally in proportion to that of the animal. In certain ovoviviparous snakes and lizards, corpora lutea form from ruptured follicles after ovulation and persist throughout pregnancy. These probably secrete a hormone necessary for maintenance of pregnancy.

Birds. Although both ovaries are present during embryonic development in most birds (except many birds of prey) the right ovary degenerates and only the left is functional. In birds, stalks extend from the surface of the ovary, and each stalk contains many ovarian follicles in various stages of development. A mature ovum escapes from the ovarian follicle through a preformed nonvascular band, the stigma or cicatrix, located on the surface of the follicle opposite the stalk. As in reptiles, only the yolk of the egg represents the true ovum. Increase in the number of hours of daylight stimulates ovarian activity in many birds. Even brief exposure to intense and bright light during hours of sleep increases production of eggs in the domestic fowl. This effect is undoubtedly mediated through stimulation of the pituitary gland. If the functional ovary of the domestic fowl is removed, the rudimentary right gonad will develop into a testislike organ, but without germ cells. *See* SEXUAL DIMORPHISM.

Mammals. The ovaries of mammals are located in the lumbar or pelvic regions and are small in comparison to the size of the body. The relationship of the microscopic mammalian ovum to the ovarian follicle differs somewhat from conditions in other vertebrates. Follicles in various stages of development, with the youngest near the surface of the ovary, are depicted in **Fig. 1**. At periodic intervals one or more follicles grow to maturity, rupture, and liberate their ova into the body cavity. In such animals as the rabbit, cat, and ferret, ovulation will not occur unless the animal copulates. Following ovulation certain cells of the follicle undergo a transformation and the entire structure becomes a more or less solid body, the corpus luteum.

If pregnancy does not occur, the corpus luteum persists only for a short time. If pregnancy does ensue, the corpus luteum usually persists throughout pregnancy. In either case it ultimately degenerates. The corpus luteum is of primary importance as an endocrine gland secreting a hormone called progesterone. *See* PREGNANCY.

Oviducts and associated structures. Oviducts, except in teleosts and a few other fishes, are modifications of Müllerian ducts formed early during
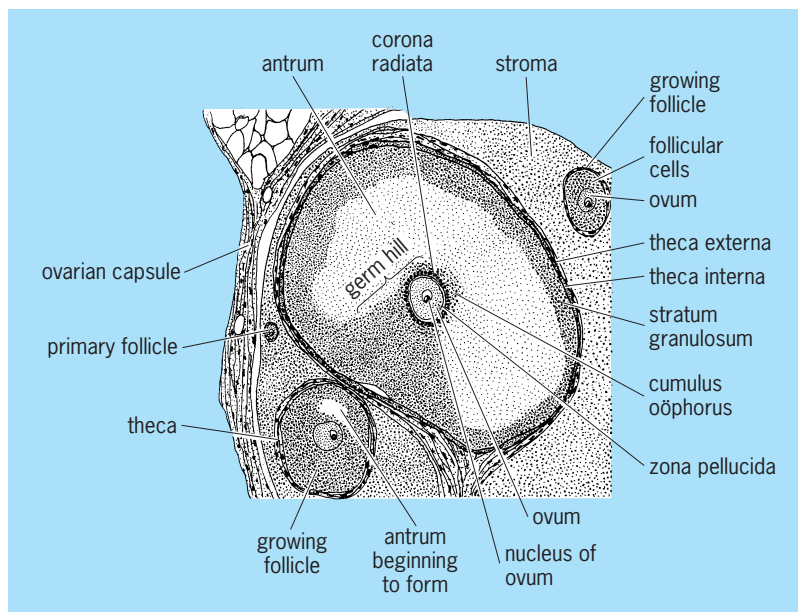
Fig. 1. Section of portion of cortex of rat ovary, showing ovarian follicles in various stages of development, the youngest being near the surface of the ovary. (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)

embryonic development. In most cases each duct is formed by an invagination of the peritoneum that covers the ventrolateral part of the opisthonephros or mesonephros. The edges of the groove fuse to form a tube which joins the cloaca or forms a uterus posteriorly and remains open anteriorly to become the ostium tubae. Although Müllerian ducts also form in the male, they ordinarily degenerate except for a few vestigial remnants. In some they persist as prominent but nonfunctional structures.

Cyclostomes. Oviducts are lacking in cyclostomes. Ova pass from the coelom through genital pores into a urogenital papilla and then to the outside.

Fishes. Much diversity exists in the oviducts of fishes. In some teleosts, and in a few other fish, eggs escape from the body cavity through modified abdominal pores. In elasmobranchs the two Müllerian ducts may fuse at their anterior ends so that only a single ostium tubae connects with the coelom. An enlargement, known as the shell gland, is present in the upper part of each oviduct. Beyond the shell gland the Müllerian duct enlarges on each side to form a uterus which opens into the cloaca.

The oviducts of most teleosts are short and continuous with the cavities of the saccular ovaries. It is doubtful whether they are true Müllerian ducts because they are formed in a different manner. A cloaca is lacking in teleosts, and the oviducts open independently to the outside. The two oviducts usually fuse, continuing posteriorly as a single structure which may open to the outside through a genital pore or else at the tip of a genital papilla. Most teleosts are oviparous, but many are ovoviviparous. Although the young may develop within the cavities of the ovaries, intrauterine development is more common. The size of the oviducts fluctuates markedly with the seasons. They are naturally largest during the breeding period.

Amphibians. Oviducts in amphibians are paired elongated tubes, each with an ostium tubae situated well forward in the body cavity. The posterior end of each oviduct is enlarged slightly to form a short uterus which opens into the cloaca. In almost all forms the uteri serve only as temporary storage places for ova that are soon to be laid. In some toads the two uteri unite before entering the cloaca by a common orifice. Marked fluctuation in size of the oviducts is apparent at different seasons (**Fig. 2**). During the breeding period they become elongated and coiled. The glandular lining secretes a clear gelatinous substance (jelly) which is deposited about each ovum as it passes down the oviduct. External fertilization is the general rule in frogs and toads, but in salamanders, with few exceptions, internal fertilization takes place. No copulatory organs are present. A diverticulum, the spermatheca, of the salamander's cloaca serves as a storage place for spermatozoa. The males deposit spermatophores (small packets of spermatozoa) which are taken into the cloaca of the female by muscular movements of the cloacal lips. Internal fertilization occurs in caecilians, the male of which has an eversible cloaca which may serve as a copulatory organ.

Reptiles. The paired oviducts of reptiles open into the coelom through large, slitlike ostia. Each oviduct is differentiated into regions which mediate different functions in forming the layers of materials deposited about the ova prior to laying. The eggshell, which is of parchmentlike consistency, except in some lizards and in crocodilians in which it is hard, is formed in the uterus. Fertilization is always internal in reptiles, copulatory organs being present in the male, with the exception of the primitive *Sphenodon*. Most reptiles are oviparous, but many snakes and lizards are ovoviviparous.

Birds. The right oviduct usually degenerates in birds and only the left one is functional (birds of prey being exceptions). The egg enters the coiled oviduct through the ostium and passes through a glandular region, an isthmus, and a uterus. Two layers of albumen are deposited about the ovum in the glandular region; inner and outer shell membranes and more albumen are laid down in the isthmus; the hard calcareous shell is formed in the uterus. In the hen these
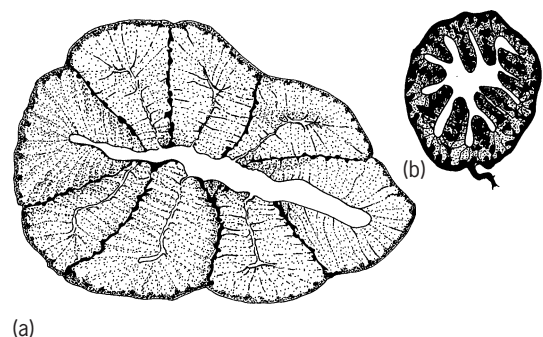
Fig. 2. Cross section of oviduct of the salamander *Eurycea bislineata*. (*a*) During breeding season. (*b*) After breeding season. (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)

processes take 21–23 h. Fertilization is internal in birds and, in a majority of species, is accomplished by cloacal apposition of the two sexes.

Mammals. Paired Müllerian ducts develop in all mammalian embryos. Each differentiates into an anterior, nondistensible Fallopian tube and a posterior, expanded uterus. In all mammals except monotremes the uterus leads to a terminal vagina which serves for the reception of the penis of the male during copulation.

The lower part, or neck, of the uterus is usually telescoped into the vagina to a slight degree. This portion is referred to as the cervix.

In the primitive monotremes a cloaca is present in the adult, and each uterus terminates independently in a urogenital sinus anterior to the region where ureters and bladder enter. Marsupials retain the primitive paired condition, but two vaginas open into a common urogenital sinus. In some mammals, the kangaroo for example, the vaginas fuse at their upper ends to form a vaginal sinus which extends posteriorly as a blind pocket or tube. It is sometimes referred to as a third vagina and may serve as the birth canal. Its posterior end connects with the urogenital sinus. In the event that this pouchlike structure has no opening, its wall ruptures at the time of delivery, thus permitting the young, lodged in it, to pass directly into the urogenital sinus. Placental mammals have a single vagina which represents a fusion of two; only the pika, which is a lagomorph, has a cloaca in the adult form.

The uterine portions of the Müllerian ducts fuse to varying degrees, resulting in different types of uteri (**Fig. 3**). The simplex type, found in apes and humans, represents the greatest degree of fusion. Only the bilaterally disposed Fallopian tubes indicate the paired origin of this type of uterus. Anomalous uteri of the duplex, bipartite, and bicornuate types are occasionally encountered in the human being.
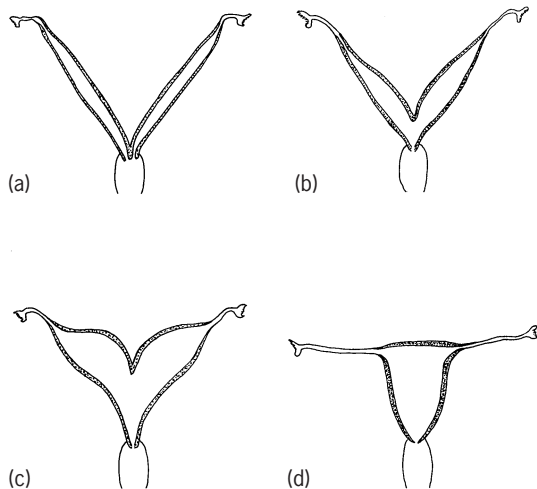
Fig. 3. Degrees of fusion of the uterine portions of the Müllerian ducts in four types of mammalian uteri. (*a*) Duplex (rat). (*b*) Bipartite (pig). (*c*) Bicornuate (horse). (*d*) Simplex (human). (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)
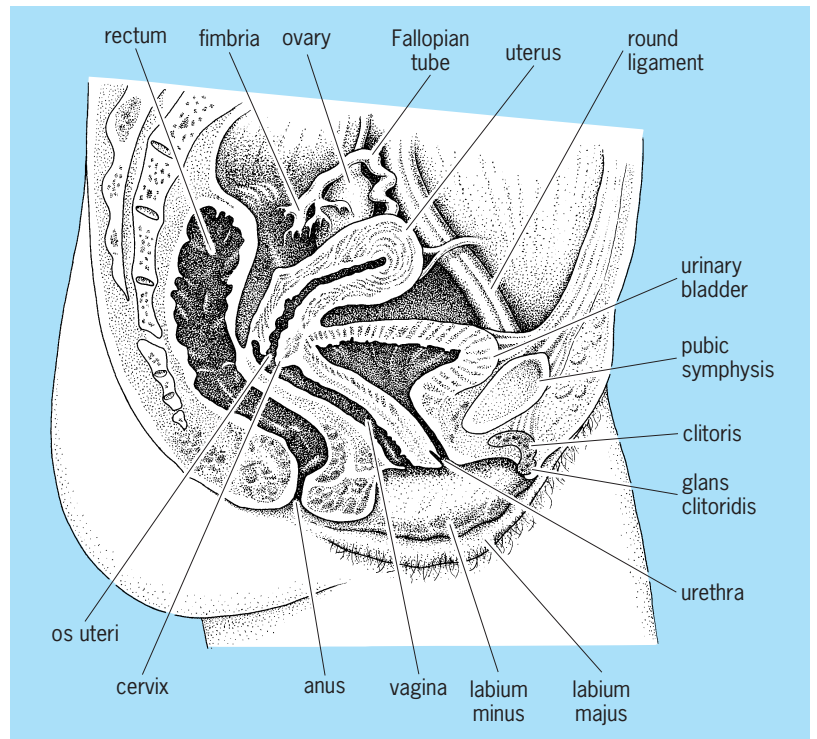
Fig. 4. Urogenital system of the human female. (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)

The urethra, coming from the bladder, may join the vagina to form a urogenital sinus, or vestibule, which opens to the outside. Often, a more or less complete fold of mucous membrane, called the hymen, marks the border of the vagina where it opens into the vestibule. The external part of the female reproductive system is called the vulva. In primates, two folds of skin, the labia minora, are located about the margins of the opening of the vestibule. In some apes, and in the human female (**Fig. 4**), two additional outerfolds, the labia majora, make up part of the vulva. A small erectile organ, the clitoris, homologous with the penis of the male, lies ventral to the vaginal orifice. It differs from the penis, however, in that it has no connection with the urethra except in a few forms such as rats and mice.

In addition to the above structures, certain glands, as well as rudiments remaining from the degenerated mesonephros, are associated with the female reproductive system. Glands of Bartholin correspond to Cowper's glands (bulbourethral glands) of males. They open into the vestibule near the hymen and secrete a clear, viscid fluid under sexual excitement. This serves as a lubricant during copulation. Paraurethral glands (of Skene), corresponding to the prostate glands of males, are occasionally encountered in females but are inconstant in appearance and of doubtful function. Small, mucus-secreting, vestibular glands, homologous with the glands of Littré of the male, are located around the opening of the urethra and on the clitoris. Among the remnants of the degenerated mesonephros that are associated with the reproductive organs of the female

is the epoophoron, a complex of degenerate anterior mesonephric tubules connecting to a persistent portion of the Wolffian duct. The paraoophoron, a similar group of more posterior tubules, is located farther caudad. Both epoophoron and paraoophoron are situated in the broad ligament near the ovaries. A canal of Gärtner, located in the wall of the uterus or vagina, represents a vestige of the Wolffian duct proper. None of these structures is functional.

The epithelium that lines the Fallopian tube consists of two kinds of cells: ciliated cells, which are most numerous on the fimbriated funnel surrounding the ostium tubae, and others of a glandular nature. No true glands are present here, however. The beat of the cilia is abovarian in direction, and this, together with the peristaltic muscular contractions, serves to transport ova down the Fallopian tube. In most species of mammals it takes 3 or 4 days for fertile ova to pass down the Fallopian tube to the uterus. If fertilization does not occur, the ova die and probably disintegrate before they have even reached the uterus.

Testes. The typical testis is a compact organ which varies greatly in shape in different groups of vertebrates. In all but a few primitive forms each testis is composed of numbers of seminiferous ampullae or seminiferous tubules which connect by means of ducts to the outside. Spermatozoa are formed within the tubules or ampullae. *See* TESTIS.

In seasonal breeders the size of the testes fluctuates with the seasons, being largest just before the breeding season. After this period they shrink to only a fraction of their former size. Each testis is suspended from the middorsal body wall by a membrane, the mesorchium.

Cyclostomes. The gonad of the male lamprey differs little in general appearance from that of the female. It is a single organ, representing a fusion of two, and is suspended from the middorsal body wall by a single mesorchium. Even when fully developed the testis does not become as voluminous as does the ovary in the female. Spermatozoa are discharged into the body cavity from which they escape via genital pores through a urogenital papilla. The hermaphroditic gonad of the hagfish has already been mentioned.

Fishes. In elasmobranchs, the paired testes are relatively small compact structures located at the anterior end of the coelom. In most other fishes they are elongated and often lobulated. Ducts transport spermatozoa to the outside. Male gonads vary greatly in size and are extremely large at the breeding season. In some teleosts, interstitial cells lie in spaces among the seminiferous ampullae or tubules; in others they form part of the walls of ampullae and show seasonal variation in appearance and activity.

Amphibians. The shape of the testes of amphibians is roughly correlated with that of the body. Thus, in caecilians, each testis is elongated and resembles a string of beads. The enlargements consist of masses of seminiferous ampullae connected by a longitudinal collecting duct. In salamanders the testes are shorter and irregular in outline; in frogs and toads they are small, oval, compact structures. A pronounced differ-

ence in size is apparent during the breeding season. Fat bodies are associated with the gonads of male as well as of female amphibians.

Reptiles. Reptilian testes are round, oval, or pyriform in shape and contain seminiferous tubules that are long and convoluted. In snakes and lizards one testis usually lies farther forward in the body cavity than the other. Periodic fluctuations in size of the testes are typical.

Birds. The round or oval shape of bird testes is characteristic. In the domestic fowl, the testes function throughout the year and no periodic variations in size are obvious. Most birds, however, are seasonal breeders, the testes enlarging conspicuously at the approach of the breeding season. Increase in the number of hours of daylight stimulates spermatogenesis in certain birds and hence brings about testicular enlargement, this generally occurring in spring.

Mammals. In all mammals except monotremes the oval-shaped testes move from their place of origin to the pelvic region, where they may remain permanently, or they may descend farther into a pouchlike scrotum. In many seasonal breeders the testes are located in the scrotum only during the breeding period. The scrotum serves as a temperature regulator, providing an environment for the testes several degrees below that of the body. This seems to be a requirement for normal development of spermatozoa. In marsupials the scrotum lies anterior to the penis but in other mammals it is posterior to that organ. In several species a relation between the number of hours of daylight and testicular activity has been demonstrated.

Male ducts. The ducts which in most vertebrates serve to transport spermatozoa to the outside of the body are the archinephric ducts or Wolffian ducts formed in connection with the opisthonephros or mesonephros, respectively. Their original function is elimination of urinary wastes. In some fishes and amphibians certain modified kidney tubules are employed in carrying spermatozoa from the testis to the archinephric duct. The tubules are known as efferent ductules, and the duct is termed the ductus deferens. The male ducts undergo profound changes in size in seasonal breeders (**Fig. 5**). Reproductive ducts are lacking in cyclostomes.

Fishes. A variety of conditions is encountered in the reproductive system of male fishes. In elasmobranchs small efferent tubules course from the testis through the mesorchium, connecting with some anterior kidney tubules along the medial border of the opisthonephros. The kidney tubules lead to the archinephric duct, which now serves almost entirely as a ductus deferens. The ductus deferens courses along the ventral side of the opisthonephros. In young specimens it is a straight tube with a urinary function; in older individuals it is highly convoluted. The posterior end is markedly dilated to form a seminal vesicle. The two seminal vesicles open into a common urogenital sinus which enters the cloaca through an aperture at the tip of a urogenital papilla. A pair of blind sperm sacs projects forward from the
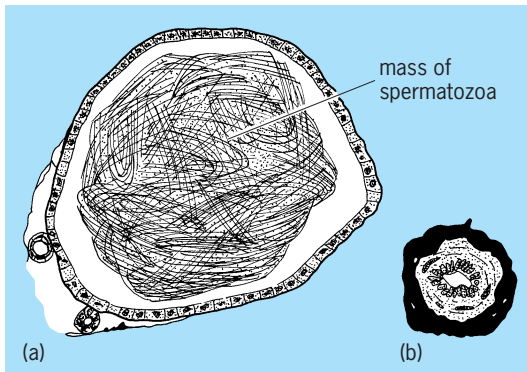
Fig. 5. Cross section of ductus deferens of male salamander *Eurycea bislineata*. From specimen obtained (*a*) just prior to breeding season; (*b*) several weeks after breeding season. (*After C. K. Weichert*, Elements of Chordate Anatomy, *3d ed., McGraw-Hill, 1967*)

ventral wall of the urogenital sinus. These sacs may be remnants of the Müllerian ducts which persist in the male.

In most other fishes the kidney ducts serve only for the passage of urinary wastes. The sperm duct, which is not a true ductus deferens because it is formed in a different manner, may be entirely independent of the kidney duct, although the two may have a common opening to the outside.

Amphibians. The arrangement of the male ducts in amphibians rather closely resembles that of elasmobranch fishes. In salamanders the ductus deferens courses outside the lateral border of the kidney (**Fig. 6**). Small efferent ductules from the testis join Bidder's canal, a duct lying medial to the kidney. This
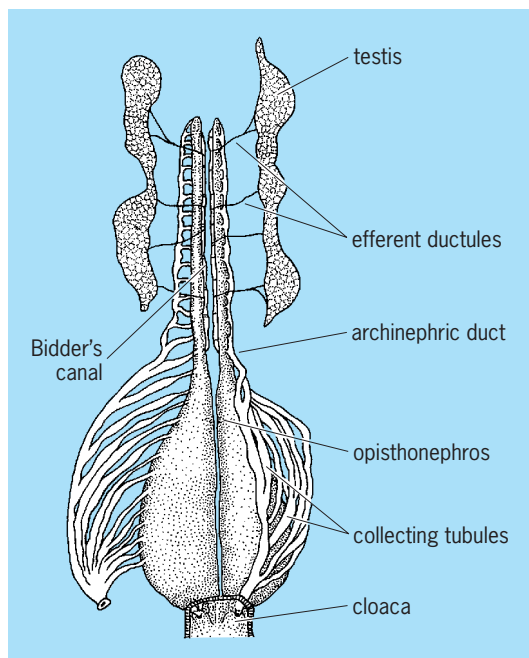
Fig. 6. Urogenital organs of male salamander, ventral view. The collecting ducts on the right side are shown detached from the cloaca and spread out for clarity. (*After C. K. Weichert*, Elements of Chordate Anatomy, *3d ed., McGraw-Hill, 1967*)

in turns connects to the ductus deferens by means of modified kidney tubules. In frogs and toads, both Bidder's canal and the ductus deferens lie within the opisthonephros. A Bidder's canal is also present in the female, but its function, if it has any, is obscure. In caecilians the longitudinal collecting duct, which connects the lobules of the testis, gives off small, transverse canals between successive lobules. These pass to the kidney to join another longitudinal duct which runs along the lateral edge of the opisthonephros. Spermatozoa then pass through a second series of transverse canals to join kidney tubules which transport them to the archinephric duct which joins the cloaca posteriorly. In males of several species of frogs and toads a dilation of the archinephric duct, or ductus deferens, as it nears the cloaca, forms a seminal vesicle in which spermatozoa may be stored temporarily. Ductus deferens, and seminal vesicles when present, undergo striking seasonal variations in size, being controlled by changes in the level of testosterone, the hormone secreted by the testes. The cloacal glands of male urodeles, which secrete a jellylike material used in forming the spermatophores deposited by the male, become markedly enlarged as the breeding season approaches.

Reptiles. When the embryonic mesonephros degenerates in male reptiles, its duct (archinephric or Wolffian duct) persists as the reproductive duct. The end near the testis becomes greatly convoluted and is called the epididymis. This connects with the testes by means of a few persistent mesonephric tubules which serve as efferent ductules. The remainder is the ductus deferens, sometimes straight, sometimes convoluted. In snakes and lizards this joins the ureter of the metanephros before entering the cloaca. In turtles and crocodilians these ducts open at the proximal end of a groove which carries spermatozoa to the free end of the penis. Seasonal variations under endocrine control are obvious in the epididymides and deferent ducts of most reptiles. In many, some of the posterior urinary tubules of the metanephros enlarge during the breeding season, producing an albuminous secretion which contributes to the seminal fluid. Glandular secretions from the cloacal walls in snakes and lizards pass into the grooves of the hemipenes. Nonfunctional Müllerian ducts, usually much reduced in size, commonly persist in male reptiles.

Birds. The reproductive ducts of male birds are essentially similar to those of reptiles but open independently into the cloaca. In the few birds that possess a penis, a groove on the upper surface carries spermatozoa to the apex. In some passerine birds, a nodule, composed of a tightly coiled portion of the ductus deferens, protrudes into the cloaca. The temperature of the nodule is somewhat lower than body temperature. This may be of importance in the maturation of spermatozoa.

Mammals. A few persistent mesonephric tubules connect the seminiferous tubules of each testis to a compactly coiled epididymis which is continuous with the ductus deferens. In man the portion of the
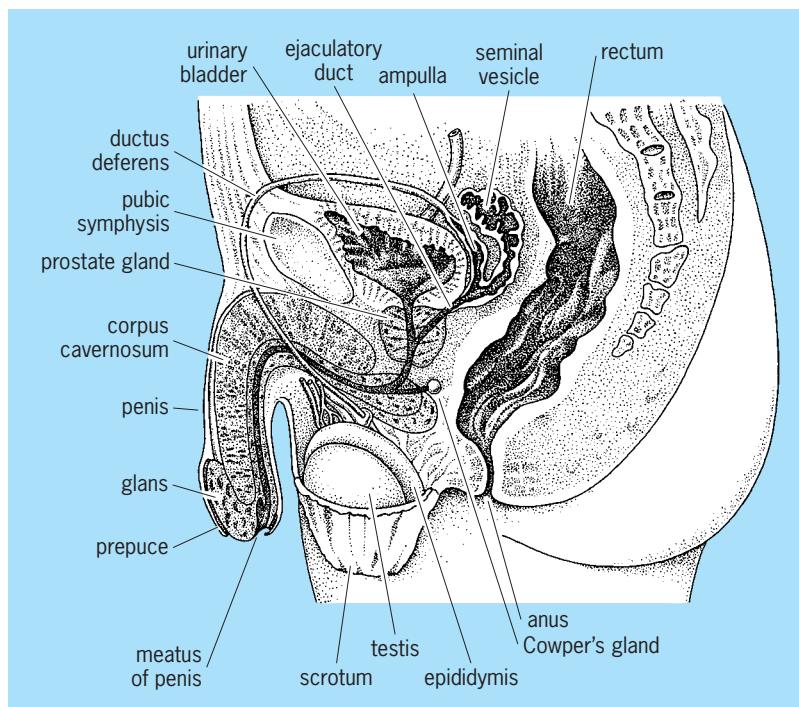
Fig. 7. Urogenital system of the human male. (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)

Wolffian duct included in the epididymis averages 20 ft (6 m) in length. Experiments have demonstrated that, during their passage through the epididymis, spermatozoa, although remaining quiescent, acquire the capability of becoming fully motile. In mammals having scrotal testes, the ductus deferens enters the body cavity lying between the peritoneum and the body wall. It crosses in front of the ureter, loops over that structure, and then courses posteriorly for a short distance before joining the urethra (**Fig. 7**). This peculiar arrangement is the result of the descent of testes from their original abdominal location. In many mammals the ductus deferens enlarges at its posterior end to form an ampulla. A glandular seminal vesicle may arise on each side as a saccular diverticulum of the ductus deferens. It does not store spermatozoa, but its secretion contributes to the seminal fluid in which spermatozoa are suspended. Seminal vesicles are absent in monotremes, marsupials, carnivores, and whales. The lower portion of the ductus deferens, between seminal vesicle and urethra, is often called the ejaculatory duct. The urethra, coming from the bladder, extends the length of the penis, opening at the tip of that organ through a small meatus. It serves both for passage of urine and seminal fluid. Accessory glands associated with the urethra include the prostate gland, which contributes to the seminal fluid; Cowper's glands, which secrete a clear, viscid fluid during sexual excitement; and small, mucus-secreting urethral glands. In some mammals, as in the rat, when the secretions of the seminal vesicles and prostate gland are intermingled they coagulate, so that just after mating a so-called copulation plug is present in the vagina. This aids in retaining spermatozoa in the reproductive tract of

the female. The development and functioning of the accessory sex organs in mammals are clearly under control of the hormone testosterone. A few remnants of the mesonephros may persist in male mammals in close relation to the reproductive systems. Among these are the paradidymis, a few aberrant ductules, and the appendix of the epididymis. Persistent homologs of the Müllerian ducts may also be present in males. The small appendix of the testis and the prostatic utricle represent such remnants.

Copulatory organs. Fertilization always takes place in a fluid medium. In many aquatic vertebrates external fertilization takes place, with water furnishing the medium in which sperm travel to gain access to ova. Terrestrial forms and even numerous aquatic species have internal fertilization. Secretions supplied by both sexes furnish the necessary fluid medium. In a number of terrestrial vertebrates, spermatozoa are transferred from male to female by cloacal apposition. In most, however, as well as in many aquatic species, intromittent, or copulatory, organs are employed to deposit spermatozoa, suspended in seminal fluid, within the reproductive tract of the female.

Fertilization usually takes place in the upper part of the oviduct. In forms in which the egg is surrounded by a shell, fertilization must take place before the shell has been formed about the ovum.

Fishes. In those fishes having internal fertilization, the copulatory organs are modifications of fins, pelvic fins in elasmobranchs and anal fins in teleosts. Claspers, which are modifications of pelvic fins, are used as intromittent organs by male elasmobranchs. Each clasper is essentially a scroll-like tube through which seminal fluid may be ejected with some force into the cloaca of the female. It is probable that only one clasper is inserted at a time. In teleosts having internal fertilization, the anterior border of the anal fin is elongated posteriorly to form an intromittent organ, the gonopodium. In some teleosts, modifications of the hemal spines of certain caudal vertebrae form a copulatory organ.

Amphibians. Although internal fertilization occurs in most salamanders, copulatory organs are lacking. By muscular action of the cloacal lips, the female is able to pick up spermatophores, or packets of spermatozoa, deposited by the male. The eversible and protrusible cloaca of some male caecilians may be used as a copulatory organ when the cloacae of both sexes are in apposition.

Reptiles. Sphenodon is the only reptile lacking copulatory organs. In others, two types of structures are evident. Snakes and lizards employ paired hemipenes, which are saclike structures, devoid of erectile tissue, lying under the skin adjacent to the cloaca. Each bears a groove for the passage of spermatozoa. These organs are everted during copulation and spermatozoa are deposited in the cloaca of the female. Hemipenes are not homologous with the single penis of turtles and crocodilians, which is basically similar to that of mammals, and contains erectile tissue which becomes distended with blood during sexual excitement. A homologous structure,

the clitoris, is present in rudimentary form in females. Paired, thickened ridges in the anterior and ventral walls of the cloaca are called corpora cavernosa. They are composed of connective and erectile tissue. A groove along the dorsal surface provides for the passage of spermatozoa. During the act of mating the corpora cavernosa are filled and distended with blood and the penis is firm and greatly enlarged. It is then said to be erect. This property of erectile tissue makes it possible for the penis to serve as an intromittent organ. Without erection, copulation is impossible. During erection the veins which normally drain blood from the erectile tissue become compressed. There is an increased supply of arterial blood supply accompanied by an impeded venous drainage. Return of the penis to its flaccid state after erection is termed detumescence.

Birds. Most birds copulate by cloacal apposition. A penis is present only in ducks, geese, swans, and ostriches. It is a single structure of the same type as that of turtles and crocodilians. A clitoris is present in females of these species. In many other birds a rudimentary penis can be identified.

Mammals. In monotremes the single penis lies on the floor of the cloaca. It is similar to the organ in turtles, crocodilians, and birds except that the groove on the dorsal side has become a closed tube. This is surrounded by another mass of erectile tissue known as the corpus spongiosum. The canal in monotremes possibly carries only seminal fluid since the urethra has a separate opening into the cloaca. The monotreme urethra thus differs from that of all other mammals. Marsupials lack a cloaca. The marsupial penis is covered by a sheath which opens to the outside of the body just beneath the anus. It may be protracted and retracted. The scrotum in marsupials is anterior to the penis. There are two erectile corpora cavernosa, separated by a septum, and a single erectile corpus spongiosum surrounding the urethra. Usually in marsupials three pairs of Cowper's glands open into the urethra at the base of the penis.

In higher mammals there is a tendency for the penis to be directed forward, and in all forms possessing a scrotum the penis is located anterior to the scrotum. In most mammals it lies within a sheath from which it can be protracted and retracted. In primates, the penis is permanently exserted. Its distal end bears a sensitive, swollen glans. Among mammals there are many variations in shape and structure of the glans, which contains erectile tissue continuous with the corpus spongiosum. In some forms, as in the cat, the glans bears numerous horny papillae or spines on its surface; these undoubtedly function as a sexual irritant. The glans is covered by a sheath, the prepuce, or foreskin, which may be retracted. Preputial glands secrete a sebaceous material called smegma about the base of the glans underneath the prepuce.

A bone is present in the penis of many rodents, carnivores, bats, whales, and lower primates. It is situated in the connective tissues between the corpora cavernosa and corpus spongiosum. In whales there is but a single corpus cavernosum. Anomalous development of the vertebrate penis is occasionally encountered. The most common anomalies are referred to as hypospadias and epispadias, in which the urethra opens on the under or upper sides, respectively, of the penis, rather than being continued to the tip of the glans.   Charles K. Weichert

Histology in Humans

The human reproductive system comprises many different organs, and therefore all the primary types of tissue are identified.

Male reproductive system. In the following sections the histological features of the testes, scrotum, excretory ducts, and auxiliary glands are discussed.

Testes. The testes have two important structural elements, the contorted seminiferous tubules and the interstitial cells which are situated between the tubules. The contorted seminiferous tubules have two kinds of cells, nutrient cells, which also serve as supporting cells (the cells of Sertoli), and germinal cells, which develop into spermatozoa by the process of spermatogenesis. The interstitial cells of the testis are seemingly modified connective-tissue cells. They resemble epithelial cells and hence are correctly designated epithelioid cells. They produce testosterone, the male sex hormone, which chemically is a steroid compound, in mammals and perhaps in other vertebrates. *See* SPERMATOGENESIS.

Scrotum. The scrotum consists of skin, smooth muscle, and connective tissue. The scrotal skin is rich in sweat glands; it is thus able to act as a thermoregulator for the testes. The human scrotal temperature is about $13°F$ ($7°C$) lower than that of the abdominal cavity. The smooth muscle is a "skin muscle" called the dartos tunic; its relaxation lengthens the scrotum and promotes loss of heat, whereas its contraction shortens the scrotum and reduces loss of heat. *See* SWEAT GLAND; THERMOREGULATION.

Excretory ducts. The excretory ducts which convey spermatozoa from the contorted seminiferous tubules to the urethra are either intratesticular (tubuli seminiferi recti and rete testis) or extratesticular (epididymis, ductus deferens, and ejaculatory duct). From the seminiferous tubules, spermatozoa enter the epididymis, a tubular structure consisting of head, body, and tail. The head of the epididymis has 12 or more efferent ductules (ductuli efferentes testis), the epithelium of which bears cilia, or little hairlike projections, which assist in moving sperm into the coiled part of the epididymis. The body and tail of the epididymis constitute the ductus epididymis, a coiled tube which, when uncoiled, is about 15–20 ft (4.5–6 m) long. The ductus deferens in humans is about 18 in. (45 cm) when uncoiled. Its proximal end is dilated to form the ampulla ductus deferentis.

Auxiliary glands. The auxiliary glands in humans include two seminal vesicles, one prostate, and two bulbourethral glands. Their secretions together with spermatozoa and the small amount of secretion from the excretory ducts constitute the semen. In the male, one ejection of semen, known as ejaculation,

has a volume of 3–5 ml, and each milliliter contains about 6×10^7 spermatozoa. The seminal vesicles, named in accordance with the erroneous belief that they are receptacles for spermatozoa (after death, sperm may be found in the vesicles), are hollow glands which lie along the back wall of the prostate. They are about 2 in. (5 cm) in length, and, developmentally, each gland is an outgrowth of the ampulla of the ductus deferens. The right and left seminal vesicles open into the prostatic urethra via the ejaculatory ducts. The prostate is a solid gland which surrounds the urethra (prostatic urethra) at its origin from the urinary bladder. It is about 1–1.5 in. (2.5–4 cm) in diameter, and has several small ducts which open individually into the prostatic urethra. The bulbourethral glands, or glands of Cowper, are about 0.25 in. (6 mm) in diameter. Each opens into the urethra at a site slightly below the prostate.

In such species as the rat, guinea pig, and rhesus monkey, the mixture of the secretions from the auxiliary glands coagulates, as in the case of ejaculated semen. In the mating of rats and guinea pigs, the coagulated semen produces the vaginal plug which temporarily partly occludes the vagina of the female partner.

Female reproductive system. The histological discussion of the female system includes the ovaries, uterine tubes, uterus, vagina, and external genitalia.

Ovary. The ovary (ovarium), or egg producer, is largely covered by peritoneum; it has follicles with egg cells (ova), a stroma of connective tissue, and, in maturity, corpora lutea.

Maturation of the ovum. The process of the origin, growth, and formation of the ovum in its preparation for fertilization is known as oogenesis. The young sex cell is an oogonium. Growth produces the primary oocyte; only the beginning of the first maturation division occurs before ovulation. After ovulation, the primary oocyte divides to produce two unequal elements: a secondary oocyte and a smaller body, the first polar body. Then the secondary oocyte divides to yield the mature ovum (rarely called ootid) and the second polar body. The mature ovum has the haploid number of chromosomes; the two polar bodies eventually disintegrate. *See* OOGENESIS.

Corpus luteum. The corpus luteum is a yellow endocrine body which originates at the site of a ruptured Graafian follicle and which produces progesterone, one of the female sex hormones. It originates by the metamorphosis of granulosa cells and thecal cells into lutein cells which are epithelioid. The corpus luteum of ovulation persists about 14 days, whereas the corpus luteum of pregnancy lasts several months. The degeneration of a corpus luteum produces a white, fibrous scar in the ovary, the corpus albicans.

Fallopian tube. The human Fallopian tube is a muscular structure about 4.5–5 in. (11.4–12.7 cm) long, and has three regional subdivisions: a funnellike upper part, the infundibulum (infundibulum tubae uterinae); a dilated part below the infundibulum, the ampulla; and a constricted part near the uterine junction, the isthmus. The mouth of the tube, or os-

tium, communicates with the peritoneal cavity, and is guarded by a fringe, the fimbriae tubae. Fertilization of the ovum occurs in the tube.

Uterus. The uterus is covered in part by peritoneum. It has an outer layer of smooth muscle, the myometrium, and an inner layer of connective tissue and epithelium, the endometrium. Thickness of the endometrium varies during the menstrual cycle. *See* MENSTRUATION.

Vagina. The musculomembranous organ situated between the cervix uteri and the external genitalia is the vagina. During copulation, it ensheathes the penis. Ventrally and dorsally, the anterior and posterior fornices, respectively, overlap the uterine cervix. In youth, the lower end of the vagina is constricted a bit by a membranous shelf of mucosa, the hymen vaginae.

External genitalia. The external genitalia include the clitoris, vestibule of the vulva (fossa vestibuli vaginae), paired lips of the vulva (labium majus pudendi and labium minus pudendi), and a pair of small vestibular glands (glandulae vestibulares minores). The large lips of the vulva are essentially folds of skin, whereas the small lips are covered with a mucous membrane. The vestibular glands secrete mucus, and, in cases of gonorrhea, may become inflamed and then form large cysts.   Lemen J. Wells

Embryology

The embryology of the reproductive system shows great similarities in all vertebrates, with the exception of teleost fishes, in which it is less specialized. It proceeds by three steps: sex determination; organogenesis of gonads; and later, during development, differentiation of the genital tract.

Sex determination. At the time of fertilization, the sex of each individual is genetically determined by certain genes carried by the sex chromosomes contained in the gametes. If both the male and the female germ cells have an identical X chromosome, the egg, upon fertilization, receives two X chromosomes (homozygous egg). If one parent germ cell contains an X and the other a Y chromosome, the egg cell receives an XY assortment (heterozygous egg). In mammals and in frogs, for instance, the male is the heterozygous XY sex. In birds and several urodeles the condition is reversed and the female sex is the heterozygous XY sex.

Sex determination should be considered as the first step in sexual differentiation of each individual. Techniques have been developed which permit direct study of the chromosomes, particularly the sex chromosomes, in dividing cells of any tissue. Chromosomal anomalies, such as the absence of one chromosome (for example, XO) or the presence of one or more supplementary chromosomes (for example, XXY) can be recognized in humans or in animals. Mosaicism results from the presence of two or more types of cells, for instance, cells with XX and cells with XY chromosomes, in the same individual. In such individuals the genetic basis of sex is disturbed, and embryological anomalies may occur. *See* CHROMOSOME; SEX DETERMINATION.
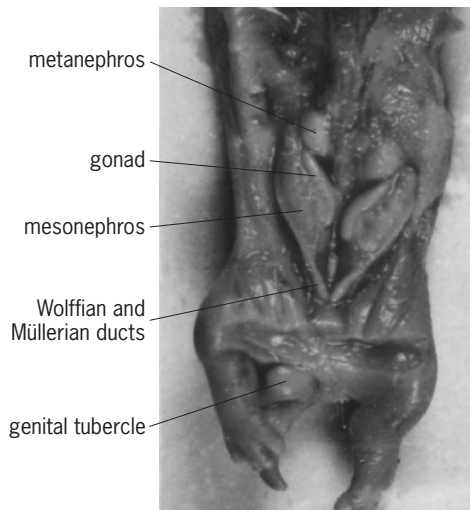
metanephros

gonad

mesonephros

Wolffian and
Müllerian ducts

genital tubercle

Fig. 8. Dissection of a 19-day-old rabbit fetus showing the undifferentiated condition of the genital tract.

ovary
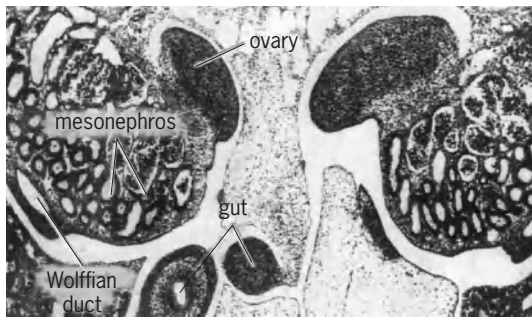
mesonephros

gut

Wolffian
duct

Fig. 9. Section through the body of a 17-day-old rabbit fetus, showing the ovaries at an early stage of differentiation located on the internal side of the mesonephroi.

Development of gonads. The development of the gonads is a progressive process, which may be divided into three main phases: (1) appearance of a genital ridge, (2) organization of an undifferentiated gonadal anlage, and (3) sexual differentiation of this primordium.

Genital ridge. The genital ridge appears on the mesolateral side of the mesonephros as a thickening of the coelomic epithelium covering the mesonephros. In reptiles, birds, and mammals it consists of a layer

female

Müllerian
duct

rete
ovarii

Wolffian
duct

uterus

urogenital
sinus

male

seminiferous
tubule

interstitial
cells

ovarian
cortex

Müllerian
duct

rete
testis

mesonephros

Wolffian
duct

seminal
vesicle

prostatic bud

urogenital
sinus
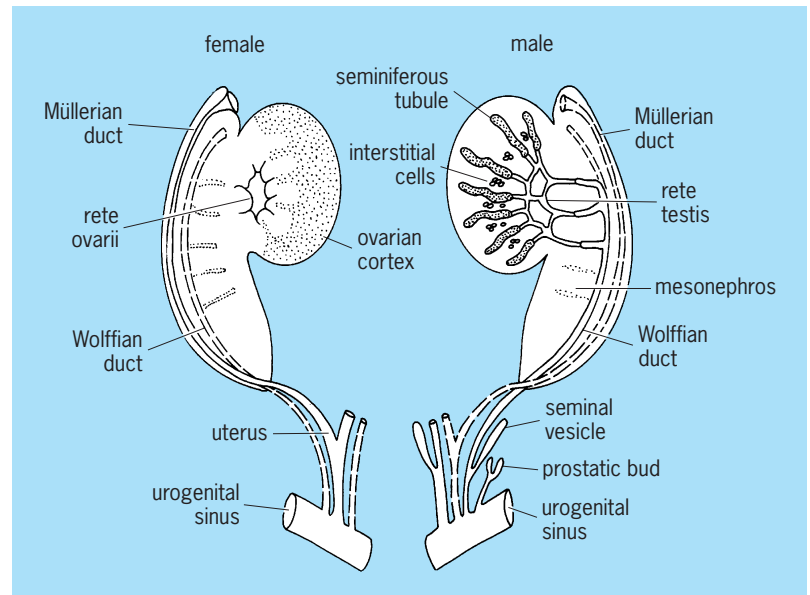
Fig. 10. Homologies in the development of the male and of the female genital tract of mammals. The gonads are shown on the inner side of the mesonephroi, and the ducts on their outer border. (*After A. Jost, General outline about reproductive physiology and its developmental background, in H. Gibian and E. J. Plotz, eds., Mammalian Reproduction, pp. 4–32, Springer, 1970*)

Role of urinary system. The development of the genital tract is intimately correlated with development of the urinary system (**Figs. 8–11**); the urinary system itself is derived from successive excretory organs. The first pair of kidneys, the pronephroi of the early embryo, develops mainly in the future neck region. The excretory (drainage) duct of the pronephros grows posteriorly and reaches the cloaca. Shortly afterward the second pair of excretory organs, the mesonephroi, differentiates approximately in the middle of the trunk, in connection with the primitive pronephric duct. The pronephros then retrogresses, and its duct becomes the mesonephric duct, also called the Wolffian duct. In higher vertebrates (birds and mammals) the mesonephroi are later replaced by a third pair of kidneys, the metanephroi.

Among the vertebrates, except for the teleost fishes, the sex glands differentiate in connection with the mesonephros. They retain these connections in male amphibians, in which the mesonephros remains the adult kidney (Fig. 11). The testicular tubules are connected with specialized mesonephric tubules. An adult male frog displays conditions which are similar to those found in the early bird or mammalian embryo. *See* URINARY SYSTEM; UROGENITAL SYSTEM.

genital part
of the
kidney

testis

Wolffian
duct

Müllerian
duct
(oviduct)

collecting
duct of
the rete

lumbar
kidney

collecting
ducts of
the kidney

(a)

genital part
of the
kidney

Müllerian
duct
(oviduct)

Wolffian
duct

ovary

lumbar
kidney

(b)

Fig. 11. Drawings of the right half of the reproductive system of the newt *Triton* (*Triturus*) *cristatus*. (*a*) Male. (*b*) Female. (*After J. de Beaumont, Wilhelm Roux Arch. Entwicklungsmech. Organ., 129:120, 1933*)

of enlarged coelomic cells, which at first is two or three cells thick. This contrasts with the flat cells of the other parts of the mesonephric coelomic wall. Primordial germ cells, the cells which will give rise to the germ cells, come in close contact with the so-called germinal epithelium and even penetrate between its cells. These primordial germ cells have an extragonadal origin.

Extragonadal origin of germ cells. In the chick embryo, the primordial germ cells are first seen as large undifferentiated cells, located in the germinal crescent, a part of the extraembryonic area of the blastoderm situated anteriorly to the embryo's future head. They reach the level of the gonadal anlage by way of the bloodstream. In sauropsidians and mammals their extragonadal origin has also been established. They originate in the primitive gut and usually migrate through the dorsal mesentery toward the gonadal area. In mammals, and in human embryos, they can be traced with histochemical techniques, showing their high content of alkaline phosphatase. In the frog the primitive germ cells have been traced from the very beginning of the cleavage of the egg; their early destruction led to the development of sterile gonads.

Undifferentiated gonad. In Amphibia, the primordial germ cells are first disposed in clumps near the inner border of the mesonephros. The genital ridge is formed when these primordial germ cells protrude into the abdominal cavity with some connective tissue and the covering coelomic epithelium. The anlage is invaded by clusters of cells coming from the mesonephric blastema. The gonadal anlage has then attained a stage known as the indifferent stage, which is the same irrespective of the genetic sex of the animal.

The indifferent sex gland consists of the coelomic epithelium at the periphery, the cells of which are flat and scattered and cover the voluminous primitive germ cells. Some connective tissue surrounds the central mass of small cells which came from the mesonephric blastema. The outer zone is often called cortex, and the inner part is called medulla. Both constituents participate in the formation of the definitive sex glands.

In mammals and in birds the undifferentiated primordia of the gonads are also composed of three major components: a superficial epithelium, inner cell masses, and germ cells. The limit between the superficial layer and the inner cells usually disappears and becomes poorly defined. Therefore, the origin and the significance of the constituents of the indifferent gonad gave rise to various interpretations which still prevail. In order to understand the issue, it must be kept in mind that in the mature female animal the ova are released from the ovary through its surface, into the abdominal cavity, in a centrifugal way, and that they are then collected by the tubes. Accordingly, in the development of the ovary the superficial layers play the major part. On the contrary, in males the sperm cells are released into an intratesticular and complex system of channels derived from the mesonephros, in a centripetal way. Accordingly,

the seminiferous tubules develop in the center or in the depth of the organ (Fig. 10).

This explains why embryologists more than a century ago introduced the concept that the undifferentiated gonadal primordium was hermaphroditic. It was supposed to comprise a superficial female component and a central male component, the latter originating from the mesonephric renal structures. Only one of these sexually specialized parts was assumed to differentiate in each sex, while the other degenerated. Another long-lived classical theory assumed that the superficial epithelium of the genital ridge proliferated so-called sex cords into the underneath undifferentiated mesenchyme. A first set of these cords was held to constitute the male medullary cords, while in females the superficial epithelium, the cortex, proliferated a second set of sex cords, which developed in an ovarian cortex, after the involution of the medullary cords.

It is still difficult to assess the exact history of the various cell types involved in the development of the sex glands, since no cellular markers are available. In any case, there is no indication of a hermaphroditic constitution of the primordium, since no separate male or female component can be distinguished.

Sexual differentiation of the gonads. Testes differentiate earlier than ovaries, and for a while female embryos can be recognized only because they do not develop testes.

The initial step of testicular differentiation consists of the appearance of a new type of cells in the undifferentiated gonad, the primordial Sertoli cells. These cells aggregate together, and with germ cells they delineate the early testicular seminiferous cords. (In the adult testis, the Sertoli cells support the germ cells in the seminiferous tubules.) In the meantime, at the surface of the testis, the cells become rarefied, flatten, and form the albuginea of the testis. A second crucial step in testicular differentiation is the differentiation of the Leydig cells, between the seminiferous cords. These cells produce the steroid male hormone. *See* HORMONE; STEROID.

At the same time that these events occur in males, the presumptive ovaries in females maintain a morphologically undifferentiated condition. The first clear-cut sign of ovarian differentiation is the onset of the meiotic prophase. In fetal ovaries, the germ cells enter the preparatory phases of chromosomal reduction (or meiosis), a nuclear process which will be completed during adulthood when oocytes are released from the ovary. A long phase of nuclear quiescence takes place between the fetal preliminaries and their completion at the time of ovulation in adulthood. *See* MEIOSIS.

Meiosis seems to be a critical phase of ovarian development. Many germ cells entering meiosis degenerate; a few of them survive and become surrounded by follicular cells. The primary ovarian follicles are constituted in that way. It seems very likely that the follicular cells are homologous to the testicular Sertoli cells. The question as to why the ovarian germ cells enter the meiotic prophase whereas in males of the same age they do not, is not yet

solved, though it has been suggested that a meiosis-inhibiting substance is produced by the testis and a meiosis-inducing substance by the ovary. The reason why so many germ cells degenerate at incipient meiosis also remains unknown.

Gonadal differentiation is well established in reptiles or birds before hatching, or in placental mammals at birth. In the marsupials, such as the opossum, intrauterine pregnancy is very short and is followed by a period of development in the marsupial pouch. Sexual organogenesis takes place after birth during pouch life.

Exceptions to general scheme. Many animals have developmental patterns that deviate from the scheme outlined above. In a great number of birds, only one ovary becomes functional in the adult. In the female chick embryo, for instance, dissymmetry of the gonads is conspicuous at an early stage, because the left anlage is much larger than the right one. The right gonad, composed of some vestigial medullary tubules, remains, and is a nonfunctional rudiment in the hen. It may develop as a small testis if the left ovary is removed.

In toads of either sex, the anterior part of the gonadal anlage acquires the structure of a persistent rudimentary ovary, Bidder's organ, above the functional gonad. In adult animals, Bidder's organ may develop into a functional ovary when the actual gonads are surgically removed. Males deprived of their testes undergo a slow feminization and may lay eggs.

In frogs, there are racial differences. In animals living in cold countries or at high altitude, males and females both differentiate early. In most frogs from temperate climates, the process of sexual differentiation is different. During the months before metamorphosis, in all individuals, whatever their genetic sex, the gonads first differentiate in a feminine way, since testicular differentiation takes place and several germ cells increase in size and enter the meiotic prophase. However, at metamorphosis, in half the individuals the largest germ cells degenerate and testes differentiate, while the other animals follow the normal feminine line. Thus, males suffer from a delayed testicular organogenesis. Animal species or strains which show such a transitory feminine phase of the male gonads are known as indifferent strains. Several amphibians, teleosts, and cyclostomes also develop indifferent strains.

Control of gonadal differentiation. In mice and in humans, evidence that the presence of a Y chromosome is a prerequisite for morphological differentiation of testes has accumulated. Thus, exceptional XO mice are females and XXY mice are males. The problem of how the sex genes control gonadal organogenesis has not yet been fully explained on cellular or biochemical bases. *See* SEX DETERMINATION.

The effects of administering hormones to developing embryos are profound, but vary among organisms (**Table 1**). Such experiments do not, however, warrant the generalization that sex hormones control sex differentiation. In lower vertebrates, sex differentiation may be influenced by external factors. Experiments made on turtles, alligators, and the

TABLE 1. Effect of administered sex hormones on developing gonads

Animal	Estradiol (on males)	Testosterone (on females)
Fishes		
Lebistes	Feminization	Masculinization
Amphibians		
Frogs, various species	Feminization (low dosage) Masculinization (high dosage), paradox effect	Masculinization
Alytes	Feminization	No effect
Xenopus	Complete feminization	No effect
Bombina	No effect	No effect
Amblystoma	Complete feminization	Feminization, paradox effect
Pleurodeles	Complete feminization	Feminization, paradox effect
Birds		
Chick	Feminization	Slight masculinization
Duck	Slight feminization	No effect
Marsupials		
Opossum	Feminization	No effect
Placental mammals		
Several species	No effect	No effect

newt *Pleurodeles* showed that breeding the eggs at different temperatures resulted in the development of young which were all of the same sex, depending on the temperature. Genetically, male *Pleurodeles* which were sex-reversed into functional females by breeding at high temperature could be mated with normal males and the progeny could be studied.

Sex differentiation of the gonads is controlled by genetic factors and the question arises of how and in which cells these genetic factors express themselves. As mentioned above, the gonadal primordium has long been assumed to be hermaphroditic, and it was accepted that the genetic factors gave prevalence to one sex component over the other, possibly via the dominance of one of the two antagonistic sex inductors produced by the sexually opposed gonadal components. In 1965, Jost introduced the concept that the sex glands were programmed to develop according to the homogametic sex (female in mammals, male in birds) and that in the heterogametic sex a special triggering mechanism imposes alternate sexuality on the primordium. This scheme was largely accepted, but the trigger has yet to be identified. A theory propounded in 1975 by Stephen Wachtel and others had some success. It proposed that the HY-histocompatibility antigen (discovered by grafting male skin in female mice and characteristic of individuals possessing the Y chromosome) was a triggering protein responsible for the differentiation of the sex gland in the heterogametic sex. This concept does not, however, fit with Humphrey's fundamental experiments mentioned above or with other observations. The chemical control of gonadal sex differentiation remains to be definitively elucidated.

Gonadal abnormalities. Important abnormalities include absence of the gonad, hermaphroditism, and sex reversal.

Absence of the gonad. The entire gonad may fail to develop or may retrogress at very early stages, leaving only some indistinct remnants which are impossible to recognize as testes or as ovaries. This condition, called gonadal agenesis or gonadal dysgenesis, is known among humans and other animals. In humans, gonadal dysgenesis is a part of the polymorphic Turner syndrome, which occurs in patients having the abnormal chromosomic formula XO. Since it also occurs in patients who have other normal or abnormal chromosomal formulas, the exact correlation between the sex chromosomes and gonadal dysgenesis is not yet clear.

Hermaphroditism. The presence of male and female gonadal tissue in the same individual is not normal among vertebrates except in some species of fishes. It can occur as an abnormal condition more or less frequently among other vertebrates. It is rather frequent in some frogs and toads, and was considered to be the result of an incomplete dominance of either the cortex or the medulla of the indifferent gonad. It might also, however, result from incomplete expression of the testicular determining system. In humans, several cases of true hermaphroditism have been reported in sterile individuals. Several cases have been correlated with chromosomal abnormalities such as XX/XY mosaicism, and in one case it was observed that the ovarian tissue contained mainly XX cells and the testicular tissue XY cells.

Genital tract development. The sex ducts become sexually specialized some time after the sexual differentiation of the sex glands. The male or female conditions develop from an indifferent condition which is identical in both sexes in early stages.

Indifferent stage. The gonads are already recognizable as presumptive ovaries or testes and are located on the anterior part of the mesonephros. The mesonephric or Wolffian duct is the ureter, and opens posteriorly into the cloaca in lower vertebrates or into the urogenital sinus in mammals. Another duct, the oviduct or Müllerian duct, parallels the mesonephric duct. The oviduct arises from a funnel which opens into the coelomic cavity. The blind end of this primordium proliferates and extends progressively caudally.

In selachians and urodeles, the funnel from which the Müllerian duct originates corresponds to a pronephric nephrostome, the coelomic opening of the primitive urinary tubules. Because the pronephros is located near the neck of the larva, and because the ostium of the oviduct retains this position, the oviducts open into an anterior part of the body cavity (Fig. 11).

In birds and mammals the origin of the oviduct from pronephric remnants is not as clear, but it is obvious that the oviduct develops in the region of the nephric field. The early funnel is located on the top of the mesonephros, and the ostia tubae open above the ovaries.

Differentiation of female genital tract. The Müllerian ducts differentiate into the female ducts. Depending upon the animal species, either a simple secretory oviduct (amphibians) or a more complicated structure de-
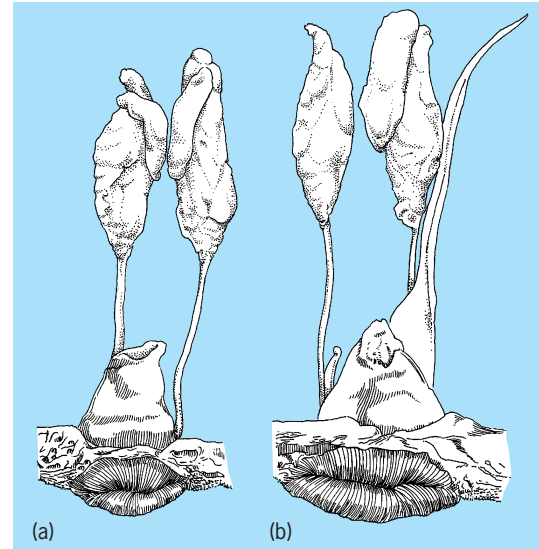
Fig. 12. Reproductive system of normal chick embryos, shortly before hatching. (*a*) Male (two testes present). (*b*) Female (sole left ovary and left oviduct). (*After E. Allen, ed., Sex and Internal Secretions, 3d ed., Williams and Wilkins, 1961*)

velops from this simple unicellular layered duct. It is divided into several specialized sections involved in the secretion of albumen or shell, as in selachians, reptiles, and birds. In those birds in which the female has only one functional ovary, only one oviduct develops (**Fig. 12**).

In mammals, the Müllerian ducts give rise to the oviduct, or tube, and the uterus. Usually the embryonic Müllerian ducts fuse posteriorly, but the extent of fusion is variable. In humans and monkeys a single uterus is formed from the fused part of the ducts; the tubes correspond to the nonfused part. In rodents, the two Müllerian ducts fuse only in the upper vagina, and form two uterine horns and two tubes. In ruminants, an intermediary condition is realized in which the uterus is composed of an inferior stem and two horns which largely communicate (Fig. 3).

The posterior part of the mammalian female genital tract is constituted by the vagina. The embryology of this organ displays great variability from one animal species to the other, and this makes an accurate interpretation difficult.

In the undifferentiated stage, the Müllerian ducts terminate blindly at the wall of the urogenital sinus (or embryonic urethra), between the two Wolffian ducts which open into the urogenital sinus. They may retain this primitive connection, as is the case in the rabbit (**Fig. 13***a*), and the urogenital sinus then becomes a urethrovaginal duct.

In other animals a cord of cells, the vaginal cord, may detach from the dorsal wall of the urogenital sinus and grow progressively caudally below the end of the Müllerian parts (Fig 13*b*). In the rat or the mouse this sinusary vagina finally opens independently of the definitive female urethra in a vaginal opening (Fig. 13*c*). In other animals, such as the mare, the vaginal cord remains connected with the posterior part of the urogenital sinus. The
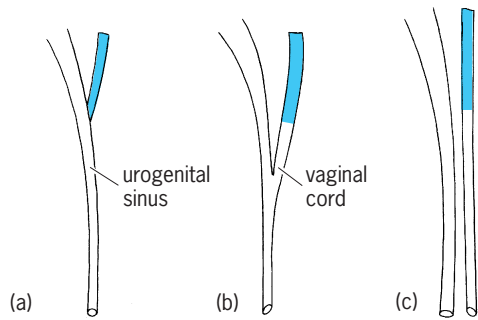
Fig. 13. Schematic interpretation of vaginal organogenesis in three species. Müllerian ducts are shaded. The three conditions are explained in the text.

vestibulum then remains as a common opening to both the vagina and urethra.

Comparative embryology of the vagina helps in understanding abnormalities of the human genital tract. Absence of the vagina and opening of the female ducts into the urethra may result from the preservation of the primitive connections.

The Wolffian or mesonephric duct of the female embryo is never incorporated into the genital tract. In animals that keep their mesonephros as a functional kidney in adulthood (selachians and amphibians), the mesonephric duct remains as the excretory channel (Fig. 11). In female birds and mammals, the mesonephric duct, as well as the mesonephros, disappears, leaving only some minor vestiges.

Differentiation of male genital tract. The male sex ducts are derived from the mesonephric or Wolffian ducts. In selachians and amphibians the mesonephros remains the functional kidney, and the mesonephric ducts function as pathways for urine as well as for sperm. Even in such animals, however, the anterior part of the kidney often becomes specialized as the sexual part. The posterior part of the kidney then produces urine, and several excretory tubes may bring this urine directly to the cloaca. In such cases the Wolffian duct of the male is only a genital canal, as in the newt *Triton cristatus* (Fig. 11).

The Wolffian ducts of birds differentiate into an undulated vas deferens, whereas in mammals the ducts differentiate into the epididymis at one end and the seminal vesicles at the other (Fig. 10).

The Müllerian duct has no function in males, and as a rule it disappears. In some amphibians, such as newts and toads, the oviducts persist in a rudimentary condition (Fig. 11); they may be activated under appropriate hormonal stimulation in adult males.

In male mammals the urogenital sinus becomes the definitive male urethra. Several accessory glands, such as the prostatic glands, bud from it, and display great variations from one animal species to another, but all open into the urethra (Fig. 10).

Copulatory organs. Copulatory organs are well developed in mammals and reptiles. In selachians they are often a specialized part of the fins, the claspers. As a rule no copulatory organ is present in amphibians and birds; only a few male birds, such as the duck, possess a penis.

In mammals the copulatory organ develops from an undifferentiated genital tubercle which is identical in both the male and female embryos and lies above the opening of the urogenital sinus. The male penis encloses the penile urethra and increases in size, whereas the homologous female clitoris remains more like the primitive tubercle.

Hormonal control. The genital tract differentiates after sexual differentiation of the sex glands, and time relationships support the view that sexual specialization of the genital tract is controlled by hormones produced by the developing sex glands. This has been shown experimentally by depriving embryos or young animals of their gonads. It was established that in the gonadless body, the genital tract becomes identical whatever the genetic sex of the individual. This means that in the absence of the sex glands, sexual dimorphism, which is a characteristic feature of almost all vertebrate species, does not appear. This identical, hormoneless aspect of the body is known as the neutral form. However, the gonadless sexual type is not the same in mammals and birds.

In mammals, gonadless sexual organization has been studied in rabbit fetuses which were surgically castrated in utero. It was found that the neutral condition is essentially feminine. This means that no embryonic gonad is necessary to produce the feminine sexual structures. In males, the testes prevent persistence and development of the female structures (tube and uterus) and impose masculinity on the whole genital apparatus (**Fig. 14**).

In a duck embryo castrated by a beam of x-rays, E. Wolff noticed masculine differentiation of such sex characters as the penis or the voice organ (syrinx). During normal sexual differentiation the embryonic ovary prevents such characters from becoming masculine. In addition, the neutral aspect of the
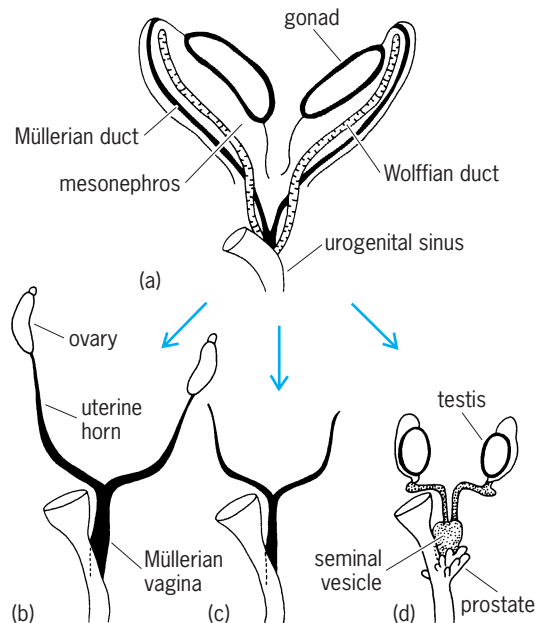
Fig. 14. Sexual differentiation of the genital tract of the rabbit fetus: (*a*) undifferentiated condition; (*b*) female; (*c*) castrate; (*d*) male. (*After A. Jost, Mem. Soc. Endocrinol., no. 7, 1960*)
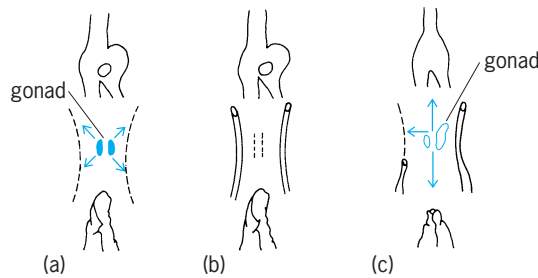
Fig. 15. Evolution of sex characters in the duck embryo: (*a*) male, **(*b*)** castrate, **(*c*)** female. Top sketch, voice organ or syrinx, well developed in males. Middle sketch, Müllerian duct (compare with Fig. 13); arrows symbolize inhibiting gonadal actions. Bottom sketch, genital tubercle. (*After E. Wolff, Compt. Rend., 229:428, 1949*)

genital tract is also characterized by the presence of oviducts. These structures are normally absent in males in which they are inhibited by an embryonic testicular hormone (**Fig. 15**). A similar trend toward maleness was observed in young newts, *Triturus cristatus*, castrated before sex differentiation. In that species females are heterozygous, as in birds.

It appears that in some vertebrates the neutral type is feminine (mammals) and in others predominantly masculine (birds and newts). In the absence of gonadal hormones, a basic developmental program prevails that corresponds to the homozygous sex, female in mammals, male in birds and newts. In the heterozygous sex, the gonads counteract this program and impose respectively maleness or femaleness in mammals and in birds and newts.

As mentioned above, in some amphibians complete sexual reversal of the gonads can be obtained under hormonal influences (testes developing in genetic females or ovaries in genetic males). In such animals the genital tract develops in agreement with the actual sexual structure, and function of these gonads and fertility are ensured.

Such a complete sexual reversion has not yet been produced in higher vertebrates. The condition prevailing for placental mammals deserves special attention. Sex hormones administered to the pregnant mother animal or injected into the fetus itself modify the development of the genital tract with the exception of the gonads themselves. The young thus become intersexes and definitely abnormal. Estrogenic hormones given to pregnant rats or mice feminize the male fetuses to a large extent. In addition to unaltered testes, large parts of the Müllerian ducts can be retained and a vagina and feminine external genitalia develop. Under the influence of androgenic hormones, female fetuses acquire several masculine structures, such as male external genitalia, urethra, accessory sex glands, vasa deferentia, and seminal vesicles. But they retain their ovaries, tubes, and uterus. Similarly, in castrated rabbit fetuses given testosterone or similar hormones, masculine organs develop but the Müllerian ducts are not suppressed. Androgenic steroid hormones do not inhibit the Müllerian ducts. This suggests that the fetal testis produces two kinds of substances, one stimulating the development of masculine char-

acters, another inhibiting the female ducts. The dual nature of the hormonal control of the differentiation of the genital tract is now established. In addition to steroidal androgens such as testosterone, the fetal testis produces a Müllerian inhibitor called anti-Müllerian hormone or Müllerian-inhibiting substance. This hormone, as produced by the fetal calf testis in cell cultures, was shown to be a glycoprotein of a molecular weight of 124,000. The two fetal testicular hormones are produced by two different cell types, the Sertoli cells for the anti-Müllerian factor and the Leydig cells for androgens.

Another line of important facts concerns the target cells for androgens. It is established that androgens can masculinize undifferentiated primordia only if the cells are provided with appropriate androgen receptors. Moreover, in several target organs testosterone has to be converted into dihydrotestosterone by the intracellular enzyme 5α-reductase, before it acts in the nucleus. Absence of receptors or of proper conversion of testosterone results in the same effect as absence of hormone. A partial cellular defect produces an intermediary degree of masculinization.

Abnormalities of genital tract. Many types of genital developmental abnormalities are known in animals and humans. Genetically female fetuses may be partially virilized (masculine or ambiguous intermediary aspect of the external genitalia) under the influence of androgenic substances either given to their pregnant mother or produced by their own abnormal adrenals (adrenogenital syndrome).

Genetically male fetuses may suffer from incomplete masculinization by their own testes. The discovery of a dual hormonal control of the differentiation of the genital tract and of the necessity of hormonal receptors in target tissues gave explanations for many sex abnormalities. Males can be incompletely masculinized in the following cases:

1. The testes are absent. In this case the genital tract is completely feminine, but remains infantile because no ovarian hormones are produced. This condition resembles that of castrated rabbit fetuses.

2. The testes do not produce testosterone, or not enough. Several enzymes are involved in the synthesis of testosterone from acetate or cholesterol, and several enzymatic defects depending on discrete genes are known. This results in a condition of pseudohermaphroditism.

3. The target cells lack androgen receptors. Since androgens are not effective, no male sex duct (vas deferens), and no male organs are present, except endocrinologically normal testes. Since the Müllerian inhibitor is produced, the tubes and uterus do not persist. This condition prevails in the syndrome called testicular feminization in humans or in animals.

4. The target organs lack 5α-reductase, a defect that does not permit appropriate masculinization (such as at the level of external genitalia). This is another type of pseudohermaphroditism.

5. The testes fail to produce the Müllerian inhibitor (or possibly the Müllerian ducts lack

receptors). In these subjects testes and male sexual characters are well developed, but a set of female organs persists (persistent Müllerian ducts).

6. In true hermaphrodites, when testicular tissue is present only on one side of the body, masculinization of the internal genital structures may be limited to the same side.

In conclusion, multiple causes of genital abnormalities are known. In females, they usually result from androgens that should not be present. In males, partial or total lack of masculinization may be due either to lack of androgen secretion or action, or to lack of Müllerian inhibitor secretion or action, or to both.   Alfred Jost

Physiology

The physiological process by which a living being gives rise to another of its kind is considered one of the outstanding characteristics of plants and animals. It is one of the two great drives of all animals: self-preservation and racial perpetuation. In contrast to other physiological processes, reproduction in vertebrates can be achieved only by participation of two individuals, the male and the female. Each produces germ cells called gametes. The male produces spermatozoa and the female ova, which carry the biochemical information arranged as genes in the chromosomes for the transmission of inherited characters. No matter how discrepant the pairing gametes may be in size and in form within a species, they contribute the same number of chromosomes. Although reproductive devices are quite different from one species to another, all serve one end, the bringing together of the spermatozoon and the ovum, each containing half the number of chromosomes of the parent cells. After the union of the chromosomes at fertilization, the newly constituted embryo possesses the number of chromosomes of its race and then divides, differentiates, grows and develops into an individual either outside the mother (oviparity) or inside the mother (viviparity). The obvious advantage in bisexual reproduction lies in the fact that young produced from the mingled genes of two ancestral lines will not be a direct copy of either parent, but will represent different combinations of ancestral traits with great potential for the survival of the fittest. Reproduction is facilitated or inhibited by environmental and nutritional factors as well as by all the other physiological activities, but it is controlled predominantly by the endocrine system and mediated at least in part by the nervous system.

Breeding season. The season when animals perform their overt reproductive functions is known as the breeding season. This sexual periodicity is a general phenomenon common to plants and animals. In general, sexual periodicity is predominant in lower vertebrates, but all birds and wild mammals are seasonal breeders. As nutritional and environmental conditions improve, the reproductive season is not so restricted; cattle, the domestic rabbit, and humans all illustrate this fact; nevertheless their fertility is higher in spring than in winter. Males of many mammalian species are capable of sperm production

and copulation at any time and rarely experience a true sexual period such as rut in the deer. In mice and rats, the testes do not descend to the scrotum until puberty and in certain other rodents they descend only during the breeding season. Most vertebrates breed only in spring or summer at the time when food, temperature, and light are optimal; they are then in the best physiological condition for reproduction. Others, for example, sheep, breed at a time which allows parturition to occur at the season of year most suitable for successful survival of the young. *See* REPRODUCTIVE BEHAVIOR.

All environmental factors play a part, but an increase of temperature for lower vertebrates and an increase or decrease of daylight for mammals play a major role in determining the onset of their breeding seasons. For instance, most fish and amphibians breed when temperature increases. The mink and horse breed at the time of increase of daylight, in the spring, and their offspring are born in the summer or the spring, respectively, of the following year. The armadillo, deer, goat, and sheep breed during the decrease of daylight in the autumn and their offspring are born the next spring. Transportation of sheep from the Northern to the Southern Hemisphere changes their breeding season to accord with the seasons of the new environment. The testes of the cottontail rabbit return to a completely immature condition in the fall and their germ cells are all of the spermatogonial type. In late November, the testes begin to grow; in early spring they reach 50–100 times their weight in the inactive stage and the growth of accessory glands follows that of the testes. Although artificial increase of daylight brings wild rabbits and ferrets into breeding condition in midwinter, permanent short-day lighting does not entirely prevent breeding in the spring. Furthermore, artificial increase of daylight has so far failed to show any definite effect in advancing the breeding season of the domestic rabbit, guinea pig, hedgehog, or ground squirrel.

Among fishes the duration of the breeding season varies considerably according to the group to which they belong and where they live. The ova of the elasmobranchs are deposited singly or in pairs at varying intervals throughout a great part of the year, but some species appear to have regular recurrent breeding seasons. Most bony fishes, producing millions of ova, breed only in the spring, and most of them migrate to a suitable locality for the deposition of their gametes. In the frog and many other amphibians the ova are produced during the winter hibernation when the animals eat very little. Similarly, the genital organs of salmon develop during migration, when the fishes cease to feed. The spermatogenesis of the amphibian *Rana temporaria* is determined by the gonadotrophic activity of the pituitary and by the sensitivity of the germinal epithelium to gonadotrophin during the spring; this activity and sensitivity are influenced by temperature rather than light. In the autumn and winter their testes are not sensitive to gonadotrophin under natural conditions. Reptiles that hibernate usually begin to breed shortly

TABLE 2. Average length of the ovulatory cycle of some familiar mammals

Animal	Length
Opossum	28 days
Guinea pig	16 days
Rat	5 days
Mouse	5 days
Sheep	16 or 17 days
Pig	20–22 days
Cow	21 days
Horse	21 days
Macaque monkey	24–26 days
Human	28 days
Chimpanzee	37 days

after the beginning of the warm weather that terminates the hibernating period. Spring and summer are the seasons when most birds pair, build their nests, and incubate their eggs. The migration of some birds is invariably associated with an increase in the size of ovaries and testes.

Estrous and menstrual cycles. The cyclic changes of reproductive activities in mammalian females are known as estrous or menstrual cycles (**Table 2**). Most mammalian females accept males only at estrus (heat). The follicles in the ovaries grow at a rather constant rate before estrus. Just before or at estrus the follicles grow rapidly, resulting in a higher output of estrogen, a female hormone. During or toward the end of estrus the follicle, or group of follicles in animals producing litters, ruptures and releases the ovum (ovulation). The follicle is transformed into the corpus luteum, a glandular tissue which produces another female hormone, progesterone. If fertilization occurs, the corpus luteum persists during all or most of the gestation period according to the species. Without mating or fertilization the corpus luteum persists for a relatively short time and then degenerates. The growth of the next crop of follicles begins again.

Estrus. Estrus in mammals can occur several times in one breeding season; the mare, ewe, and rat come to estrus every 21, 16, and 5 days respectively if breeding does not take place. This condition is called polyestrus. The bitch is monestrous; she has only one heat, or estrus, to the breeding season and if not served then, she does not come into heat again for a prolonged interval, 4–6 months according to different breeds. The duration of estrus varies according to the species, such as 4–9 days in the mare, 16–18 h in the cow, and 14 h in the rat. In the ferret, estrus is shown by predominant swelling of the vulva and lasts several weeks unless copulation takes place.

In monestrous and seasonally polyestrous species the period of sexual quiescence between seasons is called anestrus. In sheep and deer this occurs in the summer, while in the ferret and mink it is restricted to the winter. Proestrus denotes a short period just before the time of coming in heat. It is very marked in the bitch; the genital organs swell and become congested, and there is bleeding from the uterus. Following estrus, the ensuing period in which sexual

activity declines is called metestrus. After metestrus a relatively longer period of diestrus occurs and occupies the bulk of the estrous cycle before proestrus recurs. Most females ovulate spontaneously, and shed their ova close to the end of estrus: about 2–4 days before the end of estrus in the mare, 13 h after the end of estrus in the cow, 10 h after the beginning of estrus in the rat. Some species, such as the rabbit, ferret, and cat, ovulate following the stimulation of copulation. If fertilization takes place, pregnancy occurs. If mating is unsuccessful, pseudopregnancy, a physiological condition very similar to pregnancy but which lasts a relatively short period, may occur in some species, particularly the dog and the rabbit. *See* ESTRUS.

Menstruation. The reproductive cycle of the female in the primate and human is well marked by menstruation, the period of blood flow from the vagina. Menstruation does not correspond to estrus but occurs between the periods of ovulation at the time the corpus luteum declines precipitously. As a result of the collapse of the superficial capillaries in the endometrium, blood is extravasated in quantity and about two-thirds of the endometrium is desquamated, or shed. After menstruation there is a period in which the uterus is replenished under the influence of estrogens by the growth of the epithelium and its capillaries. This is known as the proliferative stage and is equivalent to proestrus and estrus in other mammals. It is accompanied by growth of the vaginal epithelium and some cornification, but these changes are gradual and are not as well defined as in many other mammals. The Graafian follicle ruptures near the midcycle, about 15 days before the beginning of the next menstruation. Ovulation is spontaneous and is followed by the formation of the corpus luteum, whose secretion of progesterone causes great glandular growth of the endometrium. This is known as the luteal or secretory phase and is equivalent to diestrus in other mammals. In New World monkeys, menstruation is confined to the appearance in the vaginal lavage of a few red blood cells. Tissue destruction and hemorrhage in the uterus of the elephant shrew *Elephantulus* of South Africa at the end of diestrus has been described, but the bleeding at proestrus in the bitch, at metestrus in the heifer 14 h after ovulation, and at metestrus occasionally in the guinea pig is due to intensive congestion of blood vessels in the uterus or in the upper part of the vagina and is not true menstruation.

Mating. Mating, also called copulation or coitus, is the synchronized bodily activity of the two sexes which enables them to deposit their gametes in close contact. It is essential for successful fertilization because sperm and ovum have a very limited life-span; as examples, the fertilizing capacity of trout and salmon sperm lasts about 30 s in fresh water, that of rabbit sperm about 30 h in the female tract. Trout and frog ova become nonfertilizable soon after contact with water, whereas rabbit ova remain fertilizable for 6–8 h after ovulation. Human spermatozoa have a fertilizing life-span of about 72 h, and ova a fertilizable life of about 24 h. Mating behavior is a very

common phenomenon in the animal kingdom. Even in the primitive vertebrate the lamprey, some sort of mating takes place. The female is seized by the male, who winds his tail around her and discharges his sperm over the ova as they extrude from her body. In the dogfish, which practices internal fertilization, contact between male and female is necessary for the claspers of the male to convey his sperm into the cloaca of the female. In the majority of the bony fish physical contact may be slight, yet the male follows the female closely or presses his body to hers and deposits his sperm over the ova either soon after they are laid or at the same time. Certain salamanders walk slowly ahead of the female and deposit spermatophores which the female straddles and secures with the lips of her cloaca. In other low vertebrates a rudimentary penis exists either as a single intracloacal organ (crocodile, turtle, and duck) or as a paired structure placed behind the cloaca (snake and lizard) and sperm are deposited into the female's cloaca at mating. In mammals the copulatory organs are well developed, varying a great deal in structure. Patterns of mating are different; the time required for mating varies from 3 h (ferret) to a few seconds (sheep and rabbit), and the number of intromissions varies from 1 (bull) to 50 (hamster).

Mating in many birds and cold-blooded vertebrates is the culminating event of a more or less complicated series of love antics. It is quite probable that these sexual displays have a direct bearing upon ovulation and that in many species ovulation would not occur in the absence of such sexual experiences. Sexual excitement and violent physical exertions of both partners may also be essential for a fertile mating.

Ovulation. Ovulation can be considered as the culmination of a series of events in the process of oogenesis, which begins in the embryo and continues into adult life. Oogenesis is the formation, growth, and maturation of the oocyte within the ovary. The time of first ovulation at puberty varies among species, occurring as early as 30 days after birth in mice and as late as 11+ years in humans. The stages in the development of the ovum are, however, similar in most mammals. By birth, most of the primordial germ cells have become oogonia anad developed into primary oocytes. Most of these oocytes remain in this stage until, under the stimulation of gonadotropic hormones (especially follicle-stimulating hormone), one or more (depending on the species) follicles containing the primary oocytes begin to enlarge and differentiate into Graafian follicles. During this process, the primary oocyte undergoes a first meiotic division, in which the chromosome number is halved from the diploid to haploid state with the extrusion of the unwanted chromosomal material in the first polar body via unequal cytoplasmic division, and commences the second meiotic division. This secondary oocyte is released when the follicle is stimulated to rupture under the influence of the gonadotropin, known as luteinizing hormone. Ovulation has been likened to an inflammatory process, and it is thought that certain mediators of inflammation, such as prostaglandin $F_{2\alpha}$ 12-HETE, may

be involved in the final rupture of the follicle wall, thus permitting release of the ovum. The ovum remains arrested at metaphase II of the second meiotic division until sperm activation. It is surrounded at ovulation by two cellular layers, the bulky and sticky cumulus oophorus and the corona radiata, the highly adherent radially arranged three-deep layer of cells with connective processes joined to the vitelline membrane. *See* ENDOCRINE MECHANISMS.

Sperm transport in mammals. The logistics of sperm transport to the site of fertilization in the oviduct present many interesting features in mammals, but it is important to distinguish between passive transport of sperm cells in the female genital tract and sperm migration, which clearly attributes significance to the intrinsic motility of the cell. The rapidity (a few minutes in most mammals) with which spermatozoa reach the upper portions of the oviduct after mating or artificial insemination has emphasized that sperm motility alone could not have made a major contribution to passage over the relatively long distances involved. It is generally accepted that the role of sperm motility becomes critical only in certain regions of the tract such as the cervix and uterotubal junction and during penetration of the egg membranes. The journey for the spermatozoa is largely dependent upon contractions in the smooth-muscle layers of the tract, which may be enhanced by the action of oxytocin released after coital stimulation. Smooth-muscle stimulants in the semen, such as prostaglandins of the E series, may also contribute to the process of sperm transport in species where semen is deposited in the anterior vagina (for example, humans and sheep). Nonetheless, in animals in which the duration of estrus is long and mating takes place early in the receptive period, migration of spermatozoa from storage regions such as the lower oviduct must be of physiological significance, especially since the excited state of the uterus and oviducts promoted by coitally released oxytocin would long since have ceased.

In comparison with the number of spermatozoa deposited in the female at mating, there is a massive reduction in sperm numbers toward the upper portions of the oviduct. This sperm gradient reduces the chances of polyspermy by restricting the numbers of spermatozoa that achieve the site of fertilization. During this marked diminution in sperm numbers the cells are progressively removed from the seminal plasma, and they become strongly influenced by the uterine and tubal fluids. This situation increases the metabolic activity of ejaculated spermatozoa, and may also be directly associated with development of the capacitated state, discussed below.

Explanations for the fact that spermatozoa "ascend" the oviducts while the eggs are "descending" arise in part from viewing the isthmus as a compartment functioning in a manner different from the ampulla, a view that is emphasized by the results of intraperitoneal insemination. However, viable spermatozoa are actively motile, and although myometrial contractions play a major role in sperm transport through the uterus, progressive motility does

contribute to migration into and within the oviducts. Waves of peristaltic contraction in the isthmus proceed principally toward the ampullary-isthmic junction, and this activity should therefore assist transport of sperm cells. These waves of contraction are propagated through the circular muscle, and may be sufficiently powerful to occlude the lumen and, for example, to transport droplets of oil or Indian ink upward to the peritoneal cavity, but the contractions tend to fade beyond the ampullary-isthmic junction. Whether the relatively high concentration of 19-hydroxyprostaglandin E_1 or other prostaglandins of the E series in primate semen benefit sperm transport by means of some alteration in the myosalpinx remains to be clarified.

Cilia in the rabbit and pig isthmus may beat actively toward the ampullary-isthmic junction at the time of ovulation, thereby creating local currents of fluid. However, such ciliary currents have not been detected in the isthmus of cat, cow, sheep, monkey, or human.

The rate and efficiency of transport of spermatozoa are influenced by the timing of coitus relative to the moment of ovulation. Transport of cells as far as the tubal isthmus usually follows semen deposition shortly before ovulation, but it is uncertain whether sufficient viable spermatozoa would have reached the ampulla before the actual time of ovulation to be immediately available for penetration of the freshly ovulated egg or eggs. Spermatozoa may be largely sequestered in the isthmus in the preovulatory interval, only to be released for further proovarian transport when ovulation has occurred and the products of ovulation have entered the tube. The inference here is that there may be some synchrony between release of the egg from the Graafian follicle and a redistribution of spermatozoa from the constricted lumen of the isthmus.

Even though a specific attractant substance for spermatozoa has not yet been demonstrated to be released from mammalian eggs or their investments, some form of chemotaxis may contribute to the final phase of sperm transport and orientation toward the egg surface. Alternatively, maturation of the spermatozoa within the oviduct may induce intrinsic motility, which is then responsible for further progression of the spermatozoa along the oviduct to the site of fertilization.

Capacitation of spermatozoa. Although in most mammalian species the oocyte is shed from the Graafian follicle in a condition suitable for fertilization, ejaculated spermatozoa must undergo some form of physiological change in the female reproductive tract before they can penetrate the egg membranes. The interval required for this change varies from about 1.5 to 6 h according to species, and the process is referred to as capacitation. The precise changes that constitute capacitation remain unknown, although there is strong evidence that they are—at least in part—membrane-associated phenomena, particularly in the region of the sperm head, that permit release of lytic acrosomal enzymes with which the spermatozoon gains access to the vitelline surface of the egg. However, increased respiratory and flagellar activity are also noted after capacitation, the changes in motility being presumed to facilitate sperm penetration of egg membranes.

Capacitation is now understood to precede and permit a coordinated acrosome reaction, thus being an essential preliminary to activation or release of inner acrosomal enzymes and penetration of the zona pellucida; it may involve some adjustment in the distance separating the plasma and outer acrosomal membranes to permit the vesiculation reaction to proceed.

An understanding of the posttesticular changes occurring in spermatozoa in the male reproductive tract has provided some insight into the possible nature of capacitation. During passage along the epididymal duct, spermatozoa undergo a series of modifications that confer increased stability on the head, including a compaction of the nuclear chromatin and a tighter application of the plasma and acrosomal membranes to the sperm nucleus. Secretions within the epididymal duct also influence the maturing spermatozoa, for the antigenic nature of the sperm surface changes during the 1 to 2 weeks spent in epididymal transit. There is good evidence that enzyme inhibitors are "added" to the spermatozoa while still in the male, although whether the principal phase of addition occurs in the epididymis, during the preejaculatory mixing with whole seminal plasma, or in the initial passage of spermatozoa within seminiferous and rete testis fluid remains to be clarified. Neutralization or removal of inhibitors of acrosomal enzymes in the female tract would seem a logical stage in the capacitation process, or an immediate sequel to it, and could explain how the enzymes of fundamental importance in fertilization are stabilized in the sperm head until spermatozoa are within reach of the eggs.

Although capacitation seemingly involves both membranous and metabolic changes in the sperm cell, evidence suggests that modification in the glycoprotein and other moieties of the plasma membrane is a critical step. For example, as uterine incubation of ejaculated rabbit spermatozoa proceeds, the ability of the cell surface to bind lectins disappears progressively from the tip of the sperm head, suggesting that carbohydrate-containing moieties are being removed progressively from the plasma membrane. This alteration in the cell surface would be expected to prepare it for fusion with the outer acrosomal membrane, possibly by reducing net negative charges or by increasing membrane fluidity. As a sequel to the acrosome reaction, at least in hamster spermatozoa, there is a modification in the postnuclear cap which facilitates sperm head incorporation by the egg plasmalemma.

Although conventional descriptions of the acrosome reaction emphasize that membrane fusion occurs as a means of releasing proteolytic enzymes critical for sperm penetration of the egg investments, there is evidence that hyaluronidase may escape from viable sperm cells before membrane vesiculation can be detected in the electron microscope.

A second point concerning acrosomal enzymes is that acrosin—the putative zona lysin—is apparently packaged in zymogen form and needs to be activated from proacrosin during the capacitation process.

There is evidence from at least three mammalian species (rabbit, hamster, pig) that different compartments of the female tract act synergistically to promote capacitation, and that sequential exposure to the uterus and then oviduct accelerates the process of capacitation. A specific action of the oviducal environment is therefore suggested and, in the golden hamster, the effects on stimulated respiration and achievement of the capacitated state are so distinct that they can be seen through the thin wall of the ampulla in terms of a massive increase in sperm motility; even though hamster sperm reach the oviducts in under 1 h, another 3 h are required for this dramatic change to be accomplished. Precise evidence for the capacitation of primate spermatozoa is not available; if it occurs it must be rapid, since human ejaculated spermatozoa after washing, but without any exposure to female reproductive tract fluids, can penetrate matured oocytes in culture within 4 h.

The evolutionary significance of capacitation can be viewed in two ways. Physiologically, it can be regarded as a process whereby the sperm cell becomes competent to undergo an orderly sequence of membrane changes, collectively termed the acrosome reaction, to permit release of the lytic enzymes. Biologically, capacitation probably evolved to meet requirements associated with internal fertilization in mammals, and might best be regarded as an effective means of gamete selection during sperm transport through the different regions of the female reproductive tract; only spermatozoa at the optimal stage of their maturation process are able to penetrate the egg.

Fertilization. Fertilization is a process that begins with penetration of the ovum surface layer (zona pellucida) and binding to the vitelline membrane by the fully matured and acrosome-reacted spermatozoon, followed by incorporation of the spermatozoon into the ovum cytoplasm and activation of the ovum. The haploid chromosome complements of the sperm nucleus and ovum are then transformed into male and female pronuclei, respectively. The pronuclei migrate to the center of the ovum, permitting the joining of the two sets of chromosomes, or syngamy, which is the final stage of the fertilization process. The ovum, now known as a zygote, begins to develop by cell division, or mitosis. Normally, after penetration of the vitelline membrane by one spermatozoon, the fertilizing sperm, the entry of additional spermatozoa is prevented by the cortical reaction, the breakdown of cortical granules that underline the membrane. *See* FERTILIZATION.

Even if the processes of gamete transport succeed in bringing capacitated spermatozoa and recently ovulated eggs into functional contact in the appropriate portion of the oviduct, a viable zygote may still not be formed. Anomalies of fertilization such as digyny (retention of the second polar body) or

polyspermy (penetration of the egg cytoplasm by more than one sperm) are pathological in the higher mammals, and nearly always associated with very early embryonic death. Digyny or polyspermy arises especially following postovulatory aging of the egg, a situation in which control of the numbers of spermatozoa passing through the isthmus to the site of fertilization is also less effective.

Regulation of early embryonic development and transport. As noted above, fertilization takes place in the oviducts of mammals, usually in the region of the ampullary-isthmic junction, and the fertilized eggs or embryos do not descend to the uterus for some 3 to 4 days in most species. During this interval, the embryo undergoes a series of mitotic divisions until it comprises a sphere of 8 or 16 cells and is termed a morula. Formation of a blastocyst occurs when the cells of the morula rearrange themselves around a central, fluid-filled cavity, the blastocoele. As the blastocyst develops within the uterine environment, it sheds its protective coat, the zona pellucida, and undergoes further differentiation before developing an intimate association with the endometrium, which represents the commencement of implantation or nidation. Early development of embryos in most species is largely under the control of the maternal genome, with sequential activation and utilization of oocyte components. Expression of the embryonic genome does not occur until after the two-cell stage in mice, four-cell stage in cattle, four- to eight-cell stage in humans, and eight-cell stage in sheep zygotes.

During this preimplantation period, whose duration is approximately 4 days in small laboratory species, 7 days in humans, and 2 to 3 weeks in pigs and sheep, respectively, metabolic support of the embryo depends partially on its own cytoplasmic reserves of lipids and other substrates, but to a far greater extent on constituents of oviducal and uterine secretions. The nature and volume of the secretions are known to be closely regulated by the prevailing balance of ovarian hormones, estrogen and progesterone, and since the corpus luteum is developing during this period, it follows that the secretion of progesterone will also be increasing. Thus, the dynamic situation in the pattern of ovarian steroid secretion will be reflected in a changing pattern of oviducal and uterine fluids, which is critically adjusted to meet the nutritional requirements of the developing embryo. Many experiments show that the nature of the substrate available in oviducal fluid is closely regulated by the prevailing balance of ovarian hormones and is therefore changing to match the embryo's requirement for and ability to utilize the substrates. Mouse one-cell zygotes require pyruvate or oxaloacetate as an energy source, but at the two-cell stage can also utilize phosphoenolpyruvate and lactate; only after the eight-cell stage can glucose be utilized. Human zygotes appear much less dependent on a critical composition of the surrounding medium for growth and development, since many media, both simple and complex, have been used for culture of oocytes prior to the successful replacement of the

resulting zygotes into the oviduct or uterus of the infertile woman.

In nonprimate species, when developing zygotes are delayed in the oviduct beyond the time of normal tubal transport, they fail to implant; similarly, when they reach the uterus too early, subsequent development can be compromised. In primates, however, a zygote reaching the blastocyst stage in the oviduct can implant, resulting in an ectopic pregnancy. This condition is rare in nonhuman primates but not uncommon in humans, accounting for about 1% of all pregnancies. The reason for the occurrence of this condition, which is potentially life-threatening because of hemorrhage and shock subsequent to tubal rupture, is unknown; in nonprimates, oviducal secretions may be hostile to continued zygote development after the normal period of tubal transport, approximately 72 h. *See* PREGNANCY DISORDERS.

Coincident with their influence upon secretory activity, the ovarian hormones also regulate muscular activity in the oviducts, and so control passage of the embryo into the uterus. Any interference with the ratio of estrogens to progesterone during the free-living, preimplantation stages of the embryo will adversely influence fertility through dual effects on transport and nutrition of the embryo. Administration of suitable preparations of ovarian steroid hormones, such as those contained in the contraceptive pill, thereby provide potent means of regulating fertility even when taken after coitus. In fact, in some instances, the principal mechanism whereby ovarian steroidal hormones regulate fertility is not by blocking ovulation via steroidal feedback mechanisms, but by upsetting the delicate relationship between the egg and the reproductive tract, leading to failure of fertilization or degeneration of the egg.

The factors that control muscular activity of the oviduct, and hence modulate the transport of the developing zygotes to the uterus at the correct time, have not been defined. However, it is known that oviducal motility can be influenced by prostaglandins, catecholamines, peptides (such as vasoactive intestinal peptide, neuropeptide Y, substance P, oxytocin, and vasopressin), cyclic nucleotides, and gamma aminobutyric acid. Flow of oviducal fluid and ciliary activity of epithelial cells lining the oviducts are also thought to be involved in the transport process.

Implantation. Association of the embryo with the uterine epithelium, either by superficial attachment or specific embedding in or beneath the endometrium, leads in due course to the formation of a placenta and complete dependence of the differentiating embryo upon metabolic support from the mother. Implantation and placentation exhibit a variety of forms, but in all instances the hormonal status of the mother is of great importance in determining whether or not implantation can proceed. In several species of rodent such as rats and mice, implantation is held in abeyance during lactation, while in other mammals such as roe deer, badger, and certain bears, there is an obligatory period of delayed implantation known as embryonic diapause; the in-

appropriate endocrine conditions associated with delayed implantation render the blastocyst metabolically dormant. Ovariectomy shortly before the expected time of implantation can prevent implantation, whereas ovariectomy plus replacement therapy with a suitable ratio of progesterone and estrogen permits implantation.

Despite this dominance by maternal hormonal factors, studies have indicated that the embryo itself may be modifying the local environment between its trophoblastic surface and the endometrium by means of steroid hormone, prostaglandin, and platelet-activating factor secretion and perhaps also by production of carbon dioxide (CO_2). The cells of the trophoblast possess enzymes capable of synthesizing ovarian hormones, and embryonic membranes supplied in culture with suitable biochemical substrates will secrete estrogens and progesterone. Such a pattern of synthetic activity in the late preimplantation embryo, which has also been indicated by histochemical studies, may explain in part the regional changes in the endometrium corresponding specifically to the future implantation sites, and also the local changes in capillary permeability of the uterus as demonstrated by the Psychoyos pontamine-blue reaction.

The physical relationship between the implanting embryo and the uterus is best summarized according to whether the embryo remains in the lumen of the uterus, referred to as central implantation as in cattle, sheep, and pigs, or colonizes the glandular crypts and maternal tissues leading to eccentric or interstitial implantation, as in rodents and humans. The process of invasion by the embryonic trophoblast in the mouse, for example, promotes characteristic changes in the uterine stroma termed the decidual reaction. This cellular proliferation occurs on the fourth day of pregnancy in mice in response to the presence of a blastocyst, but the precise physiological nature of the stimulus leading to these changes is not understood.

One other feature of the implantation process that influences the form of the placenta is whether maternal cell layers are bypassed during development of the implantation association. In species where the embryo is invasive and penetrates the maternal tissues to reach the wall or lumen of the blood vessels, this gives rise to an endotheliochorial (carnivores) or hemochorial (rodents, primates) form of placenta. In other words, cells of the maternal epithelium no longer interpose between the maternal vascular supply and the developing embryo, a situation which facilitates transfer of nutrients to the fetus.

Prevention of implantation. It should be apparent that the process of implantation is susceptible to interference by hormonal or physical means, and provides a major avenue for the regulation of human fertility. Apart from contraceptive treatment with steroid hormones that disturb the oviducal transport of eggs or embryos (as described above) and hence the time relationships essential to implantation, the application of intrauterine devices (IUDs) in a variety of forms has also been used to prevent implantation

and development of the embryo. Such devices consist of threads; plastic coils; and T- or 7-shaped plastic devices that release copper or T-shaped devices that release progesterone or a more potent progestational steroid, levonorgestrel; they are inserted into the uterus via the cervical canal.

Their mechanism of action in inhibiting implantation remains to be clarified, but whatever form it takes, it does not prevent the rapid restoration of fertility upon loss or removal of the device from the uterus. A component of their action is almost certainly mechanical, disturbing the spatial relationship between the blastocyst and endometrium, but recent studies on laboratory and farm species indicate that the abnormal uterine fluids produced in the presence of IUDs must also play a role directly or indirectly in preventing development of the embryo and its subsequent implantation. The observation that the addition of copper or progesterone to inert IUDs markedly increases their contraceptive efficiency suggests the occurrence of toxic or inflammatory effects in the uterus acting on the blastocyst as well as on the endometrium itself. While systematic research continues to clarify the basic mechanisms involved, particularly the manner whereby uterine prostaglandins may mediate the observed effects, the use of IUDs by many women in Western societies and in developing countries continues to give substantial protection against pregnancy, without resort to daily or weekly regimes of pill taking. In the United States the availability of IUDs has been greatly restricted because of potential legal liability.

The most widely used method of fertility control is sterilization—vasectomy in the male and tubal ligation, resection, or cauterization in the female. In general, sterilization is nonreversible. Various plastic devices (capsules, IUDs, vaginal rings) or long-acting injectable biodegradable formulations that release potent progestional agents such as levonorgestrel, alone or in combinations with an estrogen, are possible alternatives to conventional oral contraceptives. Such devices and formulations are especially suitable for use in developing countries. The most advanced method available in certain countries comprises six silastic capsules that release levonorgestrel and are inserted under the skin of the forearm of the woman. They can provide effective contraception for up to 5 years. Methods to inhibit male fertility, other than condoms and sterilization, are not well advanced.

Much research has been devoted to the development of contraceptive vaccines. The most advanced method involves using the beta subunit of human chorionic gonadotropin (hCG) or a portion joined to a foreign protein to induce antibodies to hCG, which is required to maintain luteal function (secretion of progesterone) during early pregnancy in the human. Other antigens under consideration as the basis for a contraceptive vaccine involve proteins or glycoproteins isolated from spermatozoa or the zona pellucida (the outer coating) of the ovum. Antibodies to such antigens would prevent fertilization. Difficulties involved in development of contraceptive vaccines include risks of irreversibility, induction of autoimmune diseases, risk of teratological damage at low antibody titers, and the need to make each vaccination effective for at least 1 year.

Fertility regulation after implantation. Research is also attempting to develop a means of regulating fertility subsequent to the process of implantation. The luteotrophic support required for the early maintenance of pregnancy may be vulnerable to immunological blockade, although in treatments requiring antibody administration, the shortcomings mentioned above must be taken into account. The alternative approach to this form of hormonal regulation is the use of compounds that are directly luteolytic in nature or block the action of progesterone at the uterine receptors and that rapidly terminate the endocrine support of pregnancy. The prostaglandin $F_{2\alpha}$ is effective as a luteolytic agent in animals but not in primates. However, administration of prostaglandins alone during the first 49 days of pregnancy can terminate the pregnancy by directly stimulating uterine contractility with expulsion of the fetus. Side effects with this treatment are common and, although in general not life-threatening, make the method less acceptable. More promising is a potent progesterone antagonist, mifepristone, which combined with a low dose of a prostaglandin is very effective in terminating early pregnancy (up to 49 days) with minimal side effects. However, this method is available in only a few countries.

Endocrine function in reproduction. The endocrine glands secrete certain substances (hormones) which are necessary for growth, metabolism, reproduction, response to stress, and various other physiological processes. The endocrine glands most concerned with the process of reproduction are the pituitary and the gonads. Other contributions to the proper functioning of the reproductive system are provided by the thyroid, pineal, and adrenal glands. The hypothalamus is also an important organ for regulating reproduction. *See* ENDOCRINE SYSTEM (VERTEBRATE); ESTROGEN; PITUITARY GLAND.

The formation of gametes (spermatogenesis and oogenesis) is controlled by anterior pituitary hormones. The differentiation of male and female reproductive tracts is influenced, and mating behavior and estrous cycles are controlled, by male or female hormones. The occurrence of the breeding season is mainly dependent upon the activity of the anterior lobe of the pituitary, which is influenced through the nervous system by external factors, such as light and temperature. The ratio of light to dark during the daily cycle affects the secretion of melatonin by the pineal gland in the brain. Melatonin can be either stimulatory or inhibitory for reproductive activity in a wide variety of mammals, especially those with defined periods of estrus and anestrus (for example sheep and many rodents). Its role in human reproduction is still uncertain, but it can inhibit ovulation in certain circumstances. The transportation of ova from the ovary to the Fallopian tube and their subsequent transportation, development, and implantation in the uterus are controlled by a balanced ratio

between estrogen and progesterone. Furthermore, it is known that estrogens, androgens, and progesterone can all have the effect of inhibiting the production or the secretion, or both, of gonadotrophic hormones, permitting the cyclic changes of reproductive activity among different animals.

Mammary glands are essential for the nursing of young. Their growth, differentiation, and secretion of milk, and in fact the whole process of lactation, are controlled by pituitary hormones as well as by estrogen and progesterone. Other glands and physiological activities also influence lactation, although this is largely via the trophic support of other pituitary hormones.

M. C. Chang; Michael J. K. Harper; R. H. F. Hunter

Nonmammalian species. The successful radiation of recent vertebrates has depended upon the adaptation of their reproductive cycles to the environments in which they have evolved. Since almost all environments show some degree of fluctuation (water, food supplies, temperature, light), control of the reproductive cycle of males and females of a species is essential if young are to be born at a time of year that is advantageous to survival. Such control is brought about through the brain-pituitary-gonadal axis. In this system, information about the external environment is detected by appropriate sensory structures (for example, the lateral eyes) and released to the basal hypothalamus of the brain, which regulates pituitary production of gonadotropic hormones and thus synchronizes gonadal development with environmental cues. Evidence is available for all vertebrate groups (except the degenerate hagfishes, or lampreys) to suggest that control of the gonad by the brain-pituitary system exists. This involves a brain peptide, gonadotropin-releasing hormone, that stimulates the release of two pituitary gonadotropins. One of these, the homolog of mammalian follicle-stimulating hormone, is concerned with gamete production; the other, a homolog of mammalian luteinizing hormone, is primarily concerned with gonadal hormone production and ovulation in conjunction with follicle-stimulating hormone. With few exceptions, the known testicular and ovarian steroids that are synthesized and secreted by nonmammalian gonads are identical to the major gonadal steroids of mammals. These are progesterone and testosterone and their oxygenated and hydroxylated derivatives and the estrogens, of which the most important is estradiol-17β. The testicular and ovarian sites of synthesis and secretion of these hormones appear to be similar, if not identical, at least down to the elasmobranch level of vertebrate gonadal organization. Plasma levels of the gonadal steroids of both sexes reflect the stage of germ cell development and correlated steroidogenic cell function. The plasma steroids thus undergo marked fluctuations associated with the gonadal cycle and ensure proper central nervous function and development of the peripheral structures of the reproductive tract so that behavioral responses and internal (elasmobranchs, reptiles, and birds) or external (teleosts and amphibia) fertilization occur.

Spermatogenesis. Throughout the Vertebrata the cytological sequence of changes from immature, round germ cells through mature, flagellated spermatozoa is virtually identical. From early stages, germ cells form part of an anatomically and functionally related clone of cells, all derived from the same stem cell and synchronized through development. In addition to the germ cells, the primary spermatogenic units (spermatocysts) include somatic elements (the Sertoli cells) that provide structural and nutritive support and in all ways mediate between the soma and the developing germ cell. The spermatogenic process begins with a number of premeiotic spermatogonial divisions. These mitotic stages are followed by meiosis (spermatocyte stage), during which the $2n$ (diploid) number of chromosomes is reduced to the round spermatid stage. The spermatids enter a metamorphic stage during which they are transformed into elongate, flagellated spermatozoa.

The primary unit of testicular structure is the spermatoblast or spermatocyst, which consists of germ cell clones plus associated Sertoli cells. In anamniote species, germ cell clones are anatomically discrete spherical units. In contrast, in amniotes (reptiles, birds, mammals) the spermatocyst is not closed or spherical but irregular in form and open to other clones and the tubular lumen. The individual Sertoli cells of amniotes, unlike those of anamniotes, are shared by four or five germ cell clones in succeeding generations. These nonrandom cellular associations signify functional relationships that are the basis for the cycle of the seminiferous epithelium and the spermatogenic wave of amniotes. Apart from the differences in clonal association of the Sertoli cells of the amniotes compared to anamniotes, the Sertoli cells of some anamniotes (such as fish or urodele amphibians) proliferate, and they remain associated and develop synchronously with a given germ cell clone throughout spermatogenesis; at spermiation they degenerate or are otherwise lost into the semen. By contrast, Sertoli cells of reptiles, birds, and mammals are permanent elements of the seminiferous epithelium, nurturing generation after generation of developing germ cells. In these groups, Sertoli cell mitoses are thought to be restricted to prepubertal development.

The secondary germinal compartment is delineated by a boundary wall and composed of blind-ended sacs (lobules) or open-ended lobules, both of which are continuous with the intratesticular collecting duct system and surrounded by the endocrine Leydig cells. In the elasmobranchs, however, definitive Leydig cells appear to be absent; the spermatocysts are embedded in connective tissue stroma and joined to collecting ducts via a short stalk patent only at spermiation. In some teleosts and in urodele amphibians, a secondary germinal compartment can be identified in mid-late developmental stages, with stem cells and spermatogonial stage spermatocysts embedded directly in connective tissue stroma. The secondary germinal units are continuously formed adjacent to the germinal region and then regress after spermiation. Leydig cells are limited to

lobular-tubular regions. In some teleosts and anuran amphibians and in all amniotes, Sertoli cells and germ cells are permanent components of the adult testis, containing germinal elements at all developmental stages. Interstitial Leydig cells are present.

In all species, there is a strict temporal and spatial relationship among the different germ cell stages. Species variation is evident, however, in the timing and anatomical organization of cells in the spermatogenic sequence. The unique developmental sequence of spermatogenesis is regulated at several levels: at the organismal level, by circulating gonadotropic hormones; at the organ level, by local (paracrine-autocrine) factors; and at the cellular level, by direct cell–cell interactions. Comparative studies suggest that the Leydig cells evolved primarily for the secretion of steroids into the peripheral circulation. Although of secondary importance to the Leydig cell in the supply of steroid hormones, the nonmammalian Sertoli cell has a steroidogenic capacity that is quantitatively more important than that of mammals in providing the steroidal needs of germ cell production; in elasmobranchs, Sertoli cells are the primary steroidogenic component of the testis throughout the year. The elasmobranch testis is of particular interest, inasmuch as germ cell–Sertoli cell units can be harvested at specific stages of development, and it is possible to correlate steroidogenic function with specific stages. Androgen production is correlated with the more mature stages of sperm development, and androgen synthesis and specific estrogen receptors are significantly correlated with earlier spermatogenic stages. Studies with the mud puppy, *Necturus*, support a role for estrogen in the proliferation of early germinal elements because both androgen and estrogen receptors are found in testicular zones with immature germ cells. As in mammals, in nonmammals gonadal steroid hormones are transported within the testis and blood by androgen-binding proteins, as demonstrated in trout and shark. *See* TESTIS.

Oogenesis. As in mammals, oogonial proliferation in nonmammalian vertebrates is characterized by the presence of intercellular bridges, probably important in synchronizing of mitosis. Oogonial proliferation may be limited to the embryonic (larval) period (in lampreys, elasmobranchs, some bony fishes, and birds) or to the adult (in most teleosts, amphibians, and reptiles). Little is known of the role of hormones in the regulation of oogonial proliferation, but some evidence suggests that gonadotropins and possibly gonadal steroids increase oogonial division in some adult fish, amphibians, and reptiles. Primary oocyte growth appears to be gonadotropin-independent in cyclostomes; in contrast, in amphibians and reptiles the process is gonadotropin-dependent, and atresia of primary oocytes in hypophysectomized amphibians and reptiles is prevented by gonadotropins. Folliculogenesis, or the investment of the primary oocyte with follicle cells to form a primordial follicle, is similar in all nonmammalian vertebrates. As in mammals, the trigger for folliculogenesis is not known, but considerable evidence

suggests that at least the vitellogenic stages of follicular growth in all nonmammalian species are dependent upon pituitary gonadotropins. In amphibia and teleost fish, the resumption of meiosis at the onset of oocyte maturation and subsequent ovulation is gonadotropin- and progestin-dependent, as in mammals. Although all nonmammals produce yolky eggs, the lampreys, teleosts, and amphibians generally produce large numbers of small yolky oocytes, and the elasmobranchs, reptiles, and birds produce fewer, larger yolky eggs. A hierarchical arrangement of follicles or cohorts of follicles is readily seen in some reptiles and birds. Smaller ovarian follicles represent future generations of oocytes to be ovulated sequentially.

The process of hepatic yolk protein synthesis, vitellogenesis, is regulated by the ovarian hormone $17(\beta)$-estradiol and thus indirectly by the pituitary gland. The action of estradiol on the induction of yolk protein synthesis by the liver has been well studied. The process is steroid receptor–ediated and involves the hormonal induction of specific vitellogenin genes in liver cells. Vitellogenin, a complex phospholipoglycoprotein, is secreted into the blood by hepatic cells and is sequestered from the blood by developing oocytes under the influence of gonadotropin. In reptiles, it appears that hormones other than estradiol are involved in vitellogenesis. Growth hormone is synergistic with estrogen and required for the full expression of the estradiol effect. In contrast, progesterone, testosterone, and prolactin appear to be inhibitory to vitellogenin synthesis. Of these hormones, the progesterone is of particular interest in the vertebrate groups in which viviparity (live-bearing) is a common reproductive strategy (as opposed to egg laying, such as in elasmobranchs or squamate reptiles).

The evolution of live-bearing from the primitive egg-laying mode of reproduction has occurred in parallel in all vertebrate groups except birds. Physiologically, this process requires the retention of eggs, loss of the outer eggshell membranes, and the ultimate development of a placental mechanism. The placenta, when present, is involved in fetal nutrition and ultimately supplants the process of vitellogenesis. In elasmobranchs and reptiles, species have been described in which the dependency of the nutrition of intrauterine young upon the placenta has reached a near-mammalian condition. A corollary of this development is the reduction or elimination of yolk from the eggs. Some evidence suggests that the ovarian hormone progesterone is responsible for the switch from oviparity to viviparity through an inhibitory action on the smooth muscle of the reproductive tract. This tends to favor egg retention and thus viviparity. In mammals, progesterone is of primary importance for the maintenance of a quiescent uterus during gestation. Since even elasmobranch corpora lutea (formed from follicles after ovulation) synthesize and secrete progesterone, which inhibits uterine muscle activity, it appears that many of the endocrine mechanisms associated with reproduction and reproductive tract adaptations for egg, embryo, and

fetal care were evolved early in vertebrate evolution. Coincident with the effect of progesterone on the reproductive tract and egg retention, the function of the hormone to switch off cyclic vitellogenesis is extended during evolution to a progressive inhibition of vitellogenesis as placentation takes on the role of embryo-fetal nutrition.

In general, the two layers of the ovarian follicle wall (theca and granulosa) cooperate to carry out the enzymatic conversions from cholesterol through progesterone and androgens to estrogens. However, the two components have different roles in different species. Thus, androgens are converted to estrogens by the granulosa cells of fish and mammals, but this function is performed by the theca in birds. Conversely, androgens are produced by the mammalian theca, but by the avian granulosa. In elasmobranchs, the granulosa elements are the primary steroidogenic components, being able to synthesize estrogens from cholesterol. In a squamate reptile (snake), thecal synthesis of estradiol is low in the absence of substrate in the form of progesterone supplied by the granulosa. This pattern of enzymatic activities in the snake ovarian follicle is unique in vertebrates. In all vertebrate groups, ovarian follicular steroid synthesis and secretion is gonadotropin-dependent. In at least elasmobranchs and reptiles, corpora lutea formed from the ovulated follicle are steroidogenic and produce progesterone in response to gonadotropins in the living organism and in cell culture. In general, in all vertebrates gonadal steroids have been identified in plasma, and cyclic fluctuations of plasma levels of the principal ovarian steroids—progesterone, testosterone, and estradiol—correlate well with ovarian follicular development and ovulation. There is evidence that the circulating hormones have extensive actions on the peripheral target tissues that are important in the synchronization of the reproductive process in nonmammalian species.   Ian P. Callard

Neuroendocrine function. The associated physiological activities of the nervous system and of the endocrine system that influence animal reproduction are briefly dealt with here to illustrate the intrinsic mechanism involved. In the lower forms of life, especially in those forms without a nervous system, the rhythm of reproduction may be controlled metabolically by the direct action of environmental factors—food, temperature, light, humidity, and chemical composition of the environment. In the higher forms certain external factors act through the intermediation of the nervous system. In the bird, the number of eggs in a clutch is generally constant within narrow limits; if the eggs are withdrawn shortly after they are laid, many birds will go on laying, making an attempt to lay the right number. In the pigeon, ovulation often is induced by courtship with another pigeon. The number of eggs laid has been reported to be increased if a mirror is placed in front of a pigeon cage. These instances illustrate the influence of the nervous system on reproduction. *See* NEUROSECRETION.

Stimuli. Experimental study on birds and on ferrets has shown that breeding can be induced in midwinter by artificial light. Hypophysectomized and blind ferrets do not ordinarily react to light as expected, and it is obvious that the stimulus must be passed through the eye, optic nerve, or some receptors in the pineal gland or brain region, and via changes in secretion of melatonin or other indoles, and thence to the anterior pituitary. Moreover, the rabbit, ferret, and ground squirrel normally ovulate in response to the stimulation of copulation. Since this stimulation to switch from the follicular phase to the luteal phase cannot be effected in the absence of the pituitary but can be brought about by injecting pituitary extracts or pituitarylike extracts, pregnant woman's urine (PU) or pregnant mare serum (PMS), it would seem that this stimulus is normally due to nervous reflexes mediated through the hypothalamus and pituitary. The stimulus, however, may be carried by several nervous paths, because local anesthesia of vagina and vulva, complete thoracosympathectomy, absence of any nerve pathway to the ovaries, or cervical sympathectomy do not inhibit ovulation after coitus. Because stimulation of the brain, of the lumbosacral part of the spinal cord, of the cervical sympathetic ganglion, or of the hypothalamus will induce ovulation to a certain extent, it seems that more than one nervous path and more than one mechanism for the initiation of ovulation must be involved. Furthermore, the rat, unlike the rabbit, ovulates spontaneously, but a prolongation of the life of the corpora lutea with subsequent pseudopregnancy can be induced by sterile mating, mechanical stimulation of the cervix, or electrical stimulation of the brain. Pseudopregnancy in the rat also seems to be mediated by the pituitary through nervous pathways. There is additional evidence to show that the stimulus for luteinizing hormone release in a spontaneously ovulating animal is controlled by nervous mechanisms employing cholinergic and adrenergic components.

Lesions. It is well established that lesions in the basal tuber or median eminence induce ovarian atrophy in the cat, dog, and rabbit. The role of the nervous system in establishing cyclic pituitary activity has also been emphasized. Localization of an erection center and of an ejaculation center in the hypothalamus has been reported. Appropriately placed hypothalamic lesions in the female guinea pig sometimes result in anestrus or prolonged estrous periods with sexual behavior in keeping with the gross changes of the cycle. In the male guinea pig, similar lesions induce sexual impotence, without genital regression. Hypothalamic production of oxytocin and the derivation of vasopressin from neurosecretory process have been postulated. As for mating behavior, it is assumed that advancing evolutionary status is accompanied by a progressive dominance of the nervous system and a corresponding reduction of endocrine control. To interpret the effect of sex hormones on mating behavior it has been suggested that their activity may increase the excitability of the central excitatory mechanism.

Fertility and sterility. The ability or inability to produce offspring is termed fertility or sterility. Fertility

and sterility occur in different grades among various species and among individuals of the same species. Absolute sterility is rare, but infertility of all degrees is very common, especially among higher vertebrates. The rate of reproduction in any species depends upon the average number of young born in each litter, the frequency of recurrence of breeding season, the duration of the reproductive period, and the age at which the animal starts to breed. The age as a general rule is earlier in small species than in large ones. In general, the number of young in a litter of mammals is inversely proportional to the size of the animal. For instance, a cow rarely produces twins, whereas the rat occasionally bears as many as 16 young. A theory of fertility has been proposed that states that individuation and genesis vary inversely; that is, the power to sustain individual life and the power to produce new individuals are inversely proportional. Where there is abundant food supply and a favorable environment, and the necessary expenditure of energy is relatively slight, the cost of individuation is much reduced and the rate of genesis is correspondingly increased. *See* INFERTILITY.

Factors controlling fertility. J. Hammond proposed that three factors control fertility:

1. The number of ova shed. In accordance with the genetic constitution of the species and the nutritional status, the number of ova shed is controlled by gonadotrophic hormones through the pituitary gland, but influenced by external factors. Before puberty, at old age, and during pregnancy, pseudopregnancy, and early lactation, very few mammals ovulate. Although by administering gonadotrophic hormones the possibility of increasing the number of ova shed in mammals was demonstrated, the actual number of young produced was rather low, probably as a result of limitations to support excessive embryos.

2. The number of ova fertilized. This depends upon the number of spermatozoa produced by the male, and the morphological and physiological integrity of the gametes. It also depends upon the probability of meeting between gametes provided by the male and female in the lower vertebrates, the efficiency of sperm transport to the site of fertilization, and the time of mating in the higher vertebrates. Aging phenomena in gametes have a deleterious influence on fertility.

3. The number of embryos developing into self-sustaining individuals. The probability of normal development depends upon where ova are deposited and the protection that the parents give to the zygotes in the lower vertebrates. In the higher animals, it depends upon the transportation of zygotes to the prepared uterus at the right time, and the physiological activities of embryo and of mother for proper implantation, for the maintenance of pregnancy, and for proper parturition.

Control of reproduction. The topic of controlling fertility, for example, the regulation of implantation in humans described above, should not be left without portraying the considerable technology available to modify fertility in domestic animals.

While artificial breeding of cattle, and to a lesser extent pigs, sheep, and goats, by means of insemination of deep-frozen or diluted semen samples has been extensively applied in recent years, enabling the exploitation of males of superior genetic merit, other features of contemporary reproductive technology are less widely known; attention will be devoted to two of these. The first can be termed controlled breeding, as seen in the various techniques of synchronization of estrus. There are obvious advantages in being able to decide when an animal is to be bred, not only for convenience at the time of artificial insemination or mating, but also because the time of birth, weaning, and marketing of the offspring can be predicted quite accurately. The key to controlled breeding lies in reprogramming the ovary, and treatments attempt to modify, in the short term, the influence of the secretions of the corpus luteum and the Graafian follicle expressed, respectively, in the luteal and follicular phases of the cycle. Inducing synchronized estrus in a population of animals is achieved most effectively by causing regression of the corpus luteum so that animals enter the follicular phase and estrus approximately synchronously.

The demonstration in the late 1960s that prostaglandin $F_{2\alpha}$ ($PGF_{2\alpha}$) was almost certainly the luteolytic factor in domestic species, informing the ovary that the uterus did not contain a viable embryo and inducing regression of the corpus luteum, has caused this hormone to become the basis of extensive estrous synchronization programs, replacing the progestagenic compounds previously used to simulate and extend the luteal phase prior to a synchronized follicular phase. Potent synthetic analogs of $PGF_{2\alpha}$ are now available that cause luteolysis following a single intramuscular injection. However, the corpus luteum in cattle is only susceptible to injected $PGF_{2\alpha}$ between days 5 and 15 of the 21-day estrous cycle, and so to achieve synchronization in all animals in a herd, two injections are given 10 days apart. Nearly all animals will be in estrus within 48–72 h of the second injection, when they can be inseminated as a group. Prostaglandins are being used in this manner extensively.

Another area of reproductive technology in domestic animals receiving attention is that of storage and transplantation of embryos, particularly in the context of exploiting rare or exotic breeds of cattle after procedures of superovulation to increase egg numbers. Transplantation of embryos is, of itself, not a new technique, for it was successfully performed in rabbits before the turn of the century, and since about 1950 embryo transfer has been a leading experimental tool in reproductive physiology. Two of the biological facts permitting such manipulations of the embryo are the relatively long period during which the blastocyst is free-living in the uterine lumen before implantation, especially in pigs, sheep, and cattle (ungulates), and also the fact that the embryo is not rejected as foreign material from the host uterus.

Apart from the care needed during handling and storage of embryos in the culture medium, many

experiments have indicated that the single most important factor determining the success of a transfer operation is the degree of synchronization of estrous cycles between the donor and recipient animals. In cattle and sheep, there is some degree of tolerance in the timing, but once development of the embryo is more than 24 h out of phase with that of the recipient uterus, the pregnancy rate commences to decline sharply. Transfer of single, freshly recovered embryos by a surgical means involving an incision to expose the uterus can lead to 90% of the transferred embryos developing into calves. Transfer of two embryos, one to each horn of the uterus, has been used to induce twinning in cattle, although here the success rate is closer to 70%. Nonsurgical transfer of embryos, that is, their introduction into the uterus from the vagina by means of a suitable pipette, has been much less satisfactory, rarely giving pregnancy rates of greater than 20–40%. Apart from the greater risk of introducing infection by this approach, there is also the problem of ejection of embryos which usually follows manipulation of the cervix during the luteal phase. Nevertheless, now that bovine embryos can be preserved at $-321°$F ($-196°$C), as has been the case for spermatozoa for many years, the emphasis is on overcoming the problems of nonsurgical transplantation. With synchronization of estrus a practical proposition, and deep-frozen banks of embryos in the offing, the attraction of breeding programs involving this technically straightforward means of propagation is considerable, although it can be economically justified only for valuable animals with desirable genetic traits.

In several species of farm animals (sheep, cattle, and pigs), zygotes up to the 8–16-cell stage can be halved or quartered. The separated clumps of cells can then be placed in homologous zonae, from which all cytoplasmic material has been removed, and can develop into normal young after retransfer to a suitable recipient. Not only does this technology rapidly create additional copies of the traits of a particular mother or father or both, but it also provides for identical twins, triplets, and so forth for studies of the impact of environmental and nutritional factors on the same genetic background. Such a technique will be useful primarily as a research tool rather than for production purposes, where transgenic animals will be more useful.

Such advances in control of animal reproduction have given rise to the application of these techniques to humans for the alleviation of infertility. Men with low sperm counts or poor spermatozoal motility can be assisted by collection of the semen specimen and concentration of the spermatozoa. The chance of pregnancy can be improved if the female partner undergoes artificial insemination with such treated sperm samples at the time of midcycle ovulation.

Infertile women can be induced to mature larger numbers of follicles by treatment with hormones. Ova from such follicles can then be recovered and placed in the oviduct with a prepared sperm sample; matured and fertilized in a culture dish and replaced in the oviduct; or matured, fertilized, and cultured in a dish and then replaced into the uterus. Many thousands of children have been born worldwide following such procedures.

In addition, zygotes can still develop normally after removal of one or more blastomeres. Through the use of polymerase chain reaction amplification of gene sequences from such a blastomere, prenatal diagnosis of genetic defects can be detected prior to replacement of an externally fertilized embryo in the mother. Such techniques can revolutionize the field of prenatal diagnosis, and have been used to detect a number of hereditary disorders. *See* PRENATAL DIAGNOSIS.

In animals, research has focused on the production of transgenic animals. Most experiments have been conducted in mice, although the techniques developed are being tried in farm animals. Genetic material with desirable characteristics can be isolated with the techniques of molecular biology, and the deoxyribonucleic acid (DNA) then microinjected into the ovum, or inserted into embryonic stem cells in a culture dish, which cells can then be microinjected into the cavity of a blastocyst and become commingled with the cells of the host inner cell mass. The foreign DNA can integrate into the host chromosomal DNA and be carried into both germ and somatic cells. The transgenic animal passes the foreign DNA to its offspring in a mendelian manner. After introduction of genes that are normally expressed in a tissue-specific manner, the location of DNA sequences important in developmental programming of each gene can be determined. The gene for rat growth hormone injected into mouse ova gave rise to mice that grew much larger than their noninjected siblings.

Other areas in which contemporary reproductive technology is attempting to increase animal productivity include induction of ovulation at a precise time, increasing the numbers of eggs shed (superovulation), extending the breeding season of sheep and goats, and perhaps most important of all, regulating the time of parturition to avoid perinatal losses. In the future this technology will doubtless also include the application of pheromones to modify in diverse ways the reproductive activity of farm livestock, and the possibility of predetermining the sex of offspring by insemination of X- or Y-bearing spermatozoa in fractionated semen samples. *See* GENETIC ENGINEERING; REPRODUCTIVE TECHNOLOGY.

M. C. Chang; Michael J. K. Harper; R. H. F. Hunter

Bibliography. E. Y. Adashi and P. C. Lang (eds.), *The Ovary*, 1993; W. Andrew, *Textbook of Comparative Histology*, 1959; C. R. Austin and R. V. Short (eds.), *Reproduction in Mammals*, 2d ed., vols. 1–5, 1982–1986; G. F. Bastian, *Reproductive System*, 1993; V. Blum, *Vertebrate Reproduction*, 1985; R. H. F. Hunter, *Physiology and Technology of Reproduction in Female Domestic Animals*, 1980; G. Lamming (ed.), *Marshall's Physiology of Reproduction*, vol. 2, 4th ed., 1990; W. R. Lyons et al., The hormonal control of mammary growth and lactation, *Recent Prog. Hormone Res.*, 14:219–248, 1958;

P. Pang and M. Schreibman (eds.), *Vertebrate Endocrinology*, 1991; J. R. Pasqualini, F. A. Kinel, and C. Sumida, *Hormones in the Fetus*, vol. 2, 1991; E. C. Roosen-Runge, *The Process of Spermatogenesis in Animals*, 1977; J. Rossant and R. A. Pedersen (eds.), *Experimental Approaches to Mammalian Embryonic Development*, 1988; M. Shemesh and B. J. Weir (eds.), *Maternal Recognition of Pregnancy and Maintenance of the Corpus Luteum*, *J. Reprod. Fertil.*, suppl. 37, 1989; D. Shoupe (ed.), *Contraception*, 1993; C. K. Weichert and W. Presch, *Elements of Chordate Anatomy*, 4th ed., 1975; R. G. Wolff, *Functional Chordate Anatomy*, 1991; R. M. Wynn and W. P. Jollie (eds.), *Biology of the Uterus*, 2d ed., 1989; S. S. C. Yen and R. B. Jaffee (eds.), *Reproductive Endocrinology: Physiology, Pathophysiology, and Clinical Management*, 4th ed., 1999; K. Yoshinaga and T. Mori (eds.), *Development of Preimplantation Embryos and Their Environment*, vol. 294 of *Progress in Clinical Biological Research*, 1989.

Reproductive system disorders

Those disorders which involve the structures of the human female and male reproductive systems.

Female System

Disorders of the female reproductive system may involve the ovaries, Fallopian tubes, uterus, cervix, vagina, or vulva.

Ovaries. Failure of the ovaries to form normally results in short stature, sterility, and lack of development of female secondary sex characteristics, such as breast growth, fat deposition in buttocks and thighs and mons pubis, and female escutcheon. Destruction of the ovaries after puberty results in loss of fertility, cessation of menses, loss of secondary sex characteristics, and osteoporosis. *See* OSTEOPOROSIS.

Abnormal ovarian development, or dysgenesis, includes chromosomal mosaics: true hermaphrodites, who possess both ovarian and testicular tissue, and pseudohermaphrodites. Male pseudohermaphrodites have undescended male gonads (testicles), but their external genitalia appear female because those tissues are not sensitive to the male hormone testosterone. Female pseudohermaphrodites have female gonads, but their external genitalia appear masculine because they have been exposed to male hormones. This might occur, for example, in certain enzyme defects of the adrenal cortex in which male hormone levels that are higher than normal are present in the circulation.

The ovaries undergo cyclic stimulation of pituitary hormones to produce an egg, and in some women ovulation is accompanied by a day or two of sharp unilateral lower abdominal pain. Either the follicle which produces the egg or the corpus luteum which forms after the egg is liberated can persist as nonneoplastic cysts. These may be asymptomatic except for altering the menstrual cycle, or they may be quite painful. *See* MENSTRUATION.

Physiologic cysts may be quite large and either single or multiple. Single cysts usually resolve spontaneously in the menstruating woman, but they may persist and require surgical removal. In polycystic ovarian syndrome, multiple cysts develop in both ovaries. Women with this condition may also have infertility, obesity, hirsutism (excessive or unusually situated hair growth), absent or unusually infrequent menstruation (respectively, amenorrhea or, oligomenorrhea), or heavy menses. These changes are all due to anovulation (the absence of ovulation) and unopposed estrogen stimulation. This condition is usually treated by clomiphene citrate, which produces ovulation; by oral contraceptive tablets, which suppress hormone production; or by progesterone, which opposes the action of estrogen. Surgical procedures used to be performed on the ovaries; although this may produce ovulation, it was also found to cause adhesions around the ovaries, and is no longer common practice.

Neoplastic enlargement of the ovary can be cystic or solid, benign or malignant. Adenomatous cysts are filled with watery (serous) or thick (mucinous) material, and the nature of the lining cells determines the kind of fluid produced and whether the cyst is malignant. The cysts are frequently bilateral, and even benign ones can grow large enough to interfere with the functions of the gastrointestinal or urinary systems. One of the most common ovarian tumors occurring in young women is the benign adult cystic teratoma (dermoid). This has both solid and cystic elements, and contains tissue from many different organ systems. Solid tumors can arise from the outer or germ-cell layer or the inner stroma, and vary in malignant potential and hormone production. Those tumors which produce male or female hormones can affect the menstrual cycle and produce symptoms such as hirsutism. Malignant ovarian tumors are commonly asymptomatic in their early stages, and may be quite widely spread in the pelvic and abdominal cavity before they are discovered. Surgical removal is still the mainstay of therapy for this disease. However, chemotherapy regimes, both intravenous and intraperitoneal, as well as whole abdominal radiation have been used to extend survival time and improve quality of life. The ovaries can also be the site of metastatic cancer from such locations as the stomach, breast, and colon. *See* OVARIAN DISORDERS; OVARY.

Endometriosis. Endometriosis, which is a condition involving the presence of ectopic endometrial tissue, can affect the ovaries, and is thought to be maintained by cyclic ovarian hormone production. Endometriosis can be found as large "chocolate" cysts of the ovary, called endometriomas, or small "blue domed" cysts; both are filled with old, dark brown blood. Tiny implants of endometriosis commonly called powder burns also occur on the surface of the ovaries, on other pelvic peritoneum, and over the Fallopian tubes and uterus as well. Endometriomas of the ovary may be painful, or may rupture and cause diffuse pelvic pain, while the smaller endometrial implants may cause severe pain with

menstrual periods, generalized pelvic pain, pain with intercourse, and infertility. To treat this disorder it is necessary to first consider the individual's age, reproductive goals, symptoms, and the extent of her disease. Medical management involves using chemicals that suppress ovulation such as oral contraceptive tablets; a weak antiestrogen drug; or gonadotropin-releasing hormone agonist, a synthetic brain hormone that suppresses ovulation and estrogen production. Each of these medications acts in a different manner to suppress ovulation and has different side effects. There are no medical measures available that cure endometriosis. Surgical treatment can be done through the laparoscope with removal of the endometriosis using electrocautery or the laser. An open surgical procedure can be done to cut adhesions, remove endometriosis, and clear the Fallopian tubes. Surgical cure is produced by removal of uterus, tubes, and ovaries. This will lead to menopausal symptoms, but they can usually be managed by continuous low-dose estrogen replacement.

Fallopian tubes (oviducts). Inflammation is the most common disorder of these organs, and if it is repeated or severe, destruction of the tubal lining with closure of the outer ends of the tubes can occur. A pyosalpinx may then develop, in which the tube is distended with pus, and subsequently a hydrosalpinx is left when the purulent material is replaced by a thin watery fluid. This inflammation can be caused by several organisms, such as gonococcus, *Escherichia coli*, bacteroides, and *Chlamydia trachomatis* agent. Sterility commonly results because the tube is permanently closed to passage of egg and sperm.

The second most common problem involving the tube is pregnancy. The egg is fertilized in the outer portion of the tube and descends to implant in the uterus, but in some cases the passage is delayed and the conceptus attaches to the wall of the tube. The tube has a small lumen and thin wall, and the growing pregnancy quickly enlarges and grows through the tube, leading usually to rupture and hemorrhage into the peritoneal cavity. In rare circumstances, the conceptus is aborted out the end of the tube or between the folds of the broad ligament and continues to grow. Usually, however, maternal bleeding is significant enough to cause pain and fainting, and investigation leads to surgical treatment of the problem. *See* PREGNANCY DISORDERS.

Uterus. Developmental abnormalities can occur during formation of the uterus. The Fallopian tubes might not join at all, or might join partially from the cervical end upward. Septae or walls in the vagina and uterus can also occur. These abnormalities are more common in females who are exposed to diethylstilbesterol (DES) during the mother's pregnancy.

The muscle (myometrium) and lining (endometrium) of the uterus are susceptible to various problems, including tumors, infections, and hormonal derangements. Benign tumors of the myometrium (fibroids) are a common disorder, producing irregular enlargement of the uterus and sometimes causing pain, obstruction of the urinary tract, and heavy vaginal bleeding. They are estrogen-dependent and usually regress with menopause. They can be treated with GnRH compounds to shrink them temporarily; they can be surgically removed (myomectomy); or the uterus can be removed (hysterectomy). Muscle tumors are rarely malignant. The uterine lining can form polyps, or can undergo malignant transformation, producing heavy, irregular bleeding. These tumors can occur during the reproductive years or postmenopausally, when they are thought to be stimulated by continuous high doses of estrogen. Treatment is removal of the uterus, tubes, and ovaries.

The uterus can be one site of a significant infection which produces fever and pain and usually involves other organs, such as ovaries and tubes. A more localized form of infection can occur, following delivery, in the endometrium alone, with low-grade fever and irregular bleeding.

Hormonal abnormalities resulting in anovulation can lead to overgrowth and irregular shedding of the endometrium. Exogenous hormones such as those administered to regulate bleeding or those contained in oral contraceptive tablets also affect the uterine lining and can cause irregular bleeding in some women. Hormonal replacement therapy after menopause can also cause regular uterine bleeding like menstrual periods, or irregular bleeding, depending on the way the hormones are administered. However, treatment of abnormal bleeding depends on its cause.

Pregnancy usually proceeds uneventfully, but certain accidents of pregnancy, such as threatened, incomplete, or missed abortion, can produce irregular bleeding and some uterine discomfort. A rare tumor involving the uterus is caused by the abnormal development of trophoblastic tissue into a hydatidiform mole (benign) or choriocarcinoma (malignant). Chemotherapy is currently an effective mode of treatment. *See* PREGNANCY; UTERINE DISORDERS; UTERUS.

Cervix. Infection of the endocervical glands with gonococcus or chlamydia trachomatis agent can occur. This may be asymptomatic, except for producing a mucopurulent discharge, or it may cause pain when the cervix is manipulated, particularly during intercourse. The infection can ascend from this area into the internal genital organs and adjacent structures, so that these organisms should be treated if detected. *See* GONORRHEA.

The cervix can be affected by malignant tumors which can be adenocarcinomas, or tumors of the glandular cells, or more commonly squamous tumors. Cervical cancer is associated with several factors, including early age at first coitus, multiple sex partners, especially at an early age, smoking, and infection with certain subtypes of human papilloma virus. There is no single factor that "causes" this cancer, but rather a combination of these factors is critical. Papanicolaou smears, in which exfoliated cervical cells are specially stained and examined, are used

to detect early changes suggestive of this cancer. Further evaluation is by low-power magnification, called colposcopy, and biopsy, which allows the cancer to be detected in tissue samples. Treatment includes surgery or radiation, depending on the extent and location of the disease. Malignant tumors confined to the surface of the cervix can be removed with a cold knife or laser.

Vagina. Developmental abnormalities of the vagina may include imperforate hymen, septae, both vertical and horizontal, and complete failure of development. These conditions are usually diagnosed by a combination of visual inspection, chromosome evaluation, radiologic studies, and exploratory surgery. If the problem is causing symptoms, surgical correction can be undertaken.

Inflammation of the lining of the vagina can be due to several common organisms: *Monilia*, a fungus; *Trichomonas*, a protozoon; and *Mobiluncus*, a bacterium. Specific chemical treatment is available for each of these entities. Postmenopausal women can experience atrophy or shrinkage of the vagina and mucosa secondary to estrogen deprivation. This produces itching, bleeding, and pain with intercourse. Estrogen therapy can relieve these symptoms. The vaginal support can be weakened through childbirth, or may simply be naturally poor, allowing the bladder and rectum to bulge inward, and the cervix and uterus to protrude from the introitus. This condition can cause a variety of symptoms, including involuntary loss of urine with a cough or laugh, inability to move the bowels without mechanically pushing the stool out, and a sensation of pelvic heaviness. Intravaginal devices called pessaries may improve support, but surgical correction may be necessary.

Vaginal cancer can occur, but it is rare, and cervical cancer can also involve the vagina. Treatment is usually surgical.

Vaginal trauma can lead to significant bleeding. Penetrating straddle injuries can occur, and trauma with intercourse or childbirth can produce lacerations. Surgical repair is usually undertaken.

Vaginismus, or painful spasm of the muscular sidewalls of the vagina, can occur, making intercourse painful. This involuntary muscle spasm then recurs with each episode of intercourse. Treatment involves use of desensitization, muscle relaxation techniques, and, in some cases, psychotherapy.

Chronic spasm of these muscles or muscle pain due to tension can produce chronic pelvic pain. This condition may have no secondary cause or may be caused by musculoskeletal trauma such as falls. It is treated by the use of measures for muscle relaxation, including medication, massage, diathermy, and ultrasound; and by behavioral measures, such as relaxation training, guided imagery, and hypnosis. *See* VAGINAL DISORDERS.

Vulva. Infection of the female external genitalia can be diffuse or localized, involving bacteria, viruses, or fungal agents. Local infections such as a Bartholin's abscess are treated by surgical drainage. Infections of hair follicles may require topical or systemic antibiotics. Fungus infections should be treated by topical antifungal agents. Viral ulcers caused by herpes can be reduced in frequency by oral acyclovir, an acyclic nucleoside; the virus cannot be cured. Viral warts can be treated by chemical application, freezing, or laser vaporization; the wart is removed, but the virus may still live in adjacent tissue.

Inflammation can be caused by allergic reactions to soap, powders, semen, lubricant, or even clothing. Local measures to improve perineal hygiene and decrease irritation usually reduce symptoms. These measures include use of loose cotton underwear; soaking with water and mild agents such as colloidal oatmeal or artificial seawater; air drying; and use of nonocclusive clothing. The most serious disorder of this area is carcinoma arising in the epidermis. This tumor can be unifocal or multifocal, and is usually slow-growing and relatively asymptomatic. It can, however, spread to the lymph nodes, both in the groin and abdomen. Surgical removal is the mainstay of treatment.

Male System

The principal organs of the male reproductive system are the testicles, epididymis, vas deferens and seminal vesicles, prostate gland, urethra, and penis.

Testicles. During fetal life the testicles form in the abdominal cavity near the kidneys and migrate down into the scrotum. Failure of descent permanently damages the sperm-producing cells, but allows the interstitial cells which produce hormones to survive. Sometimes only one testicle remains undescended, in which case fertility is usually normal because gametes form in the other gonad. Undescended testicles develop malignant neoplasms more commonly than normal testicles.

Other causes of male infertility are organic problems which block passage of seminal fluid or interfere with sperm production. Some common examples are metabolic derangements such as hypothyroidism or malnutrition, exposure to radiation or toxic chemicals, infection (most common with mumps virus), and destruction of testicular tissue by twisting and obstruction of the blood supply (torsion).

Absence of testosterone production by the interstitial cells can occur because of castration or disease. If it happens before puberty, there is a failure of development of male secondary sex characteristics, and the individual is eunuchoid, lacking a beard, male escutcheon, deep voice, and libido. Since testosterone is responsible for maintaining these characteristics in the adult male, a decrease in production (or an increase in binding proteins) results in a gradual decrease in these characteristics, which can be seen in men as they pass into old age.

The testes are sites of malignant tumors which carry a high mortality rate. Undescended testicles are more susceptible to malignant transformation, and for that reason should be removed when discovered, and appropriate hormonal replacement instituted. Seminomas are the most common tumors, but the cure rate approaches 80% with surgery and

radiation. Chemotherapy is utilized for recurrent or advanced disease. Other tumors such as embryonal cell carcinomas and teratocarcinomas are of more primitive cell types, occur in young men and metastasize widely. However, about 75% of these individuals are responsive to chemotherapy.

The membranes covering the testicle can become filled with fluid, a condition known as hydrocele. *See* TESTIS.

Epididymis. The major disorder of these paired organs is infection and inflammation, which can lead to scarring and permanent blockage of the ducts. Tuberculous epididymitis is a chronic inflammation, in which the organisms apparently gain access from the bloodstream, while the gonococcus causes an acute inflammation acquired during sexual intercourse and ascending the urethra, vas deferens, and finally the epididymis. Nonspecific inflammation can also occur, usually due to introduction of organisms during instrumentation of the urethra or surgery on the prostate.

Vas deferens and seminal vesicles. These organs are rarely afflicted by disease, although the network of veins surrounding the vas deferens can become engorged and tortuous, and is called a variococoele. This condition may be a cause of lowered sperm production and male infertility. Treatment is operative removal of the veins. The vasa can be surgically interrupted by a simple procedure to produce sterility in the male.

Prostate. Inflammation and development of small calculi are relatively minor ailments of this organ. Bacteria reach the prostate gland by migrating upward from the urethra or down from the kidneys, ureters, and bladder. Acute infection of this gland can cause chills, fever, and bladder irritability and may lead to abscess formation. Gonorrhea is a common cause of this problem, and appropriate antibiotic therapy should be administered. Chronic infections may produce only mild symptoms but can be accompanied by calculi formation.

Benign overgrowth of this gland is the most common and troublesome complaint. This enlargement results in constriction of the urethra, which obstructs urinary outflow and leads to an increasing residual of urine in the bladder. The symptoms of benign prostatic hypertrophy are difficulty and frequency of micturition and increased urinary tract infections because of the residual urine. The causes of this condition are unknown, and treatment is surgical removal of portions of the gland.

Cancer of the prostate is another relatively frequent problem in older males. This tumor spreads primarily to bones, where it is markedly painful, and it produces a large quantity of the enzyme known as acid phosphatase. This growth is stimulated by male sex hormones and inhibited by female sex hormones, so that treatment for advanced disease is removal of the testicles and administration of female hormones, or use of drugs to suppress adrenal androgens. Localized disease can be treated with brachytherapy (the placement of a solid radioisotopic source on the surface of the growth) using radioactive iodine, external beam irradiation, radial surgical removal of the prostate gland, or in some cases cryosurgery. *See* PROSTATE GLAND DISORDERS.

Urethra. Inflammation is the chief disorder of this organ, causing dysuria and a discharge. The gonococcus can cause an infection of this site, and before the age of antibiotic treatment, gonococcal urethritis or its treatment often resulted in scars or structures which obstructed urinary outflow. Other agents can cause urethritis as well, and it now appears that the chlamydia trachomatis organism accounts for a large proportion of those infections once thought to be "nonspecific."

Penis. Inflammation (balanitis) or narrowing (phimosis) of the foreskin, which covers the glans penis, can occur and may interfere with urination. These conditions are unknown in parts of the world where the foreskin is removed for social or medical reasons. Viral infections can produce warty growths (condyloma acuminata) or painful ulcers (herpetic lesions) of the glans or shaft of the penis, and a syphilitic infection can result in the firm, painless ulcer known as a chancre. *See* HERPES; SYPHILIS.

Disorders of erectile capability range from tumescence unaccompanied by sexual desire (Peyronie's disease, priapism) to impotence, both psychologic and secondary to old age. A number of prosthetic devices are available to surgically treat erectile dysfunction of organic cause. Psychosexual counseling to treat psychogenic disorders has had variable success. Carcinoma can arise from the surface epithelium of the penis and spread both locally and via the lymphatic system to the nodes of the groin. This disease is virtually unknown in populations where males are circumcised at birth. *See* INFERTILITY; ONCOLOGY; REPRODUCTIVE SYSTEM; REPRODUCTIVE TECHNOLOGY.   Gay M. Guzinski

Bibliography. L. Copeland, *Textbook of Gynecology*, 2d ed., 1999; A. Herbst et al., *Comprehensive Gynecology*, 4th ed., 2001; N. Kase et al., *Principles and Practice of Clinical Gynecology*, 2d ed., 1990; D. B. Seifer and L. Speroff et al., *Clinical Gynecologic Endocrinology and Infertility*, 6th ed., 1999; P. Walsh et al., *Campbell's Urology*, 7th ed., 1997.

Reproductive technology

Any procedure undertaken to aid in conception, intrauterine development, and birth when natural processes do not function normally. The most common are in vitro fertilization and gamete intrafallopian transfer.

Infertility, which is the inability to conceive during at least 12 months of unprotected intercourse, is an increasingly common problem. Hormonal therapy and microsurgery are used to overcome many hormonal and mechanical forms of infertility, but many types of infertility do not respond to such treatment. The birth of the first infant conceived by in vitro fertilization—in England in 1978—marked the beginning of reproductive technology as a clinical

tool. Thousands of infants have been born as a result of in vitro fertilization, and numerous centers have been established to perform the procedure. The related technological procedure of gamete intrafallopian transfer was introduced in 1984. *See* INFERTILITY.

In vitro fertilization. In vitro fertilization bypasses the Fallopian tubes. Immediately prior to ovulation the mature oocyte is removed from the ovary and placed, together with prepared sperm, in a petri dish for 2–3 days. Fertilization takes place during this time and the fertilized oocyte develops into a two- to eight-cell embryo, which is then transferred into the uterus.

In vitro fertilization is useful for females with absent or severely damaged Fallopian tubes; couples in which the female has endometriosis; when the male has severely reduced sperm counts; or when the couple has immunologic or unexplained infertility for a period of 2 or more years.

The four principal steps of in vitro fertilization are induction and timing of ovulation, oocyte retrieval, fertilization, and embryo transfer.

Induction and timing of ovulation. Prior to ovulation, oocytes are in a resting stage of the first meiotic division (prophase). Under the influence of the midcycle surge of luteinizing hormone, an oocyte matures to the metaphase II stage of meiosis and is released into the oviduct approximately 38 h later. An immature oocyte that has not reached metaphase II is not receptive to fertilization. In order to obtain a mature oocyte, it must be retrieved immediately prior to ovulation.

Although the natural cycle develops only a single oocyte, results can be comparable to medicated cycles if the woman is 35 years of age or less and has very regular cycles. Women over 35 and women who have irregular cycles become less successful with increasing age and are poorer candidates for a natural cycle. Women of any age whose partners have low sperm counts are also poor candidates.

The majority of centers use agents to induce ovulation for in vitro fertilization. These medications allow the simultaneous development and maturation of multiple oocytes. In addition, ovulation can be forced to occur at a specified time so that oocyte retrieval can be precisely scheduled.

Many ovulatory inducing agents are available. Human menopausal gonadotropins can be used either alone or in sequence or combination with clomiphene citrate. A purified compound of follicle-stimulating hormone is also used. The availability of a gonadotropin-releasing hormone agonist has provided another approach.

The various regimens are monitored by obtaining daily serum estradiol levels and frequent pelvic ultrasound examination to determine the size, number, and potential adequacy of the developing follicles. Ideally, estradiol levels will steadily increase, and at least two, or preferably three or more, follicles will reach a diameter of approximately 0.72 in. (18 mm). Before the natural midcycle rise in luteinizing hormone occurs, an injection of human chorionic gonadotropin is administered intramuscularly
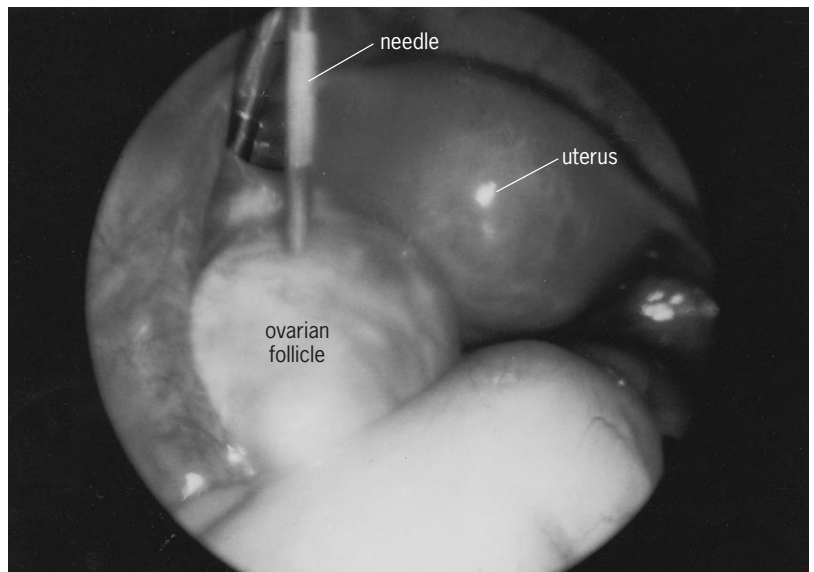
In laparoscopic oocyte retrieval the needle is inserted into the ovarian follicle. The uterus is interior. (*Courtesy of M. M. Seibel*)

to simulate the luteinizing hormone surge and initiate resumption of meiosis. Ovulation will occur approximately 36–38 h following an injection of human chorionic gonadotropin, just as it does after a natural rise in serum luteinizing hormone. Failure is typically due to a poor response of estradiol, an insufficient number of developing follicles, or a spontaneous luteinizing hormone surge.

Oocyte retrieval. Following the adequate development of preovulatory follicles, a mechanical ovulation or oocyte retrieval is performed. Oocytes are successfully retrieved more than 95% of the time. The traditional method of performing oocyte retrieval has been by laparoscopy (see **illus.**). The laparoscope is inserted in the routine fashion through a subumbilical incision that is 0.5 in. (13 mm) in length. A second small incision is made in the pubic hairline through which a grasping instrument is inserted to stabilize the ovary. A needle is then inserted either through the laparoscope itself or through a small trocar inserted 1–2 in. (25–50 mm) above the second incision. With the aid of the laparoscope, the needle is directed into the follicle, and the follicular fluid and oocyte are aspirated out by means of gentle suction.

Although laparoscopic oocyte retrieval is a highly successful technique, increasing use has been made of ultrasound-guided oocyte retrieval for several reasons. First, many individuals requiring in vitro fertilization have extensive pelvic adhesions that make the ovaries difficult or impossible to visualize by laparoscopy. Without ultrasound, a preparatory laparotomy would be required to make the ovaries accessible for laparoscopic retrieval. A second advantage is that the procedure is less invasive and can, therefore, be performed under local anesthesia augmented with intravenous analgesic agents. These advantages allow ultrasound-guided oocyte retrieval to be performed outside the operating room in an

ambulatory surgery or even an office setting. However, concealed hemorrhage can occur, so that ideally the procedure should be performed in a hospital setting or surgical facility.

The most commonly used ultrasound technique is transvaginal. A transvaginal transducer with an attached needle guide is placed into the vagina and the needle is guided through the posterior cul-de-sac directly into the follicle.

Fertilization. Following retrieval, the oocytes are allowed to complete maturation in culture because they are removed from the follicle several hours prior to ovulation. The male is asked to produce a semen specimen during this incubation period. The spermatozoa are washed in special insemination medium to separate them from the seminal plasma, which is the fluid surrounding the spermatozoa. This process allows the spermatozoa to undergo capacitation, that is, changes in the sperm plasma membrane that must be induced to allow fertilization to occur. Approximately 50,000–100,000 sperm are added to the culture medium containing the oocyte. It is not until the spermatozoa enter the oocyte that the oocyte resumes meiosis and completes its second meiotic division. The oocyte is visualized the following morning for evidence of fertilization. If two pronuclei are seen, fertilization has occurred. By 37 h (range 31–43 h), cleavage will have occurred, resulting in a two-celled stage. Embryos are usually transferred into the uterus 48–72 h following the fertilization. *See* FERTILIZATION.

Embryo transfer. Although the transfer of embryos into the uterus is technically simple, it is the stage at which most in vitro fertilization procedures fail. A speculum is placed in the vagina and the cervix cleansed with sterile saline. One or more embryos, combined with a small volume of culture medium (20–30 microliters), are loaded into a thin catheter, which is threaded through the cervical catheter to within a few centimeters of the uterine fundus, where the embryos are released. The catheter is held in place for 1 min or less and then slowly removed. The woman is asked to remain in the same position for a few minutes.

Gamete intrafallopian transfer. Gamete intrafallopian transfer, or GIFT, is similar to in vitro fertilization with a few important distinctions. The medications used to induce ovulation, and the daily blood testing and pelvic ultrasound are virtually identical to those of in vitro fertilization. The male provides a semen specimen, and it is prepared in a manner similar to that used in in vitro fertilization. However, in gamete intrafallopian tube transfer, the spermatozoa and oocytes are placed into the fimbriated end of the Fallopian tube during the laparoscopy. Indications for gamete intrafallopian transfer include unexplained infertility of two or more years' duration, cervical stenosis, immunologic infertility, oligospermia, and endometriosis. At least one Fallopian tube must appear normal; where there has been severe pelvic adhesions or distorted tubal anatomy from any cause, gamete intrafallopian transfer should not be considered. Individual consideration should be exercised when a patient has had a previous ectopic pregnancy. *See* PREGNANCY; REPRODUCTIVE SYSTEM DISORDERS.   Machelle M. Seibel

Bibliography. J. Bernstein, J. Lewis, and M. M. Seibel, Effect of previous infertility on maternal-fetal attachment, coping styles and self-concept during pregnancy, *J. Women's Health*, 3:125–133, 1994; P. Devroey, Normal fertilization of human oocytes after testicular sperm extraction and intracytoplasmic sperm injection, *Fertil. Steril.*, 62:639–641, 1994; M. M. Seibel et al., *Technology and Infertility: Clinical, Psychosocial, Legal and Ethical Aspects*, 1993; M. M. Seibel, M. Kearnan, and A. Kiessling, Parameters that predict success for natural cycle in vitro fertilization, *Fertil. Steril.*, 63:1251, 1995.

Reptilia

As traditionally defined, a class of vertebrates composed of four living orders—Testudines (turtles), Rhycocephalia (tuataras), Squamata (lizards and snakes), and Crocodylia (alligators and crocodiles)—as well as numerous extinct orders. Using recent scientific findings, however, many experts are now giving support to a very different arrangement of the reptile group that asserts that birds have a far closer evolutionary relationship to reptiles than once thought, and in fact should be considered reptiles. A classical taxonomy for the class is given below, followed by the classification scheme for extant reptiles currently accepted by most professional herpetologists. All orders other than the Testudines, Rhycocephalia, Squamata, and Crocodylia are extinct (see separate articles on each taxon listed).

Classical taxonomy:
  Class Reptilia
    Subclass Anapsida
      Order Captothinida
        Mesosauria
        Testudines (turtles)
    Subclass Diapsida
      Infraclass Sauropterygia
        Order Nothosauria
          Plesiosauria
          Placodontia
      Infraclass Lepidosauria
        Order Araeoscelida
          Eosuchia
          Rhyncocephalia (tuatara)
          Squamata (lizards and snakes)
      Infraclass Archosauria
        Order Protosauria
          Rhynchosauria
          Thecodontia
          Crocodylia (alligators and
            crocodiles)
          Pterosauria
          Saurischia
          Ornithischia
    Subclass Ichthyopterygia
    Subclass Synapsida

Order Pelycosauria
Therapsida

**Current classification scheme
for extant reptiles:**

Subclass Anapsida
Order Testudines (turtles)
Subclass Lepidosauria
Order Rhyncocephalia (tuataras)
Squamata
Suborder Sauria (lizards and skinks)
Amphisbaenia
(amphisbaenids)
Ophidia (snakes)
Subclass Archosauria
Order Crocodylia (alligators and
crocodiles)

Globally, there are approximately 8200 extant species of reptiles described (excluding birds), including 300 species of turtles, 2980 species of snakes, and 4760 species of lizards and skinks. Reptiles are found on all continents excluding Antarctica, with their greatest diversity centered in tropical and subtropical ecosystems.

Fundamental characteristics. Reptiles are considered to have certain fundamental characteristics. Most basic is an integument (skin) with well-developed scales. These scales cover most of the body and serve to reduce water loss in the terrestrial environment. The scales are made mostly of keratin, a protein-based substance that is relatively light and durable. The evolution of scales is not thought to have been driven by the need for protection, but was simply an adaptation to prevent rapid water loss through the skin. However, scales quickly began to find utility as excellent dermal protective coverings. Another fundamental characteristic of reptiles is that they are ectothermic, meaning that the predominant amount of their internal body temperature comes from external sources such as sunlight. Reptiles do generate metabolic heat, but most rely on thermoregulatory behaviors to maintain their optimal temperature, such as can be seen in the behavior of the basking turtle. Thermoregulatory behavior also means that reptiles must retreat to refugia when it becomes too hot. Reptile reproduction is as varied as their diversity and forms. Most species are oviparous, laying eggs in the sand, soil, or other materials such as decaying wood, but some species of lizards and snakes are viviparous, giving birth to live young.

Evolution. The evolution of reptiles originates with their diversification from early amphibians. Adaptations of reptiles are rooted in a more general early vertebrate move from a strong reliance on aquatic environments. One of the central events in this process was the evolution of the amniotic egg, hence the Amniota, which includes reptiles, birds, and mammals. The amniotic egg is far more complex than its anamniotic predecessor (represented today by the eggs of fish and modern amphibians). The amniotic egg has three layers: the chorion, allantois, and am-

nion. These layers are involved in protection, nutrient supply, and waste removal, all important functions when moving away from a dependence on the aquatic environment for the development of the larvae, a predominant condition in all fish and most amphibians today.

The earliest reptiles are best discussed within the context of the earliest amniotes, which date from the mid-Carboniferous period beginning approximately 330 million years before present (mybp). Based on the fossil record to date, there are two important time frames of reptile radiation. The first, from the late Carboniferous (290 mybp) through the Permian period (250 mybp), was an important phase of evolution and radiation of the ancestral groups of species we are familiar with today (turtles, snakes, and lizards), and the mammal-like reptiles (the reptile group that is the ancestral group to modern-day mammals). The second is often referred to as the Age of Reptiles. This period begins in the mid-Triassic ($\sim$225 mybp), with the increasing diversity of groups like the Ichthyosaurs, Pterosaurs, Plesiosaurs, Synapsids, Ornithischia, and Saurischia (the latter two orders are commonly referred to as dinosaurs). The success of these groups, along with other groups such as the squamates and testudines, continued until a very abrupt decline (particularly among the dinosaurs) at the end of the Cretaceous Period (65 mybp). The cause of this decline has been a topic of debate for half a century, and two more widely accepted hypotheses have emerged. First, it is proposed that increased volcanic activity globally led to excessive gases and particulate matter entering the Earth's upper atmosphere, thereby causing rapid and drastic climate shifts that affected all organisms via a cascade effect on the food web and caused drastic declines of many, particularly the large-bodied, dinosaurs. The second hypothesis asserts that a large meteor struck the Earth near the tip of the Yucatan Peninsula (often referred to as the Chicxulub Bolide Impact). This would have caused a significant plume of gas and particulate matter to enter the upper atmosphere, with the same result associated with the first hypothesis. In any case, while the outcome was catastrophic for dinosaurs, other groups began to be more successful, among them the evolutionary lines that ultimately led to modern mammals, birds, turtles, snakes, and lizards.

Living reptiles. Extant reptile groups can be distinguished easily by basic characteristics. Lizards, the most species-rich of the living reptile groups, have four limbs (with some interesting exceptions in the case of legless lizards), external ear openings, and eyelids that close. Lizards also have postanal tails that have the ability to break off when grabbed by a predator (tail autotomy).

Conversely, snakes have elongate bodies with no limbs, although members of the group have vestiges of limbs associated with their pelvic girdle. Unlike lizards, snakes have no external ear opening and no eyelids capable of closing. Snakes do, however, have a specialized transparent protective scale (called a spectacle) that covers the eye.

One highly specialized group, the amphisbaenids,

are similar in some respects to lizards and snakes, but have unique characteristics that cause some specialists to consider them a separate suborder of Squamata (Amphibaenia), apart from Sauria and Ophidia. Amphisbaenians have no limbs, with the exception of one family, and reticulate scales giving them a ringed appearance. They are highly adapted burrowers, and their head shape, often blunt or spatulate, belies their subterranean lifestyle. The highest species richness for the group occurs in South America and Africa.

Turtles are distinguished by their external shell, made up of a carapace (dorsal) and a plastron (ventral). This external covering is actually an outgrowth of the underlying vertebrae, complete with integrated ribs for support. The bone is covered by keratinized scutes for further protection. There are species of aquatic turtles, like the smooth softshell turtle (*Apalone mutica*), that have reduced bony shells covered with thick skin instead of keratinized scutes.

Crocodylians have distinctive elongate bodies, long tails and snouts, jaws with well-developed teeth, and bony plates covered with keratinized skin. Crocodylians are typically large in size and found in riverine, swamp, and marine habitats in a tropical and subtropical distribution.

Ecology. Living reptile species as a group are very successful in that they make up a significant proportion of the biodiversity of vertebrates in certain regions, and many have extraordinary adaptations that allow them to inhabit very harsh environments (such as the lizard assemblages of the Australian deserts). In fact, the most successful reptiles are snakes and lizards, making up 94% of the total number of extant reptile species (excluding birds). Reptile species have found ways to exploit many ecological niches. The gopher tortoise (*Gopherus polyphemus*) of the southeastern United States burrows meters underground, creating its own refuge and thereby creating refuge for many other (300+) species of plants and animals. Many species of snakes and amphisbaenids have fossorial or semifossorial lifestyles—that is, they either stay underground entirely or spend a significant portion of time there. Conversely, many snake species have made use of the arboreal niche and spend their time entirely in the canopy of trees.

Although one evolutionary trend in reptiles is a movement away from reliance on water, many have found great success in aquatic habitats. Some snakes, often referred to as sea snakes, are adapted to marine ecosystems, spending time in nearshore areas and estuaries or in coral reef ecosystems. Other reptiles have also demonstrated great success in marine habitats. The sea turtles, represented by two families, Cheloniidae (5 species) and Dermochelyidae (1 species), are found globally in temperate and tropical oceans, and many species migrate long distances to reach nesting beaches. The Galápagos marine iguana (*Amblyrhynchus cristatus*) lives in colonies along the rocky shorelines of the Galápagos Islands and dives into the surf to graze on dense algal mats.

Position of birds and other controversies. The study of the evolutionary course, or phylogeny, of various vertebrate groups has seen an era of change recently. The use of cladistic analysis (grouping organisms by shared common descent), particularly using molecular data (DNA) as a variable, has yielded very different hypotheses regarding the relationships of the vertebrate groups. Recent hypotheses place birds within the reptile clade (a single group of organisms related by a single common ancestor). Although this idea has met resistance, most scientists are now beginning to accept this significant step. This results in a fundamental change in our perspectives of what it means to be reptile. Specifically, scales have always been considered a fundamental characteristic of reptiles, but evolutionary morphologists point out that bird feathers are, in essence, nothing more than highly modified scales. Our view of the reptiles as ectotherms is also being modified and expanded. The metabolic physiology of reptiles is now best described as a continuum, from ectothermic to endothermic (the predominant amount of internal body heat generated by physiological processes). Another proposal that has generated great debate places turtles among the diapsid reptiles alongside squamates, crocodylians and birds. At present this step is a point of continued debate and research. Due to these new hypotheses, the taxonomy of the Reptilia is currently unresolved and in transition.   W. Ben Cash

Bibliography. S. B. Hedges and L. L. Poling, A molecular phylogeny of reptiles, *Science*, 283:998–1001, 1999; M. Laurin and R. R. Reisz, A re-evaluation of early amniote phylogeny, *Zoolog. J. Linnaean Soc.*, 113:105–225, 1995; D. W. Linzey, *Vertebrate Biology*, McGraw-Hill, 2001; F. H. Pough et al., *Herpetology*, 3d ed., Prentice-Hall, Upper Saddle River, New Jersey, 2003; O. Rieppel and R. R. Reisz, The origin and early evolution of turtles, *Ann. Rev. Ecol. Systematics*, 30:1–22, 1999; G. R. Zug, L. J. Vitt, and J. P. Caldwell, *Herpetology*, Academic Press, San Diego, 2001.

Repulsion motor

An alternating-current (ac) commutator motor designed for single-phase operation. The chief distinction between the repulsion motor and the single-phase series motors is the way in which the armature receives its power. In the series motor the armature power is supplied by conduction from the line power supply. In the repulsion motor, however, armature power is supplied by induction (transformer action) from the field of the stator winding. For discussion of the ac series motor *see* UNIVERSAL MOTOR. *See also* ALTERNATING-CURRENT MOTOR.

The repulsion motor primary or stationary field winding is connected to the power supply. The secondary or armature winding is mounted on the motor shaft and rotates with it. The terminals of the armature winding are short-circuited through a commutator and brushes. There is no electrical contact between the stationary field and rotating armature (**Fig. 1**). *See* WINDINGS IN ELECTRIC MACHINERY.

If the motor is at rest and the field coils are energized from an outside ac source, a current is induced in the armature, just as in a static transformer. If the brushes are in line with the neutral axis of the magnetic field, there is no torque, or tendency to rotate. However, if they are set at a proper angle (generally 15–25° from the neutral plane), the motor will rotate.

Repulsion motors may be started with external resistance in series with the motor field, as is done with dc series motors. A more common method is to start the motor with reduced field voltage and increase the voltage as the motor increases speed. This can be done with a transformer having an adjustable tapped secondary or a variable autotransformer.

It is also possible to doubly excite the motor; that is, the armature may receive its power not only by induction from the stator winding but also by conduction from a transformer with adjustable taps, as shown in **Fig. 2**.

Repulsion-start, induction-run motor. This motor (**Fig. 3**a) possesses the characteristics of the repulsion motor at low speeds and those of the induction motor at high speed. It starts as a repulsion motor. At a predetermined speed (generally at about two-thirds of synchronous speed) a centrifugal device lifts the brushes from the commutator and short-circuits the armature coils, producing a squirrel-cage rotor. The motor then runs as an induction motor. In Fig. 3b the curve AB represents the characteristics of an induction motor and curve CD a repulsion motor. The solid curve AD is the combined characteristic of a repulsion-start, induction-run motor. *See* INDUCTION MOTOR.

Repulsion-induction motor. This motor (**Fig. 4**a) is very similar to the standard repulsion motor in construction, except for the addition of a second separate high-resistance squirrel cage on the rotor. Both rotor windings are torque-producing, and the total torque produced is the sum of the individual torques developed in these two windings. In Fig. 4b, curve A is the characteristic of the repulsion-motor torque developed in this motor. Curve B represents the induction-motor torque. Curve C is the combined

Fig. 1. Schematic diagram of a repulsion motor.
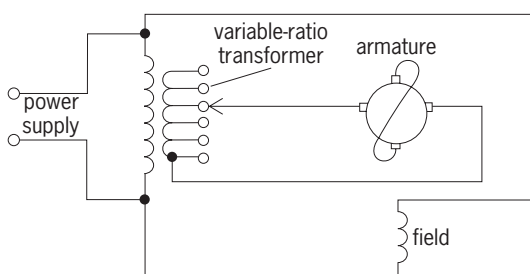
Fig. 2. Doubly excited repulsion motor.

Fig. 3. Repulsion-start, induction-run motor. (*a*) Schematic diagram. (*b*) Speed-torque-characteristic.

Fig. 4. Repulsion-induction motor. (*a*) Schematic diagram. (*b*) Speed-torque characteristic.

total torque of the motor. The advantages of this machine are its high starting torque and good speed regulation. Its disadvantages are its poor commutation and high initial cost.   Irving L. Kosow

Bibliography. I. L. Kosow, *Electric Machinery and Transformers*, 2d ed., 1991; G. McPherson and R. D. Larramore, *An Introduction to Electrical Machines and Transformers*, 2d ed., 1990; M. G. Say, *Alternating-Current Machines*, 5th ed., 1984.

Reservoir

A place or containment area where water is stored. Where large volumes of water are to be stored, reservoirs usually are created by the construction of a dam across a flowing stream. When water occurs naturally in streams, it is sometimes not available when needed. Reservoirs solve this problem by capturing water and making it available at later times. *See* DAM.

In addition to large reservoirs, many small reservoirs are in service. These include varieties of farm ponds, regulating lakes, and small industrial or recreational facilities. In some regions, small ponds are called tanks. Small reservoirs can have important cumulative effects in rural regions.

Purposes. Reservoirs can be developed for single or multiple purposes, such as to supply water for people and cities, to provide irrigation water, to lift water levels to make navigation possible on streams, and to generate electricity.

When used to supply water for a city, a reservoir can provide water during all parts of the year, and can even make it possible to populate dry regions. If cities have underground water available, they may rely on it; otherwise, they must have stored surface water from a reservoir, or they risk running out of drinking water.
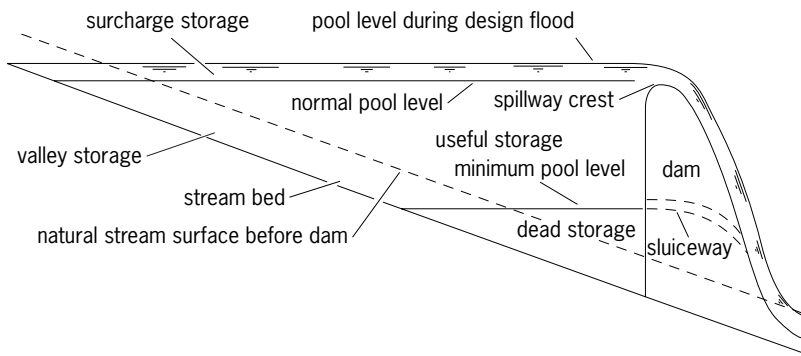
Fig. 1. Zones of storage in a reservoir. (*After R. K. Linsley et al., Water-Resources Engineering, McGraw-Hill, 1992*)

A reservoir can supply water for irrigation of crops at precisely the time required for the maximum plant growth. Many reservoirs in dry regions capture water during wet periods and release it to farm fields at times of plant growth and stress.

Reservoirs may provide for energy generation (hydropower) through run-of-river hydroelectric plants which generate power when the water is available; or if the storage is adequate, they may provide water for energy generation at any time it is needed. Another way that reservoirs aid in energy management is through pumped storage, where electric energy is stored by pumping water to a reservoir, and releasing the water later to flow downward through a generating system. Water from reservoirs is also used to cool thermal or nuclear electricity-generating plants. *See* PUMPED STORAGE.

Reservoirs also make navigation possible by maintaining river depths at required levels, usually by means of locks and dams. *See* RIVER ENGINEERING.

Another purpose of reservoirs is to control floods by providing empty spaces for flood waters to fill, thereby diminishing the rate of flow and water depth downstream of the reservoir. Such use of reservoirs has saved many lives and goods from flood damage, but reservoirs cannot control all flood problems.

Reservoirs also provide for environmental uses of water by providing water to sustain fisheries and meet other fish and wildlife needs, or to improve water quality by providing dilution water when it is needed in downstream sections of rivers. Reservoirs may also have esthetic and recreational value, providing boating, swimming, fishing, rafting, hik-

Fig. 2. Sediment accumulation in a typical reservoir. (*After R. K. Linsley et al., Water-Resources Engineering, McGraw-Hill, 1992*)

ing, viewing, photography, and general enjoyment of nature.

Characteristics and configuration. Generally, a reservoir is divided into zones of water that are reserved for different uses (**Fig. 1**). The zone for useful storage (also called working storage, multiple-purpose capacity, or operating storage) provides space for water storage applications. In this zone it is normal for the water level to rise and fall during the year. Sometimes people who rely on reservoirs only for recreational use complain about this, because they do not understand the necessity of using the full storage zone.

The surcharge storage zone enables the reservoir to be used for flood control. If flood control is an important purpose, part of the useful storage could also be devoted to that use. Surcharge storage occurs above the spillway crest, which is the elevation where water just begins to flow over the spillway's upper surface. The dam is designed so that the design flood, a maximum flood event, safely passes over the spillway without placing the dam at risk. The dead storage, normally reserved to be filled with sediment, occurs below the normal release elevation for water.

Planning. Planning for reservoir size takes into account the statistical variation of future inflows and water demands.

The reservoir sizing process carries the risk that the capacity will be too small to meet the purposes or too large for the reservoir to fill. Therefore, the planner compiles as much historical data as possible and makes studies of how the planned reservoir would have performed if it had been in place during the historical period. These are sometimes referred to as "what if" studies. Reservoirs are complex, however, because the water supplied by the stream system is not fully predictable and is subject to variations from high to low flow periods. Thus, the water reservoir is made large enough to compensate for the risk factors.

After the location of a reservoir is determined, many aspects of the construction process must be settled through the design process, an effort that involves engineers, geologists, hydrologists, and other professionals. A major consideration is the necessity to make the dam as safe as possible to avoid placing people downstream at risk due to dam failure. While there have been few dam failures, when they do occur they may cause unacceptable levels of damage.

After a dam is built, it must be operated correctly. The key person is the operator who makes decisions about when to release or store water. In the past, reservoir operating decisions were made by rule curves, which provided the operator with simple guidelines about how much water to release and what lake levels to maintain. As the science of forecasting and the use of computers became more complex, however, reservoir operation became more sophisticated. It is not uncommon to have a reservoir control center where operators use computers to monitor weather forecasts furnished from satellite data and to simulate future demands for water
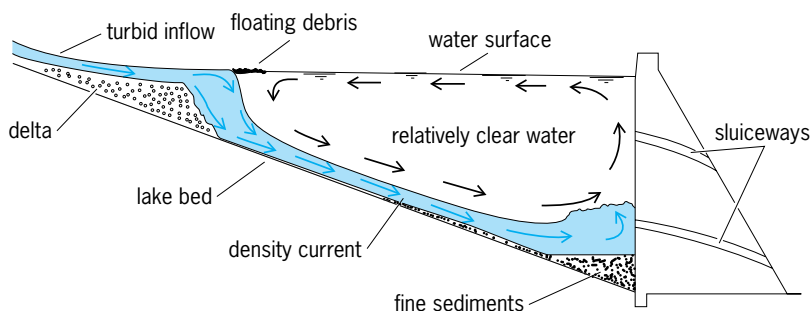
in order to make decisions about water releases. They may also be bound by legal requirements to release water for downstream users, including fish and wildlife. As the demands for water increase in developed areas, conflicts over use of the water arise, especially in water-short areas. This requires the reservoir owner or management agency to develop conflict-resolution procedures.

Reservoirs require continuing attention from owners and managers. Problems may include excessive deposition of sediments on the lake bottom which reduce the capacity to store water (**Fig. 2**), pollution of lake waters, eutrophication (aging of waters with excessive growth of algae), shoreline protection, and associated issues such as dam leakage or settlement. Sedimentation, a naturally occurring phenomenon, is a particularly difficult problem that is impossible to fully control, and it may ruin a reservoir's capacity to store water. *See* EUTROPHICATION; WATER POLLUTION.

Controversies. The construction or alteration of reservoirs is often resisted by some people because of perceived negative side effects. One side effect is reservoir evaporation, which can consume waters that could otherwise be used in the river environment for fish and wildlife or for flushing salts through the stream system. As a result of evaporation, the microclimate around the reservoir might even be altered. Seepage caused by the reservoir impoundment can change the local patterns of underground water. In the valley floor, settlement can occur because of the weight of the water in storage. Ecosystems can be changed because the schedule of water release and the quality of the water can be altered. Also, when reservoirs are built, large numbers of residents may have to be resettled. *See* WATER SUPPLY ENGINEERING.   Neil S. Grigg

Bibliography. N. S. Grigg, *Water Resources Management*, 1996; R. K. Linsley et al., *Water-Resources Engineering*, 1992; D. R. Maidment (ed.), *Handbook of Hydrology*, 1993.

Resin

Originally a category of vegetable substances soluble in ethanol but insoluble in water, but generally in modern technology an organic polymer of indeterminate molecular weight. The class of flammable, amorphous secretions of conifers or legumes are considered true resins, and include kauri, copal, dammar, mastic, guaiacum, jalap, colophony, shellac, and numerous less well-known substances. Water-swellable secretions of various plants, especially the Burseraceae, are called gum resins and include myrrh and olibanum. The official resins, according to the *United States Pharmacopeia*, are benzoin, guaiac, mastic, and resin. Other natural resins are copal, dammar, dragon's blood, elaterium, lac, and sandarac. The natural vegetable resins are largely polyterpenes and their acid derivatives, which find application in the manufacture of lacquers, adhesives, varnishes, and inks.

The synthetic resins, originally viewed as substitutes for copal, dammar, and elemi natural resins, have a large place of their own in industry and commerce. Phenol-formaldehyde, phenol-urea, and phenol-melamine resins have been important commercially for a long time. Any unplasticized organic polymer is considered a resin, thus nearly any of the common plastics may be viewed as a synthetic resin. Water-soluble resins are marketed chiefly as substitutes for vegetable gums and in their own right for highly specialized applications. Carboxymethylcellulose, hydroxyalkylated cellulose derivatives, modified starches, polyvinyl alcohol, polyvinylpyrrolidone, and polyacrylamides are used as thickening agents for foods, paints, and drilling muds, as fiber sizings, in various kinds of protective coatings, and as encapsulating substances. *See* POLYMER.
                                                        Frank Wagner

Resistance heating

The generation of heat by electric conductors carrying current. The degree of heating for a given current is proportional to the electrical resistance of the conductor. If the resistance is high, a large amount of heat is generated, and the material is used as a resistor rather than as a conductor. *See* ELECTRICAL RESISTANCE.

Resistor materials. In addition to having high resistivity, heating elements must be able to withstand high temperatures without deteriorating or sagging. Other desirable characteristics are low temperature coefficient of resistance, low cost, formability, and availability of materials. Most commercial resistance alloys contain chromium or aluminum or both, since a protective coating of chrome oxide or aluminum oxide forms on the surface upon heating and inhibits or retards further oxidation. Some commercial resistor materials are listed in **Table 1**.

Heating element forms. Since heat is transmitted by radiation, convection, or conduction or combinations of these, the form of element is designed for the major mode of transmission. The simplest form is the helix, using a round wire resistor, with the pitch of the helix approximately three wire diameters. This form is adapted to radiation and convection and is generally used for room or air heating. It is also used in industrial furnaces, utilizing forced convection up to about 1200°F (650°C). Such helixes are stretched over grooved high-alumina refractory insulators and are otherwise open and unrestricted. These helixes are suitable for mounting in air ducts or enclosed chambers, where there is no danger of human contact. *See* ELECTRIC FURNACE; INDUCTION HEATING.

For such applications as water heating, electric range units, and die heating, where complete electrical isolation is necessary, the helix is embedded in magnesium oxide inside a metal tube, after which the tube is swaged to a smaller diameter to compact the oxide and increase its thermal conductivity. Such units can then be formed and flattened to desired shapes. The metal tubing is usually copper for water

TABLE 1. Electric furnace resistor materials and temperature ranges

	Maximum resistor temperature, °F (°C)	
Material, major elements	In air	In reducing or neutral atmosphere
35% nickel, 20% chromium*	1900 (1040)	2100 (1150)
60% nickel, 16% chromium*	1800 (980)	2000 (1090)
68% nickel, 20% chromium, 1% cobalt*	2250 (1230)	2250 (1230)
78% nickel, 20% chromium	2250 (1230)	2250 (1230)
15% chromium, 4.6% aluminum*	2050 (1120)	Not used
22.5% chromium, 4.6% aluminum*	2150 (1180)	Not used
22.5% chromium, 5.5% aluminum*	2450 (1340)	Not used
Silicon carbide	2800 (1540)	2500 (1370)
Platinum	2900 (1590)	2900 (1590)
Molybdenum[†]	Not used	3400 (1870)
Tungsten[†]	Not used	3700 (2040)
Graphite[†]	Not used	5000 (2760)

*Balance is largely iron, with 0.5–1.5% silicon.
[†]Usable only in pure hydrogen, nitrogen, helium, or oxygen or in vacuum because of inability to form protective oxide.

heaters and stainless steel for radiant elements, such as range units. In some cases the tubes may be cast into finned aluminum housings, or fins may be brazed directly to the tubing to increase surface area for convection heating.

Modification of the helix for high-temperature furnaces involves supporting each turn in a grooved refractory insulator, the insulators being strung on stainless alloy rods. Wire sizes for such elements are $\frac{3}{16}$ in. (5 mm) in diameter or larger, or they may be edge-wound strap. Such elements may be used up to 1800°F (980°C) furnace temperature. *See* REFRACTORY.

Another form of furnace heating element is the sinuous grid element, made of heavy wire or strap or casting and suspended from refractory or stainless supports built into the furnace walls, floor, and roof. Some of these forms are shown in **Fig. 1**.

Silicon carbide elements are in rod form, with low-resistance integral terminals extending through the furnace walls, as shown in **Fig. 2**.

Direct heating. When heating metal strip or wire continuously, the supporting rolls can be used as electrodes, and the strip or wire can be used as the resistor.

In **Fig. 3** the electric current passes from roll A through the strip or wire to roll B. Heating by this method can be very rapid. Disadvantages are that the electric currents are large, and uniform contact between the strip and the rolls is difficult to
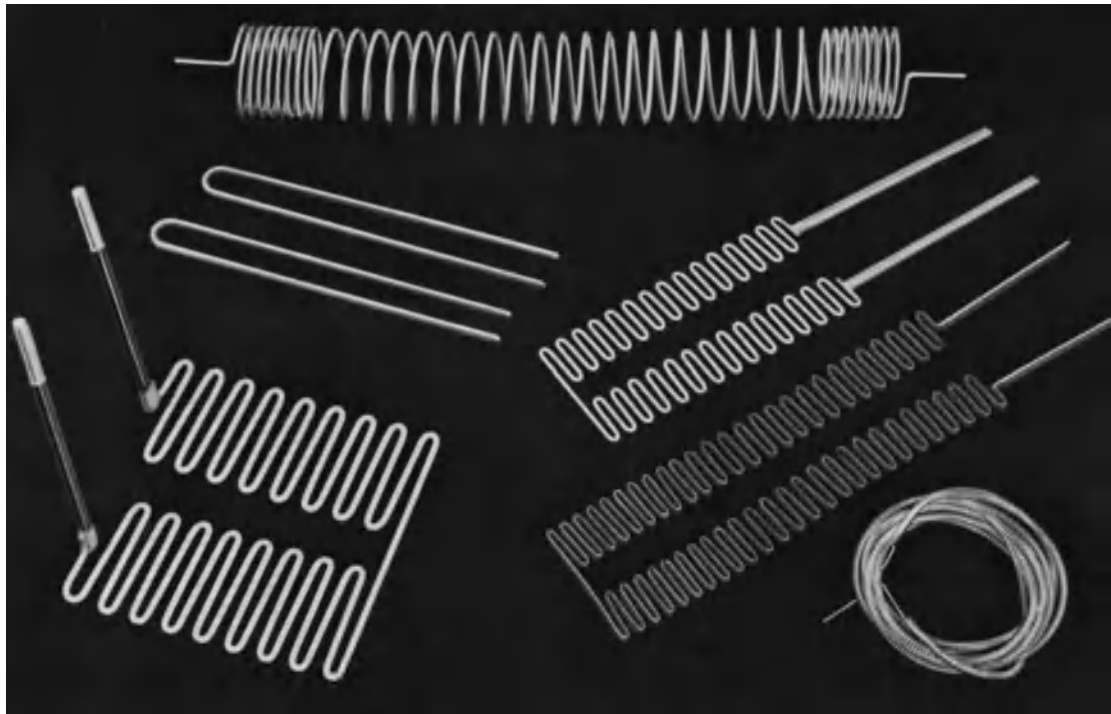
Fig. 1. Metallic resistor forms.

maintain, since both surfaces must be clean and free of oxides. For these reasons, direct heating is not used extensively.

Molten salts. The electrical resistance of molten salts between immersed electrodes can be used to generate heat. Limiting temperatures are dependent on decomposition or evaporization temperatures of the salt. Parts to be heated are immersed in the salt. Heating is rapid and, since there is no exposure to air, oxidation is largely prevented. Disadvantages are the personnel hazards and discomfort of working close to molten salts.

Major applications. A major application of resistance heating is in electric home appliances, including electric ranges, clothes dryers, water heaters, coffee percolators, portable radiant heaters, and hair dryers. Resistance heating has also found application in home or space heating; some homes are designed with suitable thermal insulation to make electric heating practicable. A general rule for such insulation is that heat loss be restricted to 10 W/ft^2 (108 W/m^2) of floor living area at 75°F (42°C) temperature difference, inside to outside. This applies to areas having 5000–7500 degree days annually. The degree day is the difference between the 24-h average outdoor temperature and 65°F (18.3°C). For instance, if the 24-h average were 40°F (4.4°C), there would be 25 degree days for that day. This includes also a three-quarter air change per hour, except in the basement, where one-quarter air change per hour is calculated. The values in **Table 2** are applied to determine overall heat loss. *See* DEGREE-DAY.

Heating system capacity. For rooms normally kept warm, the installed capacity should be 20% higher than the calculated losses. For rooms intermittently heated or for entries with frequent door openings, the installed capacity should be 50% more than the calculated losses.

A variety of electric heating units may be used for home heating. The simplest is a grid of helical resistance coils mounted in the central heating unit, with a blower to circulate the warm air through the rooms, and adjustable louvers in each room to control the temperature. Another type is the baseboard heater, consisting of finned sheathed helical
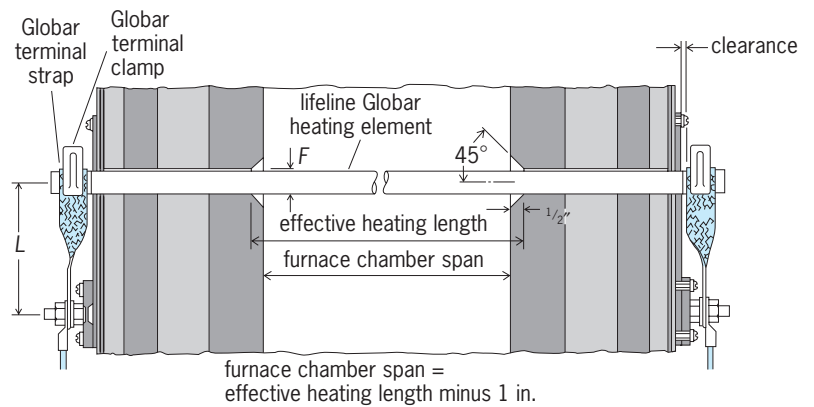
Fig. 2. Silicon carbide heating element. 1 in. = 2.5 cm.
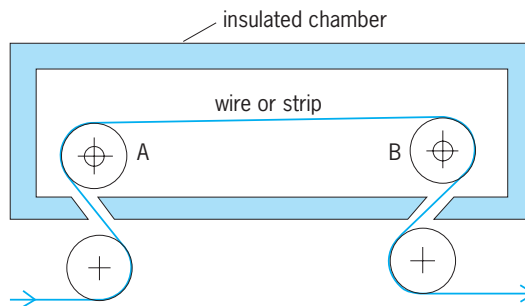
Fig. 3. Direct heating of metal strip or wire.

resistance coils, with individual room control (thermostat in each room). Still another room unit is the glass or ceramic panel in which resistance wires are embedded. Such panels are designed to radiate infrared wavelengths and are particularly suitable for glassed-in areas or bathrooms. Mounted usually in the ceiling, such panels neutralize the chilling effect of large glass exposures, thus preventing drafts. *See* THERMOSTAT.

For isolated rooms or work areas, integrated units which combine resistors, circulating fans, and magnetic contactors operated by thermostats open and close the power circuits to the heating elements to maintain desired temperature. Where central heating units involving several kilowatts of energy are used, the contactors are arranged in multiple circuits and close in sequence to minimize the transient voltage effects of sudden large energy demands.

Ovens and furnaces. If the resistor is located in a thermally insulated chamber, most of the heat generated is conserved and can be applied to a wide variety of heating processes. Such insulated chambers are called ovens or furnaces, depending on the temperature range and use.

The term oven is generally applied to units which operate up to approximately 800°F (430°C). Ovens use rock wool or glass wool between the inner and outer steel casings for thermal insulation. Typical uses are for baking or roasting foods, drying paints and organic enamels, baking foundry cores, and low-temperature treatments of metals.

The term furnace generally applies to units operating above 1200°F (650°C). In these the thermal

TABLE 2. Maximum recommended heat loss	
Structure	Heat loss, W/(ft^2)(°F) [W/(m^2)(°C)]
Ceiling	0.015 (0.29)
Wall	0.021 (0.41)
Floor	0.021 (0.41)
Floor slab perimeter	12.0/lineal foot (71/lineal meter)
Basement walls	0.045 (0.77)
Basement walls, 25% of wall aboveground	0.026 (0.50)

Requisites:
Windows—double-glazed or storm
Doors—storm doors required
Attic—ventilating area of 1 ft^2/150 ft^2 or 1 m^2/150 m^2 ceiling area
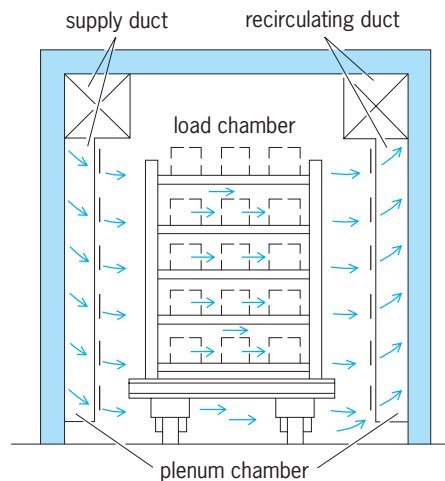
Fig. 4. Typical oven for maximum temperature of 800°F (430°C); heating is by forced convection. (*Carl Mayer Corp.*)

insulation is made up of an inner wall of fireclay, kaolin, or high-alumina or zirconia brick, depending on the temperature, with secondary insulating blocks made from such base materials as rock wool, asbestos fiber, and diatomaceous earth. Typical uses
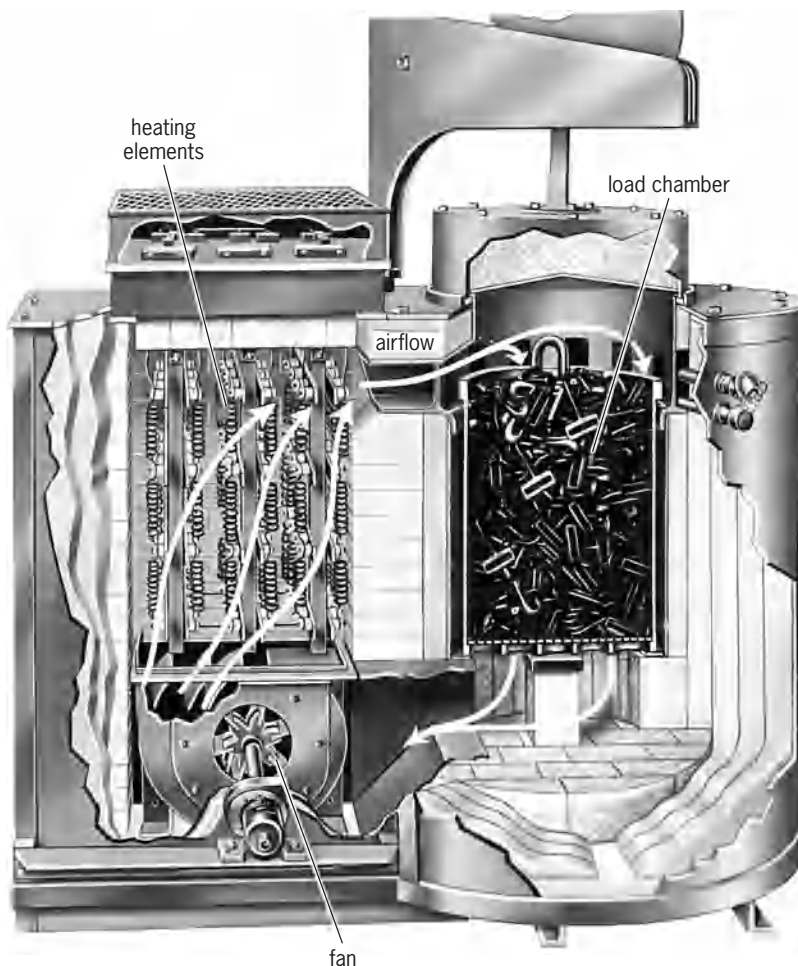

Fig. 5. Low-temperature pit-type furnace, using forced convection and operating at temperatures of 250–1400°F (120–760°C). (*Lindberg Industrial Corp.*)

of furnaces are for heat treatment or melting of metals, for vitrification and glazing of ceramic wares, for annealing of glass, and for roasting and calcining of ores. *See* ELECTRIC HEATING.

Actually, ovens and furnaces overlap in temperature range, with ovens being used at temperatures as high as 1000°F (540°C), and furnaces as low as 250°F (120°C). Electrically heated ovens and furnaces have advantages over fuel-fired units. These advantages often compensate for the generally higher cost of electric energy. The main advantages are (1) ease of distributing resistors, or heating elements, to obtain a uniform temperature in the product being heated; (2) ease of operation, since adjustments by operators are usually unnecessary; (3) cleanliness; (4) comfort, since heat losses are low and there are no waste fuel products; (5) adaptability to the use of controlled furnace atmospheres or vacuum, and (6) high temperatures beyond the range attainable with commercial fuels.

Heat transfer in ovens and furnaces. Heating elements mounted in ovens and furnaces may be located to radiate directly to the parts being heated or may be located behind baffles or walls so that direct radiation cannot take place. The heat then is transferred by circulating the furnace air or gas. Determination of which method or combination of methods to use is based on temperature uniformity, speed of heating, and high-temperature strength limitations of fans or blowers. At temperatures below 1200°F (650°C) radiation is slow, and virtually all ovens and furnaces in this temperature range use forced convection or circulation. From 1200 to 1500°F (650 to 820°C), radiation is increasingly effective, while reduced gas density and lower fan or blower speeds (because of reduced strength at high temperatures) make forced convection less effective. Therefore in this temperature range, combinations of radiation and forced convection are used. Above 1500°F forced convection is employed only when direct radiation cannot reach all parts of the loads being heated. An example of this is a container filled with bolts or a rack filled with gears.

Obviously, in vacuum furnaces, heat transfer can be only by radiation and conduction.

Ovens. **Figure 4** shows a typical oven, for a maximum temperature of 800°F (430°C); heating is by forced convection. The heated air enters through the supply duct, passing down into the plenum chamber and distributing louvers into and across the load chamber and returning through the right-side plenum to the recirculating duct. The external heater and blower, completing the circuit, are shown.

Low-temperature furnace. **Figure 5** shows a pit-type furnace used for tempering steel. It operates at temperatures of 250–1400°F (120–760°C) and uses forced convection.

The steel parts to be tempered are placed in the load chamber at the right. The electrical heating elements are in the heating chamber at the upper left. The fan, at the lower left, circulates air upward over the heating elements and into the load chamber. The hot air is forced down through the load and back to

the fan. The heating chamber is thermally insulated from the load chamber so there is no direct radiation. Such furnaces are designed commercially to hold the load temperature within 10° of the control temperature, and can be designed for closer control if desired.

High-temperature furnace. **Figure 6** shows the interior of a pit-type carburizing furnace, which uses radiant-heating elements in the form of corrugated metal bands mounted on the inside of the brick walls. The work basket (not shown) rests on the load support at the bottom. These elements are designed to operate at low voltage (approximately 30 V) so that soot deposits from the carburizing gases will not cause short circuits. *See* FURNACE CONSTRUCTION.

Electrical input. Electric ovens and furnaces are rated in kilowatts. The electrical input is determined from the energy absorption Q of the load and the thermal losses.

The average rate of energy flow into load during the heating period is Q/t, where Q is in kilowatt-hours and t is the heating time in hours.

However, the rate of energy flow into the load is high when the load is cold and decreases as the load temperature approaches furnace temperature. Therefore the energy input must be high enough to take care of the high initial heating rate. This leads to the following approximate formulas for the input, in which L is the thermal losses in kilowatts at the operating temperature:

$$\text{Batch furnaces: input in kW} = L + 1.5\, Q/t$$

$$\text{Continuous furnaces: input in kW} = L + 1.25\, Q/t$$

Operating voltage for heating elements. Heating elements are usually designed to operate at standard service voltages of 115, 230, or 460 V, if two conditions can be satisfied. These are, first, that the heating element is sufficiently heavy in cross section to avoid sagging or deformation in service, and second, that there is no appreciable electrical leakage through the furnace refractories tending to short-circuit the heating elements. The latter consideration generally limits voltages to 260 volts at 2100°F (1150°C) and to approximately 50 volts at 3200°F (1700°C), because the refractory walls of the furnace become increasingly better conductors at higher temperatures.

In vacuum furnaces, voltages are limited to 230 V by the tendency to break down into glow discharge at low pressure.

Temperature control. Almost all commercial electric ovens and furnaces have automatic temperature control. The simplest control uses magnetic contactors which open and close the circuit to the heating elements in response to temperature signals from control thermostats or thermoelectric pyrometers. A refinement of this ON-and-OFF control is to modulate it with a timer, with the ON period becoming a progressively smaller percentage as the control temperature is approached. This prevents the overshooting which results from thermal lag in the control thermocouple and the furnace.

Fig. 6. Interior of high-temperature pit-type carburizing furnace, using radiant heating. (*Lindberg Industrial Corp.*)

True proportioning control is achieved through the use of saturable reactors in series with the heating elements. **Figure 7** shows the arrangement used.

The pyrometer, through the amplifier, controls the direct current to the control winding of the reactor. When the direct current is maximum, the reactor offers virtually no impedance to the alternating current flowing through the heating element, and the normal amount of heat is generated. As control temperature is approached, the direct current is decreased, increasing the impedance of the reactor and reducing the current to the heating elements until equilibrium is reached.
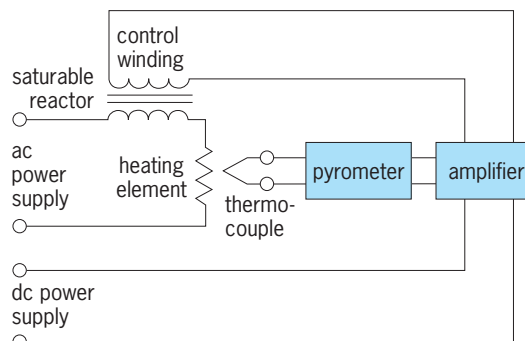
Fig. 7. Temperature control using saturable reactor.

The solid-state silicon controlled rectifier is now used as a heating contactor. Compared to magnetic contactor, the SCR has no moving parts, lower maintenance costs, and silent, vibration-free operation. Moreover, it is directly adaptable to proportioning control, obviating the need for saturable reactors. In this latter case it is cost-competitive. *See* SEMICONDUCTOR RECTIFIER.

Snow and ice melting. Low-temperature resistors with moisture-proof insulation are embedded in concrete sidewalks or strung in roof gutters to melt snow and ice as needed. Such cables also are used to prevent freezing and exposed water pipes. Analogously, very thin resistance wires may be cast in automobile windows to melt snow and ice.

Metal-sheathed units are used at railroad switches and dam locks to keep the moving parts free in freezing temperatures.

Die and platen heating. Metal-sheathed elements are embedded in grooves or holes in dies or platens to maintain these parts at desired temperatures for hot processing or molding of plastics.

Cryogenics. Metal-sheathed immersion-type units are used to supply heat of evaporation for liquefied gases such as argon, helium, nitrogen, hydrogen, and oxygen. Metal sheathing is usually of stainless steel or aluminum to withstand low temperatures. Gaskets are of poly(tetrafluoroethylene) for the same reason.   Willard Roth

Bibliography. Chemical Rubber Publishing Co., *Handbook of Chemistry and Physics*, annually; C. J. Erikson, *Handbook of Electrical Heating forIndustry*, IEEE 1994; D. G. Fink and H. W. Beatty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; B. Parsons (ed.), *ASHRAE Handbook: Heating, Ventilating, and Air-Conditioning Systems and Equipment*, 1992.

Resistance measurement

The quantitative determination of that property of an electrically conductive material, component, or circuit called electrical resistance. The ohm, which is the International System (SI) unit of resistance, is defined through the application of Ohm's law as the electric resistance between two points of a conductor when a constant potential difference of 1 volt applied to these points produces in the conductor a current of 1 ampere. Ohm's law can thus be taken to define resistance R as the ratio of dc voltage V to current I, Eq. (1). For bulk metallic conductors, for ex-

$$R = \frac{V}{I} \tag{1}$$

ample, bars, sheets, wires, and foils, this ratio is constant. For most other substances, such as semiconductors, ceramics, and composite materials, it may vary with voltage, and many electronic devices depend on this fact. The resistance of any conductor is given by the integral of expression (2), where l is the

length, A the cross-sectional area, and ρ the resistiv-

$$\int_0^l \frac{\rho \, dl}{A} \tag{2}$$

ity. *See* ELECTRICAL CONDUCTIVITY OF METALS; ELECTRICAL RESISTANCE; ELECTRICAL RESISTIVITY; OHM'S LAW; SEMICONDUCTOR.

Realization of the ohm. The ohm can be realized by means of a calculable capacitor and a series of ac and dc bridges. Following international agreement, the ohm has been maintained since January 1, 1990, with reference to a quantized Hall resistance.

Calculable capacitor. The calculable capacitor enables the ohm to be realized independently of voltage and current. The dielectric constant of free space ϵ_0 is given by Eq. (3), where $\mu_0 = 4\pi \times 10^{-7}$ henry per

$$\epsilon_0 = \frac{1}{c^2 \mu_0} \tag{3}$$

meter (a consequence of the definition of the ampere), and c is the velocity of light, which has been declared by the International Committee for Weights and Measures (CIPM) to be 299,792,458 m/s. Given this value, the capacitance in farads of any assembly of conductors can be calculated through the use of electromagnetic theory. The Thomson-Lampard theorem in electrostatics gives the result that the capacitance per unit length between the opposite members of cylinders arranged in a symmetrical quadrupole (**Fig. 1**a) is given by Eq. (4), a result which is
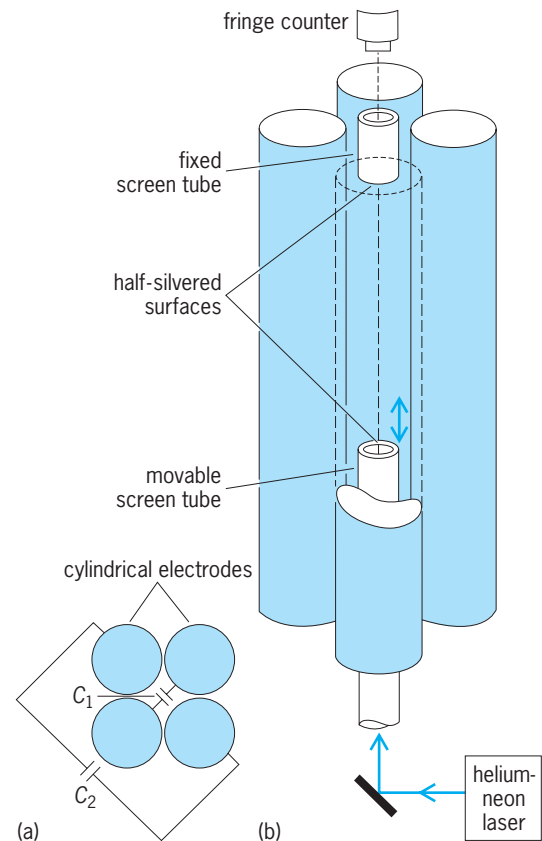
Fig. 1. Thomson-Lampard calculable capacitor. (*a*) Cross section of electrodes. (*b*) Arrangement of components.
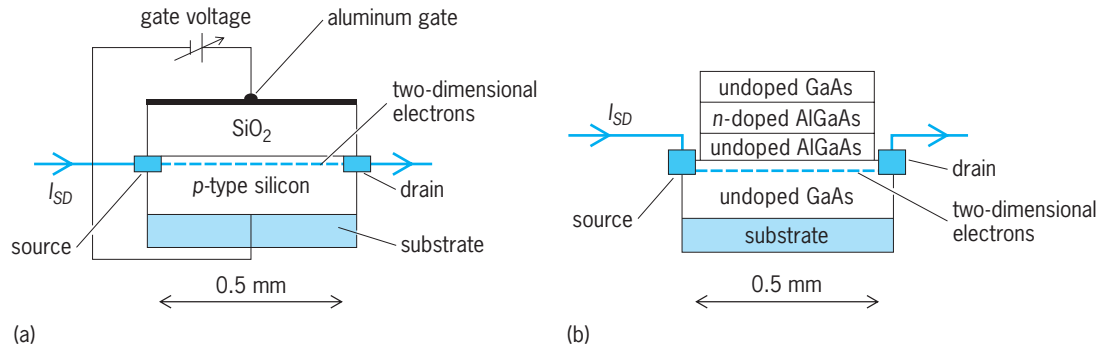
Fig. 2. Two types of semiconductor devices having quasi-two-dimensional electron layers. I_{SD} = current between source and drain electrodes. (*a*) Silicon MOSFET. (*b*) Heterostructure. (*After A. Hartland, Quantum standards for electrical units, Contemp. Phys., 29:477–498, 1988*)

independent of constant cross-sectional dimensions

$$C = \frac{\epsilon_0}{\pi}(\log_e 2) \qquad \text{farads/meter} \qquad (4)$$

of the apparatus. This result is used to construct a calculable capacitor, in which the active length is defined by two central screen tubes, one of which can be moved along the axis. The change in the capacitance between either of the opposite pairs of conducting cylinders is then given by Eq. (3). The change in effective length is measured by arranging a Fabry-Perot interferometer between the faces of the screen tubes (Fig. 1*b*).

Several such instruments exist, and produce a change in capacitance of about 0.4 picofarad with an uncertainty of about 3 parts in 10^8. Through the use of accurate transformer bridges, this change in capacitance is used to establish the value of a 1-nanofarad capacitor, and the reactance of this at a frequency of 1592 Hz (10,000 radians/s) is then used to establish the resistance value of 100 kilohms at the same frequency. Again using transformer bridges, the value of 1000 ohms (Ω) is obtained, and this in turn is used to establish a 1000-Ω dc value by using a coaxial resistor having a single thin-wire element which permits accurate calculation of the difference between the resistance values at 1592 Hz and dc. The final step from 1000 Ω to 1 Ω is made by dc comparison techniques, for example, through the use of a series-parallel Hamon network of 32×32 Ω resistors, discussed below, or a cryogenic current comparator. The overall uncertainties of this type of experiment are such that, for example, the realizations of the ohm in the United States and the United Kingdom are in agreement to within 3 parts in 10^7. *See* BRIDGE CIRCUIT; CAPACITANCE MEASUREMENT; CURRENT COMPARATOR; ELECTRICAL UNITS AND STANDARDS; INTERFEROMETRY; RADIO-FREQUENCY IMPEDANCE MEASUREMENTS; RESISTOR.

Quantized Hall resistance. The quantum Hall effect occurs at the interface between two suitable semiconducting layers at low temperatures (less than 4.2 K or −452°F). Under these conditions, a narrow potential well is formed and a two-dimensional layer of electrons is confined at the interface. Although the effect was first discovered in silicon metal-oxide-

semiconductor field-effect transistor (MOSFET) devices, the most commonly used system is a gallium arsenide–aluminium gallium arsenide heterostructure (**Fig. 2**). The inset in **Fig. 3** shows a typical device configuration. Any of the three electrode pairs, 1–4, 2–5, and 3–6, permit measurement of the transverse Hall voltage, and the Hall resistance R_H is defined as the ratio of this voltage to the current between the source electrode and the drain electrode. If a current of less than 100 microamperes flows between the source and drain electrodes, the Hall voltage and hence the Hall resistance when measured as a function of magnetic field exhibit plateaus (Fig. 3). These plateaus occur at values of the Hall resistance given by Eq. (5), where i is an integer, and

$$R_H = \frac{R_K}{i} \qquad (5)$$

R_K is the von Klitzing constant and is believed to be equal to h/e^2, h being the Planck constant and e the elementary charge. The International Committee for
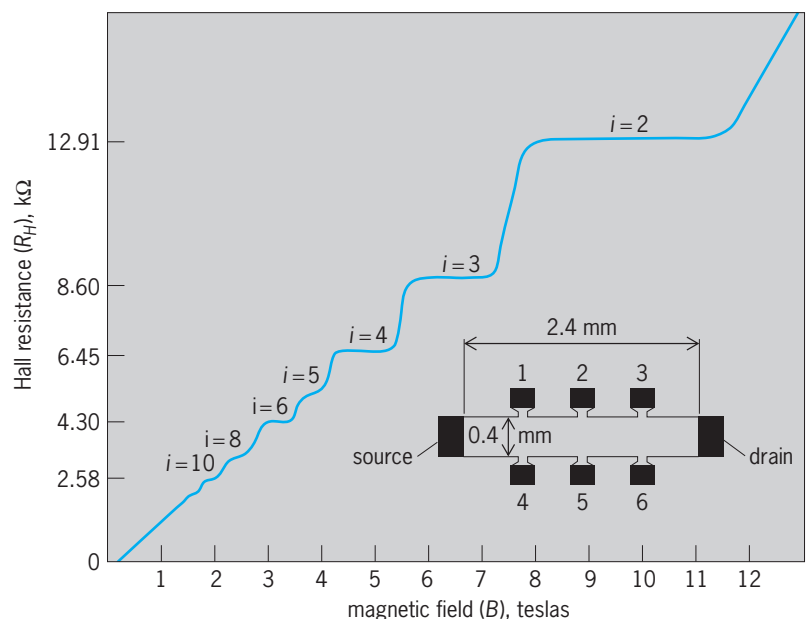
Fig. 3. Quantized Hall resistance in a gallium arsenide–aluminum gallium arsenide heterostructure at a temperature of 0.33 K. Inset is a plan view of the device. (*After A. Hartland, Quantum standards for electrical units, Contemp. Phys., 29:477–498, 1988*)

Weights and Measures has recommended that R_K be assigned the value $25,812.807$ $(1 \pm 0.2 \times 10^{-6})$ Ω (one standard deviation uncertainty). This value was obtained mainly from measurements referred to a calculable capacitor and partly from measurements of fundamental constants such as the fine-structure constant. *See* FUNDAMENTAL CONSTANTS; SEMICONDUCTOR HETEROSTRUCTURES; TRANSISTOR.

Since January 1, 1990, all resistance measurements worldwide have been referred to the quantized Hall resistance standard, which is used to maintain the ohm in all national standards laboratories. Conventional wire-wound working standards are measured in terms of the quantized Hall resistance and then used to disseminate the ohm through the normal calibration chain. These working standards can be measured in terms of the quantized Hall resistance with a one-standard-deviation uncertainty of about 1 part in 10^8. *See* HALL EFFECT.

Resistance standards. Various resistance standards are in common use.

Wire-wound standard resistors. The day-to-day primary reference standards of national laboratories are 1-Ω resistors wound either from a quaternary alloy, which is a nickel-chrome alloy with small amounts of other metals added to produce better stability and reduced variation with temperature, or of manganin, which is a copper-nickel-manganese alloy. Both materials have temperature coefficients of zero at some temperature close to $68°\mathrm{F}$ ($20°\mathrm{C}$). The wire is best wound and located in a strain-free manner, and is immersed in oil to assist temperature uniformity. In use, the power dissipated is normally kept below a few milliwatts. The best resistors of this type have been found to change by less than 1 part in 10^7 per year. Basically similar construction is used for standard resistors at decade values from 1 milliohm to at least 100 kilohms, and at 25 Ω, a value at which a standard is needed as the basis for platinum resistance thermometry. These are all four-terminal resistors (**Fig. 4**); the ends of the resistive element are connected by pairs of leads, one of which is used to conduct the current, while the other carries only the much smaller current needed by the detector.

The resistive elements are often wound in a bifilar or hairpin fashion to reduce the self-inductance. All resistors necessarily have some inductance and capacitance, and can be simply represented by the

current terminal   standard resistance   current terminal
B I B
lead resistances   zero current (ideally)   lead resistances
A A
voltage terminals
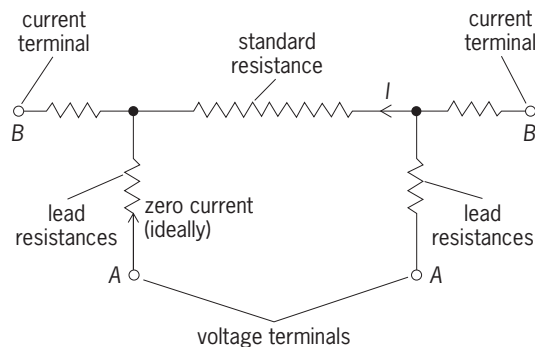
Fig. 4. Four-terminal resistor. Resistance $= V_{AA}/I$.
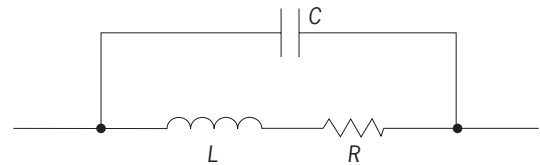
Fig. 5. Equivalent circuit of two-terminal resistor with resistance R. The capacitor represents the stray capacitance C of the resistor, and the inductor represents its stray inductance L.

circuit of **Fig. 5**. To first order, the impedance Z is given by Eq. (6), where C is the stray capacitance, L is

$$Z = R(1 - \omega^2 C^2 R^2) + j\omega(L - R^2 C) \qquad (6)$$

the stray inductance, and ω is the angular frequency. For resistors below about 300 Ω the inductive effect predominates, while resistors of more than this value are essentially capacitive. *See* ALTERNATING-CURRENT CIRCUIT THEORY.

Standard resistors below 100 kΩ in value are usually measured in thermostatic oil baths, permitting temperature stability to within a few millidegrees, and under these conditions it is possible to achieve repeatability of a few parts in 10^8. Higher values of resistance, above 1 megohm, cannot be oil-immersed because the conductance of the oil becomes significant. These are usually loosely wound on bobbins, and are two-terminal resistors. They may be fitted with a guard, which can be used to reduce the effects of leakage or stray conductance. Such resistors extend as far as 10^7 Ω. Beyond that level they become prohibitively large and expensive because of the quantity of wire involved.

Thick-film resistors. These resistors consist of elements of dimensions of only a fraction of an inch (a few millimeters), in which thick nichrome films are deposited on glass or ceramic substrates. These provide good stability with temperature by acting as compensated strain gages, in which the increase of resistivity due to change in temperature is compensated by the effect of strain produced by the differential expansion between the metal and the substrate. The chips are made and adjusted by printed-circuit techniques, and assemblies of very good overall performance can be obtained by selection of suitable elements. They have zero voltage coefficients, and overall temperature coefficients which may be as good as a few parts per million per degree Celsius. Their stability with time is also in the parts-per-million range. Because of their size, they have low inductance and capacitance, and so can be used over a wider range of frequency than wire-wound resistors. Resistors of similar form, printed onto a single substrate, are widely used in measuring instruments where accurate and constant ratios are required, since the maintenance of the ratio will normally be much better than the actual stability of the resistance.

Very high resistances. Above 10^9 Ω, the best available resistors are of protected thin-metal films, often completely encapsulated in glass. At these low conductances, any stray conductance due to moisture or dirt

becomes increasingly important. These resistances must therefore be kept clean and dry. Their resistance value is voltage-dependent to some extent, and so the voltage at which measurements are made must be specified.

Current-carrying resistors. These are needed for the calibration of resistive shunts which are used for measurement of direct current in the range from 10 to 10,000 A or higher. They are four-terminal resistors made from rods or strips of manganin, using several elements in parallel for the lower-resistance high-current values. They are mounted in oil-filled containers, fitted with some means of continuously stirring the oil. The range of resistance is from about 0.01 mΩ to 0.01 Ω, and in those intended for the larger currents the case is fitted with a water-cooled jacket. Because of the less well-defined thermal conditions, the accuracy obtainable is limited to 0.01% at the low current values and degrades to about 0.1% for high current values.

Decade resistors. These are switched assemblies of resistors, usually wire-wound, to provide variable resistances for laboratory work. They are necessarily two-terminal instruments, and hence have inbuilt errors due to lead resistances and variation of switch contacts. As in most switched decade resistance elements, each dial includes a 10 position in addition to the values from 0 to 9. This permits a ready self-calibration procedure to check linearity.

Kelvin-Varley dividers. These are used as variable voltage dividers for potentiometric work, for example, for the calibration of digital voltmeters from a fixed voltage standard. They consist (**Fig. 6**) of a series of sets of equal resistors connected by switches. In each set there are 11 resistors, 2 of which are spanned by the resistors of the next decade. This arrangement maintains an input resistance which is independent of the setting, and a low output resistance. *See* POTENTIOMETER.

Comparison methods. The value of an unknown resistance R_x is determined by comparison with a standard resistor R_s.

Wheatstone bridge. This bridge (**Fig. 7**) is perhaps the most basic and widely used resistance- or impedance-comparing device. Its principal advantage is that its operation and balance are independent of variations in the supply. At balance, Eq. (7) is satisfied. The

$$R_x = R_S \times \frac{R_1}{R_2} \qquad (7)$$

greatest sensitivity is obtained when all resistances are similar in value, and the comparison of standard resistors can then be made with a repeatability of about 3 parts in 10^8, the limit arising from thermal noise in the resistors. In use, the direction of supply is reversed periodically to eliminate effects of thermal or contact emf's.

The bridge is normally arranged for two-terminal measurements, and so is not suitable for the most accurate measurement at values below about 100 Ω, although still very convenient for lower resistances if the loss of accuracy does not matter. However, a
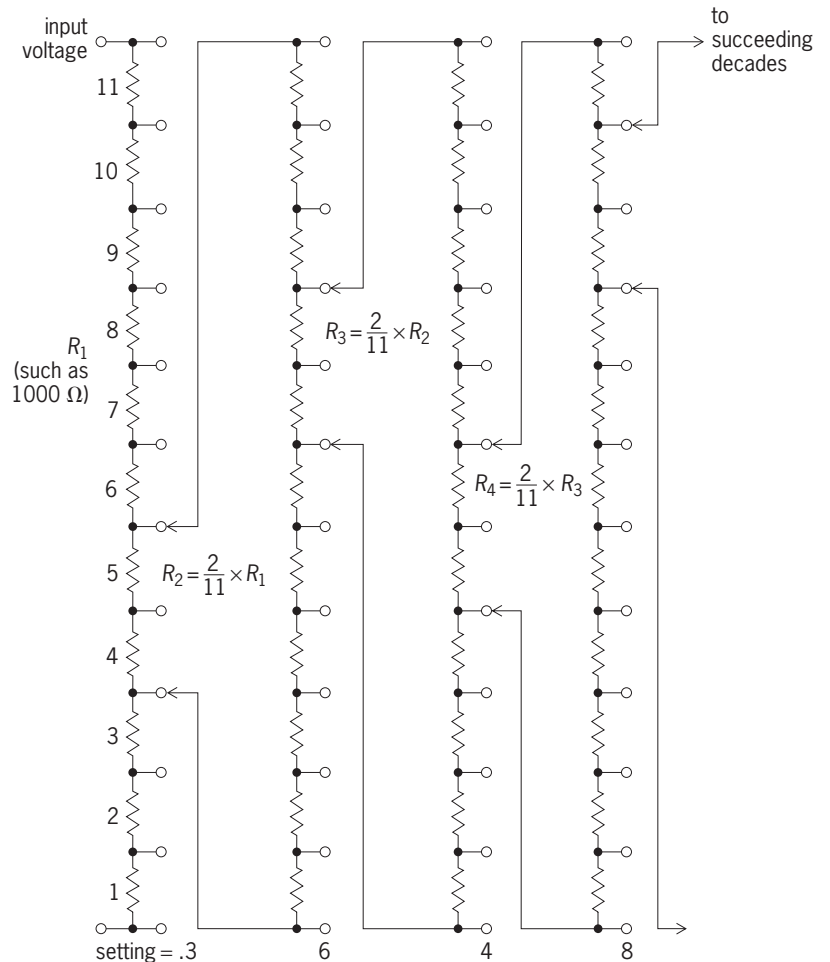
Fig. 6. Kelvin-Varley divider. Setting of divider with taps in the positions shown is 0.3648.

Wheatstone bridge has also been developed for the measurement of four-terminal resistors. This involves the use of auxiliary balances, and resistors of the same value can be compared with uncertainties of a few parts in 10^8.

Typically a bridge will have two decade-ratio arms, for example, of 1, 10, 100, 1000, and 10,000 Ω, and a variable switched decade arm of 1–100,000 Ω, although many variations are encountered. For the measurement of resistors of values close to the
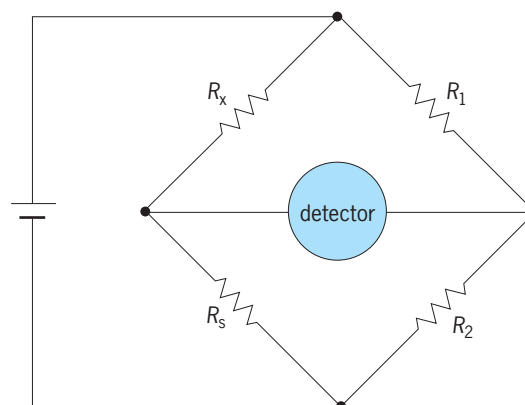
Fig. 7. Wheatstone bridge. At balance, $R_1/R_2 = R_x/R_s$.
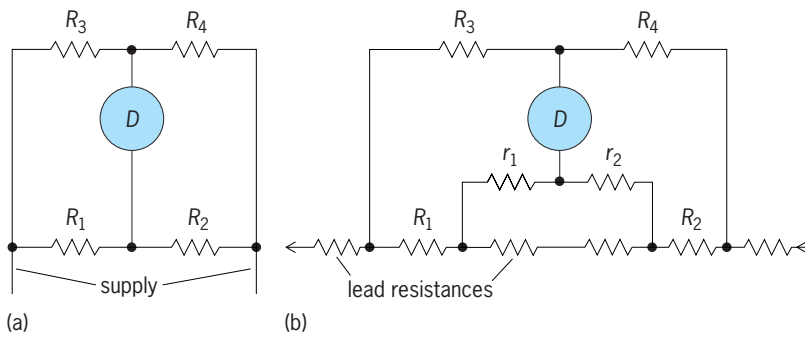
Fig. 8. Comparison of bridge circuits. (a) Wheatstone bridge. At balance, $R_1/R_2 = R_3/R_4$. (b) Kelvin double bridge. At balance, $R_1/R_2 = R_3/R_4 = r_1/r_2$.

decade values, a considerable increase in accuracy can be obtained by substitution measurement, in which the bridge is used only as an indicating instrument. The resistors being compared can be brought to the same value by connecting a much higher variable resistance across the larger of them, and the accuracy of this high-resistance shunt can be much less than that of the resistance being compared. *See* WHEATSTONE BRIDGE.

Kelvin double bridge. This is a double bridge (**Fig. 8**) for four-terminal measurements, and so can be used for very low resistances. The double condition for balance is given by Eq. (8). In addition to its use for ac-

$$\frac{R_3}{R_4} = \frac{R_1}{R_2} = \frac{r_1}{r_2} \qquad (8)$$

curate laboratory measurement of resistances below 100 Ω, it is very valuable for finding the resistance of conducting rods or bars, or for the calibration in the field of air-cooled resistors used for measurement of large currents. *See* KELVIN BRIDGE.

High-resistance measurements. Measurements of resistances from 10 megohms to 1 terohm (10^{12} Ω) or even higher with a Wheatstone bridge present additional problems. The resistance to be measured will usually be voltage-dependent, and so the measurement voltage must be specified. The resistors in the ratio arms must be sufficiently high in value that they are not overloaded. If a guard electrode is fitted, it is necessary to eliminate any current flowing to the guard from the measurement circuit. As an example, for measurement of 1 TΩ at 100 V, the ratio arms could be as shown in **Fig. 9**. The power dissipated in the 1-MΩ resistor is then 10 mW, and the bridge ratio is 10^6. The guard is connected to a subsidiary divider of the same ratio, so that any current flowing to it does not pass through the detector. Automated measurements can be made by replacing the ratio arms of the Wheatstone bridge by programmable voltage sources. An alternative method that can also be automated is to measure the *RC* time constant of the unknown resistor *R* combined with a capacitor of known value *C*. *See* INSULATION RESISTANCE TESTING.

Magnetic current comparator. This is a very effective instrument, which offers simple four-terminal measurements of resistors up to a value of 100 kΩ (or to 1 MΩ by modification). The essential feature of the

comparator (**Fig. 10**) is that the ratio of the currents in the two circuits is measured by means of balancing the fluxes they produce in a high-permeability toroidal comparator, a device like a transformer but connected so that the dc fluxes due to each winding are in opposition. The balance condition of zero flux is sensitively detected by a third winding. This balance condition is then maintained by a servo circuit controlling one of the power supplies, and is obtained when the ampere-turns products of the two windings are equal. The resistors to be compared are connected to the two circuits, and the turns' ratio adjusted by rotary switches until the emf's across the
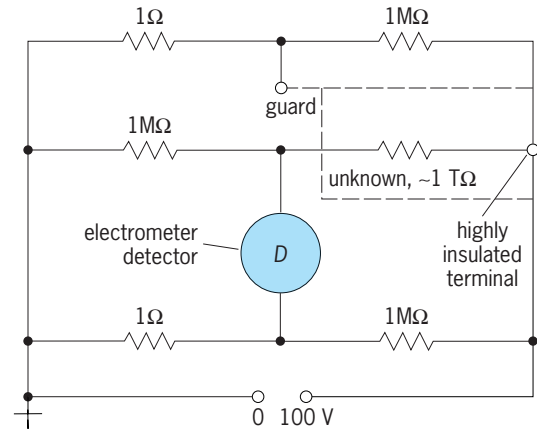
Fig. 9. Wheatstone bridge for high-resistance measurement.
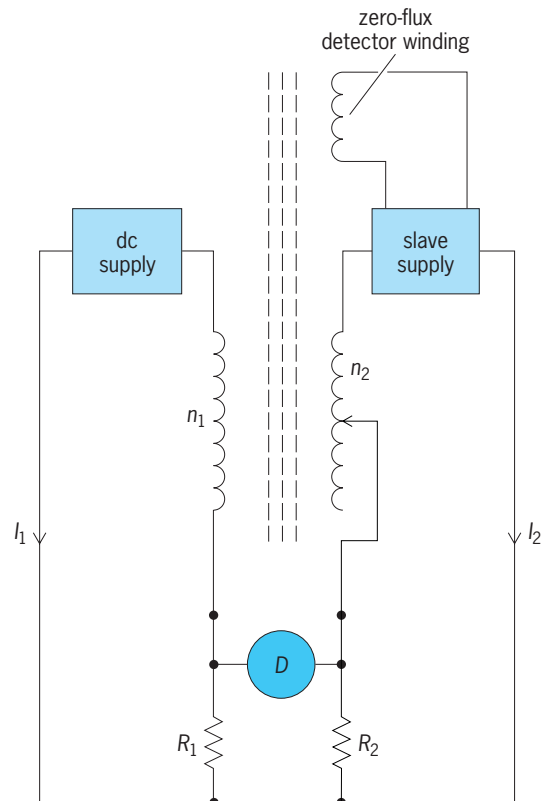
Fig. 10. DC current comparator bridge. $I_1/I_2 = n_2/n_1$; $R_1/R_2 = n_1/n_2$.

resistors balance. As with the Wheatstone bridge, a quick check can be made by reversing the positions of the two resistors being compared. An important merit of the system is that the ratio measured depends only on numbers of turns on the comparator, and not on ratio resistors, which will inevitably drift to some extent. *See* CURRENT COMPARATOR.

Cryogenic current comparator. This system operates in a similar way to the magnetic current comparator, except that the device that provides the accurate current ratio is a superconducting current comparator. The device relies on the Meissner effect in superconductivity. The principle can be described as follows, but the actual topology is more complex. A superconducting cylinder is threaded by two coils, one of one turn and one of N turns. The currents in the two coils are in inverse ratio to the number of turns. Any imbalance will give rise, via the Meissner effect, to current flowing in the surface of the superconducting cylinder. This current can be detected by another coil connected to a very sensitive superconducting quantum interference device (SQUID) detector, which acts as the feedback element in a servo system that maintains the two currents in the same precise ratio as the ratio of the turns in the two coils. The excellent shielding properties of superconductors make possible the provision of very accurate and stable ratios that can be measured to a few parts in 10^{10}. The small difference in the voltages across the two resistors being compared can then be measured by using a sensitive voltmeter. Resistance ratios can be measured with uncertainties approaching 1 part in 10^9. *See* MEISSNER EFFECT; SQUID; SUPERCONDUCTING DEVICES; SUPERCONDUCTIVITY.

Series-parallel resistance networks. The values of standard resistors at decade values are usually established from the 1-Ω level by means of buildup boxes. If n similar resistors are connected first in parallel and then in series, they provide the ratio n^2 between the series and parallel values. It can readily be shown that if each resistor is different from the mean by no more than some small amount δ (for example, 1 in 10^4), the ratio is achieved with an accuracy no worse than δ^2 (in this case, 1 in 10^8). This readily provides ratios of 100:1. A 10:1 ratio can be provided by connecting 10 resistors first in series, and then in three groups of 3 resistors in parallel connected in series with the remaining resistor, although this does not give as good accuracy as does the 100:1 arrangement. In order to connect four-terminal resistors so that only the resistive elements are truly in series, it is necessary to use special symmetrical four-terminal junctions, which are made so that the transresistance defined by the ratio of the voltage across any two connections to the current passing between the other two connections is less than 10^{-9} Ω. Resistance networks made in this way are called Hamon resistors, after their inventor B. V. Hamon. The nano-ohm junctions are made from solid copper, usually in cylindrical or cubic form, with the connecting wires brazed symmetrically into them.

Voltage and current measurement. An obvious and direct way of measuring resistance is by the simultaneous
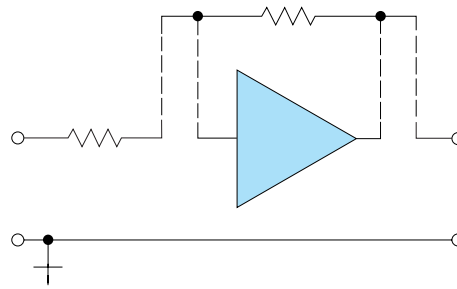
Fig. 11. Constant current resistance measurement circuit.

measurement of voltage and current, and this is usual in very many indicating ohmmeters and multirange meters. In most digital instruments, which are usually also digital voltage meters, the resistor is supplied from a constant-current circuit and the voltage across it is measured by the digital voltage meter. This is a convenient arrangement for a four-terminal measurement, so that long leads can be used from the instrument to the resistor without introducing errors (**Fig. 11**). The simplest systems, used in passive pointer instruments, measure directly the current through the meter which is adjusted to give full-scale deflection by an additional resistor in series with the battery. This gives a nonlinear scale of limited accuracy, but sufficient for many practical applications. *See* CURRENT MEASUREMENT; VOLTAGE MEASUREMENT.

AC resistance measurements. The use of inductive voltage dividers makes it possible to create simple ac resistance bridges in which the ratio arms of the bridge are composed of the two parts of the inductive voltage divider. Inductive voltage dividers are inherently very accurate and stable devices, and so do not need the extensive regular recalibration needed by resistive bridges. The simple bridge shown in **Fig. 12** is principally suitable for the comparison of resistors when their reactance (due to their stray inductance or capacitance) is negligible in comparison with the resistance, for example, at low frequencies. An important field of application is for platinum resistance thermometry, in which the resistance to be measured is usually in the range 1–1000 Ω. *See* INDUCTIVE VOLTAGE DIVIDER; THERMOMETER.

At high frequencies, two effects occur which
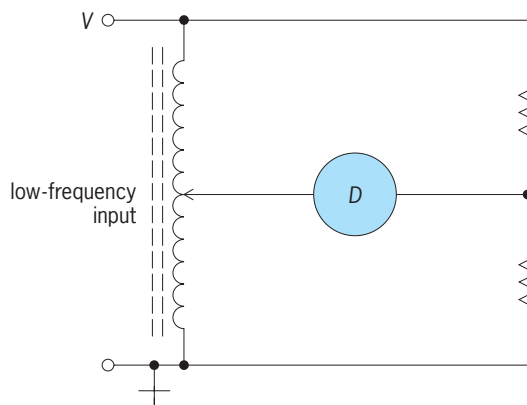
Fig. 12. Inductive divider resistance bridge.

change the measured value of resistance. The stray shunt capacitance C reduces the measured value of resistance by a factor $1/(1 + \omega^2 C^2 R^2)$ in addition to introducing resistance, and the resistance is increased due to the fact that the current flows only in the surface of the conductor. At radio and microwave frequencies, resistance becomes of importance principally in limiting the resolution or Q factor of circuits, and its measurement is often made by the measurement of this factor. *See* Q (ELECTRICITY).

Radio-frequency resistance is also used to evaluate the state of annealing of metals used in aircraft construction, and for this purpose it is measured by the coupled resistance induced in an inductor placed close to the metal. *See* EDDY CURRENT.

<div align="right">Cyril H. Dix; R. Gareth Jones</div>

Bibliography. D. A. Bell, *Electronic Instrumentation and Measurements*, 2d ed., 1997; C. H. Dix and A. E. Bailey, Electrical standards of measurement, *Proc. IEE*, vol. 122, no. 1OR, October 1975; F. K. Harris, *Electrical Measurements*, 1952, reprint 1975; A. Hartland, Quantum standards for electrical units, *Contemp. Phys.*, 29:477–498, 1988; L. Schnell (ed.), *Technology of Electrical Measurements*, 1993; L. M. Thompson, *Electrical Measurements and Calibration*, 2d ed., 1994.

Resistance welding

A process in which the heat for producing the weld is generated by the resistance to the flow of current through the parts to be joined. The application of external force is required; however, no fluxes, filler metals, or external heat sources are necessary. Most metals and their alloys can be successfully joined by resistance welding processes.

Methods. Several methods are classified as resistance welding processes: spot, roll-spot, seam, projection, upset, flash, and percussion (**Fig. 1**).

Resistance spot welding. In this process, coalescence at the faying surfaces is produced in one spot by the heat obtained from the resistance to electric current through the work parts held together under pressure by electrodes. The size and shape of the individually formed welds are limited primarily by the size and contour of the electrodes (**Fig. 2***a*). *See* SPOT WELDING.

Roll resistance spot welding. Separated resistance spot welds are made with one or more rotating circular electrodes. The rotation of the electrodes may or may not be stopped during the making of a weld.

Resistance seam welding. Coalescence at the faying surfaces is produced by the heat obtained from resistance to electric current through the work parts held together under pressure by electrodes. The resulting weld is a series of overlapping resistance spot welds made progressively along a joint by rotating the electrodes (Fig. 2*b*).

Projection welding. Coalescence is produced by the heat obtained from resistance to electric current through the work parts held together under pressure by electrodes. The resulting welds are localized at predetermined points by projections, embossments, or intersections (Fig. 2*c*).

Upset welding. Coalescence is produced, simultaneously over the entire area of abutting surfaces or progressively along a joint, by the heat obtained from resistance to electric current through the area of contact of those surfaces. Pressure is applied before heating is started and is maintained throughout the heating period.

Flash welding. Coalescence is produced simultaneously over the entire area of abutting surfaces by the heat obtained from resistance to electric current between the two surfaces and by the application of pressure after heating is substantially completed. Flash and upsetting are accompanied by expulsion of the metal from the joint. *See* FLASH WELDING.

Percussion welding. Coalescence is produced simultaneously over the entire abutting surfaces by the heat obtained from an arc produced by a rapid discharge of electrical energy with pressure percussively applied during or immediately following the electrical discharge.

Fundamentals of process. In all resistance welding processes, heat is generated at a concentrated position and is controlled by adjusting the current magnitude and the duration of current flow. The spot welding process will be used to illustrate the fundamentals involved.

The welding current from the secondary of the transformer flows through a flexible conductor, an electrode, and the workpieces, and back through an electrode and a flexible conductor to the transformer. The application of force to the electrodes is used to position the electrodes in intimate contact with the workpieces, thus completing the electrical circuit. The amount of heat generated during the current flow through the resistance of the parts may be expressed as $H = I^2Rt$, where H is the heat generated in the workpieces in joules (watt-seconds), I is the current in amperes, R is the resistance of the workpieces in ohms measured between the electrodes, and t is the time of current flow in seconds. In a typical spot welding application for joining two sheets of 0.04-in.-thick (1.0-mm) steel, the current is high (typically 10,000 A), the resistance is low (typically 0.0001 ohm), and the time is short (typically 0.15 s). In this case the heat generated in the parts, as calculated by the formula, is 1500 joules.

The values of current and time are selected such that sufficient base metal is melted at the desired position to produce a weld of the size and strength appropriate for the application.

Electrode force. The electrode force, which is used to provide intimate contact between the workpieces, reduces porosity and cracking in the fusion zone of the weld. However, when the electrode force is excessive, the weld region becomes severely indented by the electrodes.

Welding current. In the majority of cases, the current during resistance welding is in the range from 5000 to 50,000 A. Because of the high values of current required, resistance welding machines are designed

as low-voltage sources, usually in the range from 1 to 25 V. These low voltages are obtained from a step-down transformer which usually has a single- or two-turn, cast-copper, water-cooled secondary. Power supplied to the primary of the transformer is usually obtained from public utility single-phase, alternating-current sources at 440 V and 60 Hz. The step-down ratio of the transformer is approximately 100; thus the current requirements of the primary are lowered to reasonable values, which generally range from 50 to 500 A.

Power sources. Alternating current of 60 Hz is used in about 90% of the installations, although three-phase frequency converters are used to supply 3- to 25-Hz voltage to single-phase transformers. Welding current may also be supplied by a direct-current or stored-energy source. Direct current may be obtained from various low-voltage sources, such as rectifiers, homopolar generators, or storage batteries. The energy may be stored during a relatively long period and released suddenly from capacitors, magnetic fields, storage batteries, or heavy flywheels on homopolar generators. These types of power supplies eliminate large transient loads on power lines.

Welding time. The duration of the welding time is short, generally in the range from $\frac{1}{2}$ cycle ($\frac{1}{20}$ of a second) to a few minutes. In the great majority of applications, the time is in the range of 5–120 cycles of a 60-Hz source ($\frac{1}{12}$ to 2 s).

Welding electrodes. Resistance welding electrodes serve several purposes in the process: the application of the force required to bring the workpieces into alignment and into intimate contact; the production of proper contact resistance between electrodes and the workpieces; the conduction of welding current; the prevention of weld porosity and cracking; the prevention of "spitting," which is liquid-metal expulsion during welding; dissipation of heat developed at the electrode-workpiece interface; and dissipation of heat developed in the workpieces after the weld is made.

Electrodes are made of copper alloys which provide a satisfactory combination of mechanical properties and electrical and thermal conductivities. Electrodes generally are water-cooled to dissipate the heat developed, to minimize deformation of the electrode tips, and to prevent them from sticking to the workpieces.

Controls. The functions of these controls are to initiate and terminate the welding current, to control the magnitude of the current, and to sequence the operations. Controls are classified into two groups: synchronous and nonsynchronous.

In synchronous controls the timer initiates and terminates each half-cycle of the input current to the primary of the welding transformer at specific times with respect to the input voltage wave. Thus each half-cycle of welding current is identical, and the power delivered is the same for consecutive welds.

With nonsynchronous controls, the timer initiates the current to the welding transformer at random points with respect to the input voltage wave. In the worst case the nonsynchronous closure of the circuit
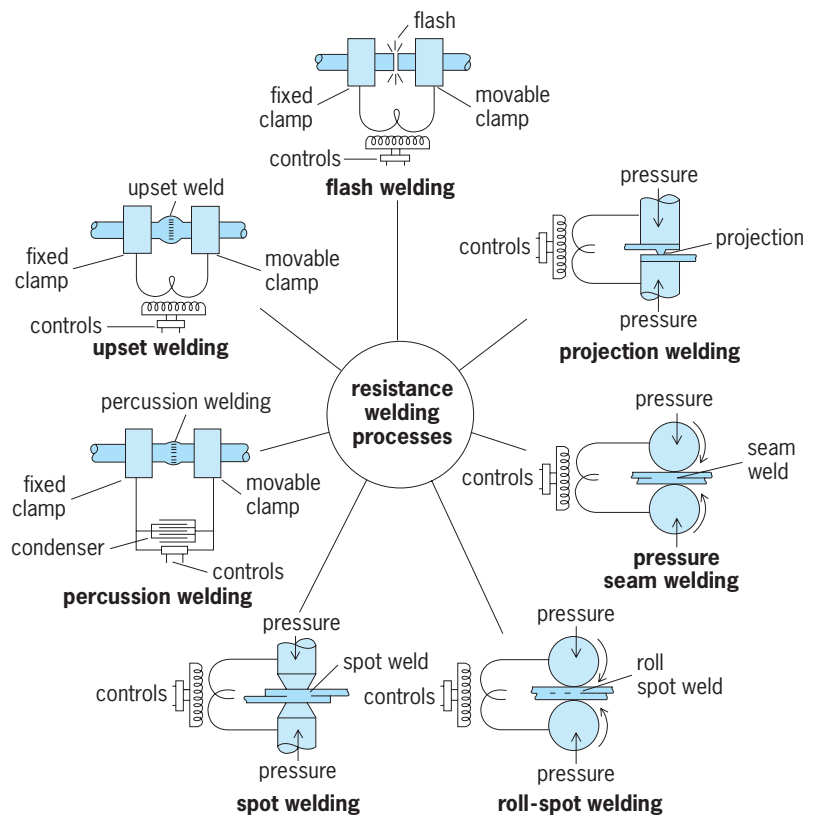
Fig. 1. Comparison of basic resistance welding processes. (*After B. E. Rossi, Welding Engineering, McGraw-Hill, 1954*)

causes a current transient with a duration of about 2 cycles and a magnitude of current twice that of the steady-stage value. The differences in current and power for nonsynchronous and synchronous controls are shown in **Fig. 3**. The variations in timing and current, which become less important as the time of welding increases, can be neglected for times
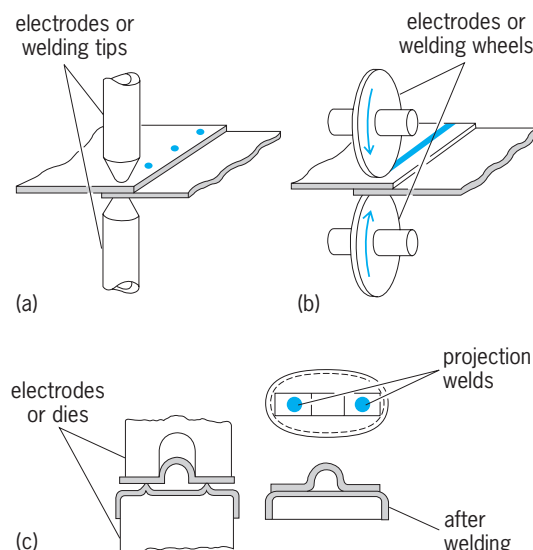
Fig. 2. Three typical resistance welding methods. (*a*) Spot welding. (*b*) Seam welding. (*c*) Projection welding. (*After W. H. Kearns, ed., Welding Handbook, 7th ed., vol. 3, American Welding Society, 1976*)
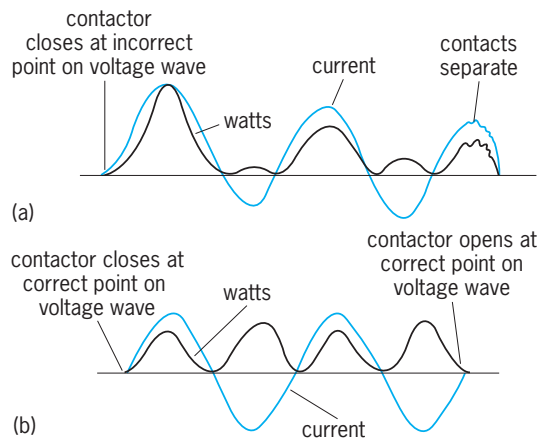
Fig. 3. Curves showing difference in current and power waves when circuit is closed nonsynchronously and synchronously. (*a*) Closure of circuit at incorrect point, with resultant decaying transient. The opening is also nonsynchronous. (*b*) Synchronous opening and closing of circuit at point of zero current corresponding to minimum transient. (*After A. L. Phillips, ed., Welding Handbook, 6th ed., American Welding Society, 1969*)

greater than about 20 cycles. However, when the welding time is less than 10 cycles, synchronous controls must be used for producing consistent welds.

The control of the magnitude of the welding current is accomplished, in coarse steps, by changing the taps on the welding transformer and, in a continuous manner, by adjustment of an electronic heat control. Synchronous controls are provided with electronic heat controls; however, these heat controls may be added to nonsynchronous controls. When a spot or projection weld is made by more than one impulse of current, the process is termed multiple-impulse welding. Multiple-impulse welding is also used to produce overlapping welds in the seam welding process.

Various levels of impulses can be used to preheat the metal, make the weld, refine the grain structure of the weld, and temper the hardened weld zone.

Machines. Welding machines are selected for use on the basis of the type of weld to be produced, the required weld quality, and production schedules. Simple, standard equipment is often utilized for production operations. However, because of economic considerations, complex resistance welding machines are used in the automotive and appliance industries, where high hourly production rates with minimum labor costs are essential.

Modern machines are designed with low-inertia systems which enable the electrodes to maintain the proper force on the workpieces at all times. If the electrodes fail to maintain adequate force during the welding operation, increased contact resistance will occur at the electrode-workpiece interface and at the workpiece interface. This increased resistance may cause excessive heating and result in deformed electrodes and expulsion of molten metal.

The clamping force of the electrodes is obtained by manual, mechanical, hydraulic, or pneumatic means. Prior to welding, electrodes should approach rapidly but in a controlled manner, so that they are not deformed by impacting the workpieces.

Surface preparation. Preweld surface preparation is usually required to control the magnitude and the variation in magnitude of the contact resistance at the electrode-workpiece interfaces and at the workpiece interface. Thus surface preparation, which removes dirt, oil, and oxide films by mechanical or chemical methods, provides for the production of consistent welds.

Metallurgical considerations. Most metals and alloys can be resistance-welded to themselves and to each other. The weld properties are determined by the metal and by the resultant alloys which form during the welding process. Stronger metals and alloys require higher electrode forces, and poor electrical conductors require less current. Copper, silver, and gold, which are excellent electrical conductors, are very difficult to weld because they require high current densities to compensate for their low resistance. Medium- and high-carbon steels, which are hardened and embrittled during the normal welding process, must be tempered by multiple impulses. *See* WELDING AND CUTTING OF MATERIALS.   Ernest F. Nippes

Bibliography. R. Carr and R. O'Con, *Welding Practices and Procedures*, 1983; A. C. Davies, *The Science and Practice of Welding*, 2 vols., 10th ed., 1993; W. H. Kearns (ed.), *Welding Handbook*, 7th ed., vol. 3, American Welding Society, 1980; Resistance Welders Manufacturers Association, *Resistance Welding Manual*, 4th ed., 1989.

Resistor

One of the three basic passive components of an electric circuit that displays a voltage drop across its terminals and produces heat when an electric current passes through it. The electrical resistance, measured in ohms, is equal to the ratio of the voltage drop across the resistor terminals measured in volts divided by the current measured in amperes. *See* OHM'S LAW.

Resistors are described by stating their total resistance in ohms along with their safe power-dissipating ability in watts. The tolerance and temperature coefficient of the resistance value may also be given. *See* ELECTRICAL RESISTANCE; ELECTRICAL RESISTIVITY.

All resistors possess a finite shunt capacitance across their terminals, leading to a reduced impedance at high frequencies. Resistors also possess inductance, the magnitude of which depends greatly on the construction and is largest for wire-wound types. *See* CAPACITANCE; ELECTRICAL IMPEDANCE; INDUCTANCE.

Classification by use. Resistors may be classified according to the general field of engineering in which they are used.

Power resistors. Such resistors range in size from about 5 W to many kilowatts and may be cooled by air convection, air blast, or water. The smaller sizes, up to several hundred watts, are used in both the power and electronics fields of engineering.

Resistor color code

Color	Figures	Multiplier	Tolerance, %	Temperature coefficient, ppm/°C
No band	—	—	20	—
Silver	—	0.01	10	—
Gold	—	0.1	5	—
Black	0	1	—	200
Brown	1	10	1	100
Red	2	10^2	2	50
Orange	3	10^3	—	15[†]
Yellow	4	10^4	*	25[†]
Green	5	10^5	0.5	—
Blue	6	10^6	0.25	10
Violet	7	10^7	0.1	5
Gray	8	10^8	—	1
White	9	10^9	—	—

*Yellow sometimes used instead of gold to indicate 5% tolerance.
[†] Orange or yellow sixth bands may indicate reliability ratings when used on a 5% tolerance device.

Instrument resistors. Direct-current (dc) ammeters employ resistors as meter shunts to bypass the major portion of the current around the low-current elements. These high-accuracy, four-terminal resistors are commonly designed to provide a voltage drop of 50–100 mV when a stated current passes through the shunt. *See* AMMETER.

Voltmeters of both the dc and the ac types employ scale-multiplying resistors designed for accuracy and stability. The arc-over voltage rating of these resistors is of importance in the case of high-voltage voltmeters. *See* VOLTMETER.

Standard resistors. These devices are used for calibration purposes in resistance measurements and are made to be as stable as possible, in value, with time, temperature, and other influences. Resistors with values from 1 ohm to 10 megohms are wound by using wire made from special alloys. The best performance is obtained from quaternary alloys, which contain four metals. The proportions are chosen to give a shallow parabolic variation of resistance with temperature, with a peak, and therefore the slowest rate of change, near room temperature. *See* ELECTRICAL UNITS AND STANDARDS.

Resistors for electronic circuits. By far the greatest number of resistors manufactured are intended for use in the electronics field. The major application of these resistors is in transistor analog and digital circuits which operate at voltage levels between 0.1 and 200 V, currents between 1 μA and 100 mA, and frequencies from dc to 100 MHz. Their power-dissipating ability is small, as is their physical size.

Preferred values. Since their exact value is rarely important, resistors are supplied in decade values (0.1, 1, 10, 100 ohms, and so forth) with the interval between these divided into a geometric series, thus having a constant percentage increase. For noncritical applications, values from a series with intervals of 20% (12 per decade) are appropriate. A series with 10% intervals (24 per decade) is often used for resistors having a tolerance of 1%. Where the precise value of a resistor is important, a series with 2.5% intervals (96 per decade) may be used.

Resistor marking. Most electronic-type resistors are marked by using a standard code consisting of a set of closely spaced colored bands which start near one end of the resistor. The code evolved from a three-digit system in which the first two bands define the two significant figures of the value, and the third band indicates the number by which the significant figures are to be multiplied in order to obtain the value, in ohms (see **table** and **illus.**). A tolerance of ±20% is assumed unless a fourth band is applied, spaced more widely from the previous bands. The code has been extended to allow three significant figures and a multiplier to be displayed in four equally spaced bands. These are always followed, after a wider gap, by a tolerance band, and there may be a sixth band to indicate the temperature coefficient.

Resistors may also be marked by using a figure code. The symbols Ω, K, and M can be used both to indicate the units in which the value is defined and to effectively serve as a decimal point. Thus 2Ω7 and 18K7 signify 2.7 ohms and 18,700 ohms. Decimal digits are sometimes used in the same way as colored

(a)   (b)   (c)
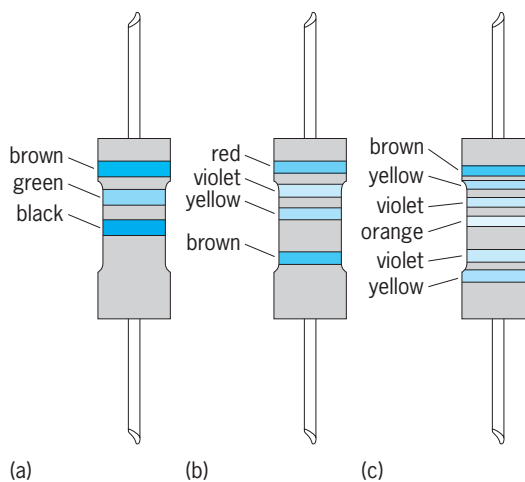
Examples of color-coded resistors. (*a*) 15 Ω ± 20%. (*b*) 270,000 Ω ± 1%. (*c*) 147,000 Ω ± 0.1%, 25 ppm/°C.

bands, so that the final digit is to be interpreted as a multiplier; for example, a 392-kΩ resistor may be marked 3923.

Classification by construction. Resistors are also classified according to their construction, which may be composition, film-type, wire-wound, or integrated circuit.

Composition resistor. This resistor is in wide use because of its low cost, high reliability, and small size. Basically it is a mixture of resistive materials, usually carbon, and a suitable binder molded into a cylinder. Copper wire leads are attached to the ends of the cylinder, and the entire resistor is molded into a plastic or ceramic jacket. The overall length of the jacketed resistor excluding the leads is $^1/_2$–$1^1/_3$ in. (13–34 mm) for resistors varying in power rating from 0.25 to 2 W. Composition resistors are commonly used in the range from several ohms to 10–20 MΩ, and are available with tolerances of 20, 10, or 5%.

Film-type resistor. This resistor is now the preferred type for most electronic applications because its performance has surpassed that of composition resistors and mass-production techniques have reduced the cost to a comparable level. The design of the film resistor further lends itself to the controlled manufacture of precision resistors of any desired value. Basically this resistor consists of a thin conducting film of carbon, metal, or metal oxide deposited on a cylindrical ceramic or glass former. The resistance is controlled by cutting a helical groove through the conducting film. This helical groove increases the length and decreases the width of the conducting path, thereby determining its ohmic value. By controlling the conductivity, thickness of the film, and pitch of the helix, resistors over a wide range of values can be manufactured. Since the resistor value may be measured while cutting is in progress, close manufacturing tolerances are possible. Standard tolerances of ±0.1% and temperature coefficients of resistance (TCR) better than 15 parts per million (ppm)/°C are readily available. Temperature coefficients as low as 1 ppm/°C are available at higher cost.

Film construction is used for very high value resistors, up to and even beyond 1 TΩ (10^{12} ohms). To avoid the problems caused by humidity, such resistors require special packaging in silicone-treated glass envelopes. A guard electrode is sometimes applied, to both the inner and outer glass surfaces. High-value resistors should never be touched by bare hands for their value will be affected by contamination.

The spiral-cut resistor displays a small inductive effect, which may be a problem at high frequencies. In order to obtain the widest possible bandwidth, film resistors are constructed by producing a serpentine path of resistance material on a flat substrate by using photographic methods, followed by mechanical or laser trimming. Such resistors can closely approach the performance of the best wire-wound types.

Wire-wound resistor. Wire remains the most stable form of resistance material available; therefore, all high-precision instruments rely upon wire-wound resistors. Wire also will tolerate operation at high temperatures, and so compact high-power resistors use this construction. Power resistors are available in resistance values from a fraction of an ohm to several hundred thousand ohms, at power ratings from one to several thousand watts, and at tolerances from 10 to 0.1%. The usual design of a power resistor is a helical winding of wire on a cylindrical ceramic former. After winding, the entire resistor is coated in vitreous enamel. Alternatively, the wound element may be fitted inside a ceramic or metal package, which will assist in heat dissipation. The helical winding results in the resistor having significant inductance, which may become objectionable at the higher audio frequencies and all radio frequencies.

Precision wire-wound resistors are usually wound in several sections on ceramic or plastic bobbins and are available in the range from 0.1 Ω to 10 MΩ. They can be supplied with a tolerance as small as 25 ppm and temperature coefficients less than 1 ppm/°C. Special selection can provide groups of components with even more closely matched temperature coefficients and values or ratios. This form of construction gives large and unpredictable inductive and capacitive effects, and so the use of devices of this type is limited to dc and frequencies below about 50 kHz. For ac standards applications, special constructional methods are used to control the reactive properties, for example, by winding the wire on a flat card, thereby reducing the cross-sectional area of the coil.

Integrated circuit resistor. These resistors must be capable of fabrication on a silicon integrated circuit chip along with transistors and capacitors. There are two major types: thin-film resistors and diffused resistors.

Thin-film resistors are formed by vacuum deposition or sputtering of nichrome, tantalum, or Cermet (Cr-SiO). Such resistors are stable, and the resistance may be adjusted to close tolerances by trimming the film by using a laser. Typical resistor values lie in the range from 100 Ω to 10 kΩ with a matching tolerance of ±0.2% and a temperature coefficient of resistance of ±10 to ±200 ppm/°C.

Diffused resistors are based upon the same fabrication geometry and techniques used to produce the active transistors on the silicon chip or die. A diffused base, emitter, or epitaxial layer may be formed as a bar with contacts at its extremities. The resistance of such a semiconductor resistor depends upon the impurity doping and the length and cross section of the resistor region. In the case of the base-diffused resistor, the emitter and collector regions may be formed so as to pinch the base region to a very small cross-sectional area, thereby appreciably increasing the resistance. The relatively large impurity carrier concentration in *n*- and *p*-type regions limits the resistance value. Resistor values between 100 Ω and 10 kΩ are common with a matching tolerance lower than ±2% but with a temperature coefficient of resistance approaching 2000 ppm/°C. The epitaxial layer resistor has a much lower impurity carrier concentration and offers resistance values up to 50 kΩ but with a higher temperature coefficient of resistance

of 3000 ppm/°C and a matching tolerance of about ±.6%. *See* INTEGRATED CIRCUITS.

Adjustable resistors. The deposited-film and wire-wound resistors lend themselves to the design of adjustable resistors or rheostats and potentiometers. Adjustable-slider power resistors are constructed in the same manner as any wire-wound resistor on a cylindrical form except that when the vitreous outer coating is applied an uncovered strip is provided. The resistance wire is exposed along this strip, and a suitable slider contact can be used to adjust the overall resistance, or the slider can be used as the tap on a potentiometer. *See* POTENTIOMETER; RHEOSTAT.

Where continuous adjustment of the resistor is intended, a ring-shaped form is generally used. For power resistors, the ring is wound with resistance wire. For compact 0.5- and 1-W resistors, the ring is coated on one surface with a resistance film. Adjustable resistors have the problem of maintaining a good, noise-free, electrical contact at the wiper, which is mounted on a shaft concentric with the ring.

For discussion of nonlinear resistors *see* THERMISTOR; VARISTOR.   R. B. D. Knight

Bibliography. A. B. Glaser and G. E. Subak-Sharpe, *Integrated Circuit Engineering*, 1977; P. R. Gray and R. G. Meyer, *Analysis and Design of Integrated Circuits*, 4th ed., 2001; F. K. Harris, *Electrical Measurements*, 1952, reprint 1975; P. Horowitz and W. Hill, *The Art of Electronics*, 2d ed., 1989; F. Mazda (ed.), *Electronics Engineer's Reference Book*, 7th ed., 1993; M. G. Say (ed.), *Electrical Engineer's Reference Book*, 15th ed., 1993.

Resolving power (optics)

A quantitative measure of the ability of an optical instrument to produce separable images. The images to be resolved may differ in position because they represent (1) different points on the object, as in telescopes and microscopes, or (2) images of the same object in light of two different wavelengths, as in prism and grating spectroscopes. For the former class of instruments, the resolving limit is usually quoted as the smallest angular or linear separation of two object points, and for the latter class, as the smallest difference in wavelength or wave number that will produce separate images. Since these quantities are inversely proportional to the power of the instrument to resolve, the term resolving power has generally fallen into disfavor. It is still commonly applied to spectroscopes, however, for which the term chromatic resolving power is used, signifying the ratio of the wavelength itself to the smallest wavelength interval resolved. The figure quoted as the resolving power or resolving limit of an instrument may be the theoretical value that would be obtained if all optical parts were perfect, or it may be the actual value found experimentally. Aberrations of lenses or defects in the ruling of gratings usually cause the actual resolution to fall below the theoretical value, which therefore represents the maximum that could

be obtained with the given dimensions of the instrument in question. This maximum is fixed by the wave nature of light and may be calculated for given conditions by diffraction theory. *See* DIFFRACTION; OPTICAL IMAGE.

Chromatic resolving power. The chromatic resolving power R of any spectroscopic instrument, including prisms, gratings, and interferometers, is defined by Eq. (1), where $\delta\lambda$ represents the difference

$$R = \frac{\lambda}{\delta\lambda} \qquad (1)$$

in wavelength of two equally strong spectrum lines that can barely be separated by the instrument, and λ the average wavelength of these two lines. It is necessary to specify more precisely the term "barely separated," and for prisms and gratings, in which the width of the lines is determined by diffraction, this is done by use of Rayleigh's criterion. **Illustration** *a* shows the contours of two similar spectrum lines which are at the limit of resolution according to this criterion. The lighter curve represents the line shape due to Fraunhofer diffraction for the wavelength λ, the dashed curve that for $\lambda + \delta\lambda$, and the heavy curve the sum of the two. Rayleigh's criterion specifies that the lines are resolved when the principal maximum of one falls exactly on the first minimum (zero intensity) of the other. Diffraction theory shows that the intensity I of either pattern at the central crossing point is $4/\pi^2$ of that at the maximum, so that the curve representing the sum dips to 81% at the center. The theory also shows that the angular separation $\delta\theta$ of the rays forming the two maxima is λ/a, where a is the linear width of the

(a)

(b)
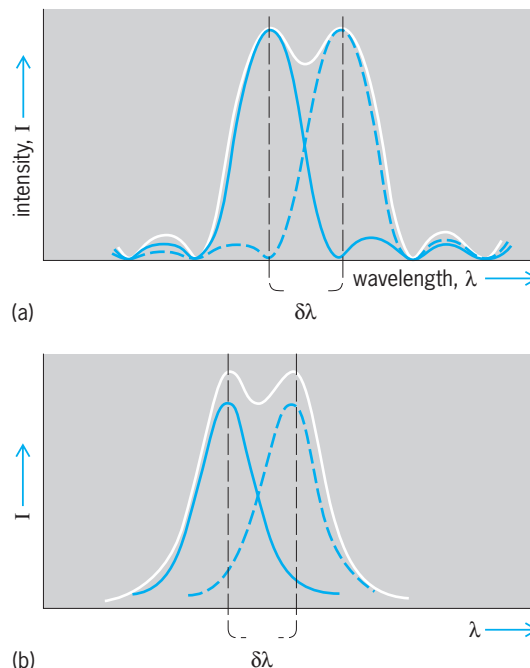
Resolution of two spectrum lines (*a*) when the shape is determined by diffraction (Rayleigh criterion), and (*b*) when the shape follows the Airy formula. The latter is applicable to multiple-beam interferometers.

beam of light emerging from the prism or grating. Hence, quite generally for such an instrument, the resolving power may be defined by Eq. (2*a*), which can be expressed in words as Eq. (2*b*). In a given

$$R = \frac{\lambda}{\delta\lambda} = \frac{\lambda}{\delta\theta}\frac{d\theta}{d\lambda} = a\,\frac{d\theta}{d\lambda} \qquad (2a)$$

Chromatic resolving power

= width of emergent beam

× angular dispersion   (2*b*)

instrument, the calculation of resolving power thus involves finding the last two quantities.

Resolving power of prisms. When a prism is used at minimum deviation, the resolving power depends on the length *b* of the base of the prism and the slope $dn/d\lambda$ of the dispersion curve giving the wavelength variation of the refractive index *n*. Thus Eq. (3) holds.

$$R = b\frac{dn}{d\lambda} \qquad (3)$$

Here the assumption is made that the prism is completely filled by the beam of light. If it is not, *b* must represent the difference in path length between the longest and shortest rays through the prism. *See* OPTICAL PRISM.

Resolving power of gratings. This equals the product of the order of interference *m* and the total number of rulings *N*. The order *m* may be expressed in terms of the grating space *s* and the angles α and β of incidence and diffraction. Thus Eq. (4) holds.

$$R = mN = \frac{Ns(\sin\alpha + \sin\beta)}{\lambda}$$
$$= \frac{w(\sin\alpha + \sin\beta)}{\lambda} \qquad (4)$$

Here *w* is the width of the ruled area of the grating. For the limiting case of grazing angles of incidence and diffraction, the maximum possible *R* is seen to be $2w/\delta$, or the number of wavelengths in twice the width of the grating. *See* DIFFRACTION GRATING.

Resolving power of interferometers. For the type of interferometer most commonly used, the Fabry-Perot interferometer, the resolving power may be expressed as the product of the order of interference $m = 2t/\lambda$, where *t* is the separation of the interferometer mirrors, and an effective number N_{eff} of interfering beams. For interferometers the line contour of the spectrum lines is not that of Fraunhofer diffraction, but is given by a relation called the Airy formula. This contour has no points of zero intensity but has the general shape shown in illus. *b*. Therefore, the Rayleigh criterion cannot be applied in the usual way. If, however, the two curves are made to cross at the half-intensity point of each, it is found that there is a dip of approximately 20% in the resultant curve. The value of N_{eff} is thereby specified, and the resolving power *R* is given by Eq. (5), where

$$R = mN_{eff} = m\left(\frac{\pi\sqrt{\rho}}{1-\rho}\right) \qquad (5)$$

ρ designates the reflectance of the interferometer plates. *See* INTERFEROMETRY.

Resolving power of telescopes. This depends on the size of the diffraction maximum produced when light from a distant point source passes through a circular aperture of size equal to that of the objective lens or mirror. A graph of the intensity in the diffraction pattern plotted against radial distance closely resembles one of the curves of illus. *a*, and hence the pattern consists of a central spot surrounded by faint rings. The angular radius of the first dark ring corresponds, by the Rayleigh criterion, to the angular separation of two point sources that are barely resolved. Theory gives this angle, which represents the resolving limit, as defined by Eq. (6) for

$$\alpha = \frac{1.220\lambda}{d}\text{ radians } = \frac{14.1}{d}\text{ seconds of arc}\quad (6)$$

$\lambda = 560$ nanometers and *d*, the diameter of the objective lens, in centimeters. *See* TELESCOPE.

Resolving power of microscopes. This is determined by diffraction of a circular aperture representing the exit pupil of the microscope objective. There are two important differences between the resolving power of microscopes and that of telescopes. First, the resolving limit of microscopes is expressed in terms of the smallest distance *l* between two points on the object that are just resolved. Second, this limit depends on the mode of illumination of the object. If the illumination is incoherent, so that there is no constant phase relation between light from adjacent points, the resolving limit is given by Eq. (7), where

$$l = \frac{0.61\lambda}{n\,\sin\alpha} \qquad (7)$$

n is the refractive index of the material (for example, oil) in the object space, and α is the angle that the extreme ray entering the objective makes with the axis of the instrument. The quantity $n\sin\alpha$ is called the numerical aperture of the objective. With coherent illumination the resolving limit is given by this formula, with 1.0 in place of 0.61, provided the illumination is central. When the object is illuminated from a point slightly to one side, the factor may be reduced to 0.5. *See* OPTICAL MICROSCOPE; SPECTROSCOPY.

<div style="text-align:right">Francis A. Jenkins; George R. Harrison</div>

Bibliography. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed., 1976; Optical Society of America, *Handbook of Optics*, 2 vols., 2d ed., 1995; W. J. Smith, *Modern Optical Engineering*, 3d ed., 2000.

Resonance (acoustics and mechanics)

When a mechanical or acoustical system is acted upon by an external periodic driving force whose frequency equals a natural free oscillation frequency of the system, the amplitude of oscillation becomes large and the system is said to be in a state of resonance.

When a simple oscillator of mass m, stiffness constant s, and mechanical damping constant R is driven by a periodic driving force $F \cos 2\pi ft$, it vibrates with a velocity amplitude given by the equation below, which implies that (1) the amplitude be-

$$V = \frac{F}{[R^2 + (2\pi fm - s/2\pi f)^2]^{1/2}}$$

comes a maximum when the driving frequency is $f = (1/2\pi) \cdot \sqrt{s/m}$, that is, at the natural free oscillation frequency of the oscillator, (2) small damping constants R are associated with large amplitudes of vibration at resonance, and (3) the smaller R, the more rapidly the amplitude decreases as the driving frequency departs from the resonance frequency. In addition, driving any vibrating system at its resonance frequency is characterized by a maximum dissipation of power.

A knowledge of both the resonance frequency and the sharpness of resonance is essential to any discussion of driven vibrating systems. When a vibrating system is sharply resonant, careful tuning is required to obtain the resonance condition. Mechanical standards of frequency must be sharply resonant so that their peak response can easily be determined. In other circumstances, resonance is undesirable. For example, in the faithful recording and reproduction of musical sounds, it is necessary either to have all vibrational resonances of the system outside the band of frequencies being reproduced or to employ heavily damped systems. *See* ACOUSTIC RESONATOR; OSCILLATOR; SYMPATHETIC VIBRATION; VIBRATION.   Lawrence E. Kinsler

Resonance (alternating-current circuits)

A condition in a circuit characterized by relatively unimpeded oscillation of energy from a potential to a kinetic form. In an electrical network there is oscillation between the potential energy of current in inductance. This is analogous to the mechanical resonance seen in a pendulum.

Three kinds of resonant frequency in circuits are officially defined. Phase resonance is the frequency at which the phase angle between sinusoidal current entering a circuit and sinusoidal voltage applied to the terminals of the circuit is zero. Amplitude resonance is the frequency at which a given sinusoidal excitation (voltage of current) produces the maximum oscillation of electric charge in the resonant circuit. Natural resonance is the natural frequency of oscillation of the resonant circuit in the absence of any forcing excitation. These three frequencies are so nearly equal in low-loss circuits that they do not often have to be distinguished.

Phase resonance is perhaps the most useful in many practical situations, as well as being slightly simpler mathematically. The following discussion considers phase resonance in passive, linear, two-terminal networks.
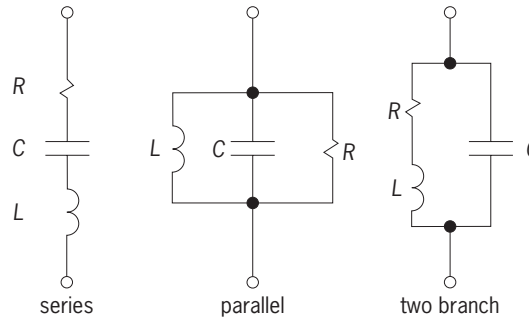
Fig. 1. **Resonant circuits.**

Resonance can appear in two-terminal networks of any degree of complication, but the three circuits shown in **Fig. 1** are simple and typical. The first illustrates series resonance and the second, parallel resonance; the third is a series-parallel resonant circuit of two branches (sometimes referred to as antiresonance). Series resonance is highly practical for providing low impedance at the resonant frequency. Parallel resonance is the dual of series resonance, but it is not practical because it assumes an inductive element with no resistance. The third example, however, shows an eminently practical means of providing the typical characteristic of parallel resonance, which is high impedance at the resonant frequency.

Use. Resonance is of great importance in communications, permitting certain frequencies to be passed and others to be rejected. Thus a pair of telephone wires can carry many messages at the same time, each modulating a different carrier frequency, and each being separated from the others at the receiving end of the line by an appropriate arrangement of resonant filters. A radio or television receiver uses much the same principle to accept a desired signal and to reject all the undesired signals that arrive concurrently at its antenna; tuning a receiver means adjusting a circuit to be resonant at a desired frequency.

Many frequency-sensitive circuits are not truly resonant, and oscillations of a certain frequency can be produced or enhanced by networks that do not involve inductance. It is difficult and expensive to provide inductance with integrated circuits, but frequency selection can be provided by the use of capacitance and resistance, a large amount of amplification being obtained from the semiconductor material employed.
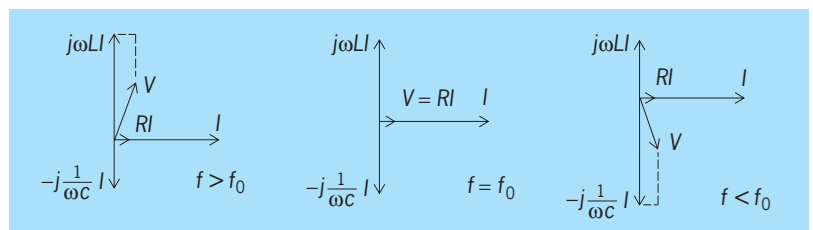
Fig. 2. **Phasor diagrams at frequencies near resonance ($Q = 5$).**

Series resonance. **Figure 2** shows phasor diagrams of the voltage, resulting from a given current (steady alternating current) in a series-resonant circuit, such as shown in Fig. 1. The component voltages across the three circuit elements add to give the total applied voltage V, as shown for a frequency slightly above resonance, for the resonant frequency, and for the frequency below resonance. It is of course possible in low-loss (high Q) circuits for the voltage across the capacitance and the voltage across the inductance each to be many times greater than the applied voltage.

Analytically, the impedance of the series-resonant circuit is given by Eq. (1). The resonant frequency

$$Z = R + j\omega L + \frac{1}{j\omega C} = R + j\left(\omega L - \frac{1}{\omega C}\right) \quad (1)$$

f_0 is the frequency at which Z is purely real (phase resonance), so $\omega_0 L = 1/\omega_0 C$, or $2\pi f_0 L = 1/2\pi f_0 C$, from which Eq. (2) obtains. A more convenient notation is expressed by Eq. (3). In Eqs. (1)–(3), $Z =$

$$f_0 = \frac{1}{2\pi \sqrt{LC}} \quad (2)$$

$$Z = R_0 \left(\frac{R}{R_0} + jQ_0\delta \frac{2+\delta}{1+\delta}\right) \quad (3)$$

impedance at the terminals of the series-resonant circuit; R, L, C = the three circuit parameters; R_0 = resistance (effective) at resonant frequency; $Q_0 = \omega_0 L/R_0$; $\delta = (\omega - \omega_0)/\omega_0$; $\omega = 2\pi f$, where f is frequency (hertz); $\omega_0 = 2\pi f_0$, where f_0 is resonant frequency.

Equation (3) is true for all series-resonant circuits, but interest is mainly in circuits for which Q_0, the quality factor at the resonant frequency, is high (20 or more) and for which δ, the fractional detuning, is low (perhaps less than 0.1). Assuming high Q_0 and low δ, which means a low-loss circuit and a frequency near resonance, Eq. (4) is very nearly the relative

$$\frac{Y}{Y_0} = \frac{Z_0}{Z} = \frac{1}{1 + j2Q_0\delta} \quad (4)$$

admittance of the series-resonant circuit.

Universal resonance curve. The magnitude and the real and imaginary components of Eq. (4) are usefully plotted in the universal resonance curve of **Fig. 3**. Since Y/Y_0 is plotted as a function of $Q_0\delta$, this curve can be applied to all series-resonant circuits. (If $Q_0 = 20$, the error in Y barely exceeds 1% of Y_0 for any δ, and is less for small δ.)

Moreover, because of the duality of the network, the curve can also be applied to any parallel-resonant circuit (Fig. 1) provided Q_0 is now interpreted as $Q_0 = R_0/\omega_0 L$. When used for a parallel-resonant circuit, the curve of Fig. 3 gives not Y/Y_0 but the relative input impedance Z/Z_0.

Finally, the universal resonance curve of Fig. 3 can also be applied (with the same slight approximations) to the two-branch resonant circuit of Fig. 1. For this purpose the curve shows Z/Z_0 (as for the
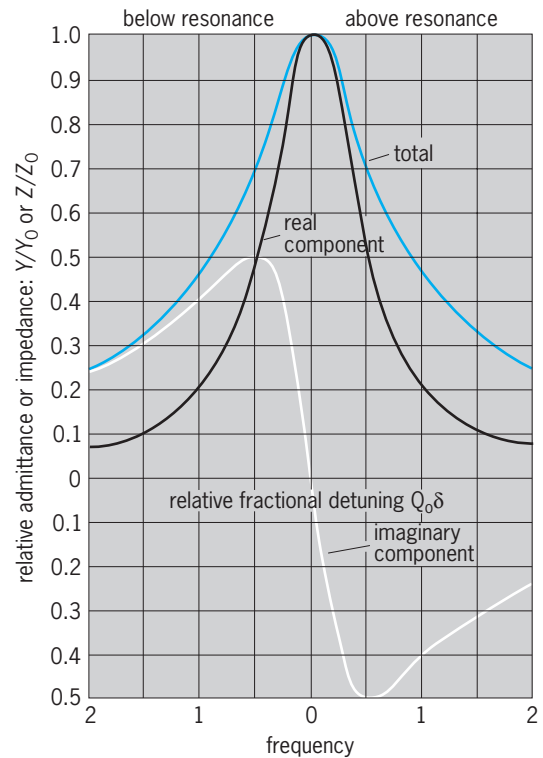
Fig. 3. Universal resonance curve. (*After H. H. Skilling, Electrical Engineering Circuits, 2d ed., John Wiley and Sons, 1955*)
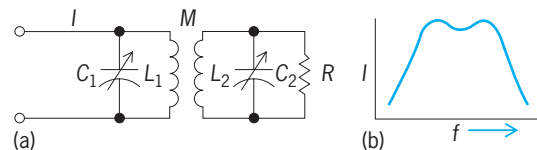
Fig. 4. Multiple resonance. (*a*) Double-tuned network with mutual inductance *M*. (*b*) Current *I* in *R* as function of frequency *f*.

three-branch parallel-resonant circuit), but the value of Q to be used is $Q_0 = \omega_0/R_0$, exactly as with the series-resonant circuit. Note that Z_0 for this circuit is given by Eq. (5) [instead of being equal to R_0 as it is

$$Z_0 = (\omega_0 L)Q_0 = R_0 Q_0^2 \quad (5)$$

in the other two circuits of Fig. 1].

Multiple resonance. If two or more coupled circuits are resonant at slightly different frequencies, many valuable characteristics can be obtained. **Figure 4** shows a double-tuned network and a typical curve of current in R, the load, as a function of frequency. *See* ALTERNATING-CURRENT CIRCUIT THEORY; NETWORK THEORY.   Hugh Hildreth Skilling

Bibliography. N. Balabanian, *Electric Circuits*, 1994; J. R. Duff and S. L. Herman, *Alternating Current Fundamentals*, 6th ed., 1999; Institute of Electrical and Electronics Engineers, *IEEE Standard Dictionary of Electrical and Electronics Terms*, 5th ed., 1992.

Resonance (molecular structure)

A feature of the valence-bond method, which is a mathematical procedure to obtain approximate solutions to the Schrödinger equation for molecules. The term came into use because the procedure is similar to that describing how weakly coupled tuning forks, pendulums, and such resonate, that is, transfer energy back and forth to one another. The valence-bond method is based on the theorem that if two or more solutions to the Schrödinger equation are available, certain linear combinations of them will also be solutions. It has this basis in common with its rival, the molecular orbital method. The valence-bond and molecular orbital approaches are both approximations and, if carried out to their logical and exact extremes, must yield identical results; nevertheless, both are often described as theories. In the valence-bond theory, combinations of solutions represent hypothetical structures of the molecule in question. These structures are said to be resonance (or contributing) structures, and the real molecule is said to be the resonance hybrid (or just simply the hybrid) of these structures. *See* MOLECULAR ORBITAL THEORY; SCHRÖDINGER'S WAVE EQUATION.

After it was demonstrated that the Schrödinger equation provided an exact three-dimensional description of the hydrogen atom, and thus that it was the key to the description of all chemical species, the resonance concept was developed to deal with the much more complex problem of finding solutions for multielectron atoms and for the multiatom molecules. The resonance theory was initially more popular than the molecular orbital approach, since it was applicable and understandable in qualitative terms so that the chemist could forego the difficult mathematics altogether. Also, the contributing structures that could be used were the classical ones that had been in use all along, and the resonance theory provided a solution for a molecule which had baffled and preoccupied chemists for a century—benzene. The principal use of resonance still lies in the qualitative description of molecules whose properties would otherwise be difficult to understand. *See* BENZENE.

Structural formulas. It had long been a tradition in chemistry, especially in organic chemistry, to portray molecules in terms of the symbols of the atoms that are part of the molecules, with lines drawn between them to indicate chemical (covalent) bonds. Thus, methane, which is the principal component of natural gas and which has the formula CH_4, can be pictured to indicate that four hydrogen atoms are bound to a carbon atom (**1**); alternatively, the slightly more complex structure (**2**) is often drawn to indicate that

(**1**)   (**2**)
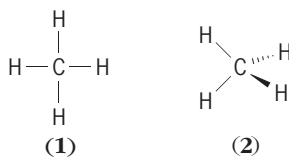
the real molecule is not planar, but tetrahedral; that

is, the four hydrogen atoms are at the corners of a (hypothetical) tetrahedron which has the carbon atom at the center. The lines drawn between the atoms denote the electron pairs which provide the bonds. Almost all molecules can be represented by such structures, and conversely, correctly devised structures may (with certain exceptions) be expected to correspond to molecules even if as yet they are not known either in nature or in the laboratory.

Formulating a correct structure requires knowledge of the number of electrons each atom has in its outer (valence) shell, and application of the so-called octet rule, which states that most atoms strive in their compounds to reach a total of eight electrons in this shell. Exceptions to this rule are the lightest few atoms—especially hydrogen (H), which can accommodate only two electrons—and the heavier ones which may have more than eight. Thus, the rule is not rigid, but is a generalization that is especially useful in biologically significant molecules, which are primarily made of atoms which obey it, namely, carbon (C), nitrogen (N), and oxygen (O). The isolated atoms, H, C, N, and O, have one, four, five, and six electrons, respectively, in their valence shells. Thus methane could also be represented as structure (**3**). The carbon atom in this molecule is

(**3**)
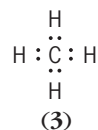
surrounded by four hydrogen atoms; it shares a pair of electrons with each of these, one contributed by the hydrogen atom and one by the carbon. In this way, the needs for all five atoms (two electrons for each H—, and eight for the C atom) are satisfied. In a similar way, the correct structures can be readily devised for such molecules as ammonia, NH_3 (**4**), water, H_2O (**5**), and ethane, C_2H_6 (**6**), another constituent

(**4**)   (**5**)   (**6**)
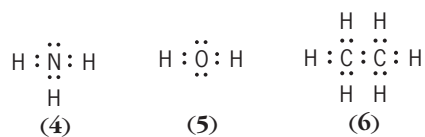
of natural gas. The unshared electron pairs in such representations are often simply omitted, but understood to be there. The structures of such molecules as ethylene, C_2H_4 (**7**), acetylene, C_2H_2 (**8**), and formaldehyde, CH_2O (**9**), represent slight

(**7**)   (**8**)   (**9**)
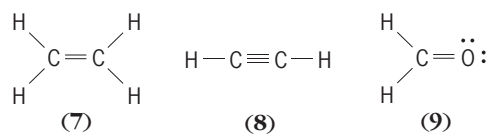
extensions in that two or even three electron pairs may be used to bind first-row atoms such as C, N, and O together. One of the interesting chemical properties of such molecules containing multiple bonds is

their tendency to form single bonds; thus, formaldehyde is readily converted into the new molecule trioxane (**10**) as shown in the reaction below.
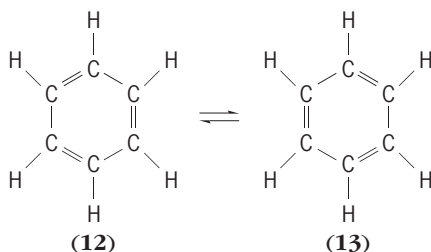

(**10**)
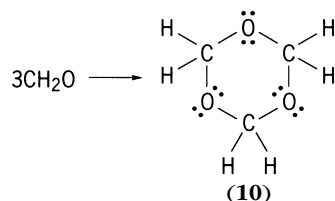
Benzene structure. Chemists had learned to deduce the atomic composition of molecules and the rudiments of the structural theory in the first half of the nineteenth century. The extension to three dimensions came in the 1870s, and that of the electron pair followed in the beginning of the twentieth century. But through most of this period the substance benzene posed a baffling challenge to organic chemists: in spite of its relatively simple formula, C_6H_6, they were unable to conceive of a suitable structure for it. While a great many structures were proposed, the properties of benzene corresponded to none of them. Thus, it was known that there were three and only three disubstituted benzenes (that is, benzenes in which two of the six hydrogen atoms had been replaced by other monovalent atoms such as chlorine, Cl). Furthermore, although it is difficult to avoid writing several multiple bonds, benzene and its derivatives have little or no tendency to undergo reactions known as additions, which typically convert such multiple bonds into single ones. Thus, structure (**11**) could be written for benzene, but

(**11**)
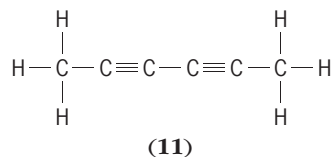
would not be correct. Such a substance should have only two disubstituted derivatives (with the substituting atoms either at the same or at opposite ends of the six-carbon chain). A compound that does have this structure is known, but it is strongly sensitive to addition reagents.

In the early 1870s F. A. Kekulé proposed a revolutionary idea: benzene must be represented by two structures (**12**) and (**13**) rather than one, and all compounds containing the benzene skeleton must be subject to a rapid equilibration (oscillation) between the two. Thus, whereas four disubstituted benzenes

(**12**)   (**13**)

(**14**)–(**17**) would be expected on the basis of a

(**14**)   (**15**)

(**16**)   (**17**)
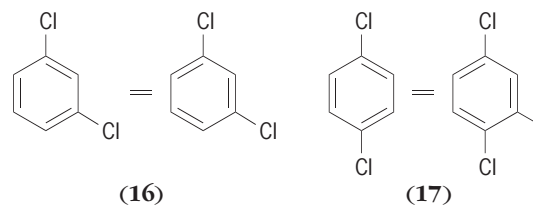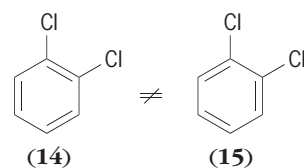
single structure for benzene, the equilibration between (**14**) and (**15**) was proposed to be so facile that only a mixture of the two can be isolated. This was a new development; chemists at that time assumed that every structure that can be drawn must correspond to an isolable compound, and that its chemical conversion to another compound requires some action on the part of a chemist: heating, the addition of chemical reagents, and so on. Many instances of such labile structures are now known; they are described as tautomers. However, benzene is not among them. *See* TAUTOMERISM.

Hybrid structure. Kekulé's description of benzene was not completely satisfactory. While it accounted for the number of isomers, it did not explain why the compound failed to exhibit reactivity indicating the presence of multiple bonds. The problem therefore continued to attract attention, but its final solution had to wait until the advent of quantum mechanics in the early part of the twentieth century. In a sense, this solution is an expansion of Kekulé's oscillating pair: the so-called activation energy (the energy which must be imparted to a molecule in order to make it overcome the barrier that keeps it from being converted into another molecule) is negative in the case of benzene with respect to the oscillation, and this molecule therefore exists neither as (**12**) nor as (**13**) at any time, but it is in an intermediate form (**18**) all the time. This intermediate structure of

(**18**)
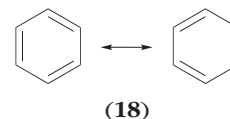
benzene is described in terms of Kekulé's structures with the symbol ↔ between them; this is intended to signify that benzene has neither structure, but in fact is a hybrid of the two. The properties of benzene are thereby indicated to be those of neither (**12**) nor (**13**), but to be intermediate between the two. Thus, since single bonds are longer than double bonds, it might be expected from the Kekulé structures that the molecule should be irregularly shaped; in fact, benzene is known to have the shape of a regular hexagon. All six of the carbon-carbon bonds are equally long, and their length is intermediate

between those of normal single and double bonds. The only property of the hybrid which is not intermediate between those of the hypothetical contributing structures is the energy: the energy of a resonance hybrid is by definition always at a minimum. This fact is responsible for the abnormal reluctance of benzene to undergo addition reactions; such reactions would lead to products that no longer have the resonance energy.

Other structures. Although benzene is the classical example of resonance, the phenomenon is certainly not limited to it; other instances include carboxylate anions such as formate (**19**) and guanidinium cation (**20**). Furthermore, the properties of all compounds

(19)

(20)
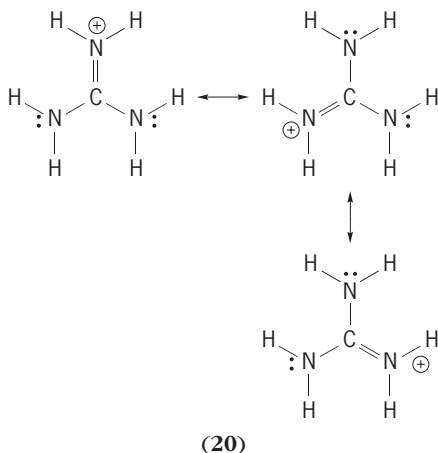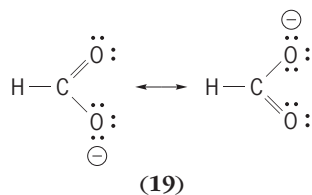
are affected by resonance to some degree. Thus, the properties of acrolein (**21**) suggest that the molecule

(21)
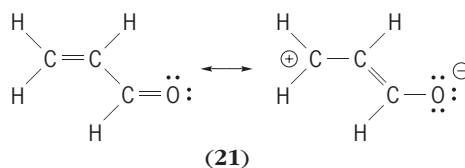
cannot be adequately described by a single structure: the single carbon-carbon bond length is rather short, and the double bonds are longer than normal. Thus, the dipolar structure is making a contribution. The main difference in the preceding cases is that the contributions are not equal; the weight of the dipolar structure is much less than half. In general, this inequality of weight will occur whenever the two structures drawn are different. *See* STRUCTURAL CHEMISTRY.

Bond order. This is a useful numerical device to indicate the occurrence of resonance in various molecules. The bond orders of single, double, and triple bonds are 1, 2, and 3, respectively; that of the carbon-carbon bonds in benzene and of the carbon-

oxygen bonds in carboxylate anions is 1.50, and that of the carbon-nitrogen bonds in guanidinium ion is 1.33. The importance of this concept is its connection with experimental observables such as the stretching frequencies of infrared absorption spectroscopy, bond lengths as determined by means of x-ray diffraction, and dipole moments. On such grounds, the bond order of the carbon-oxygen bond in acrolein may be estimated to be 1.80; in other words, the zwitterion contributes about 20% to the structure of this molecule.

No-bond resonance. Bond orders can in principle have values less than one. An example of this phenomenon is furnished by the molecule diborane, B_2H_6. Boron has three valence electrons and hydrogen has one; it might therefore be supposed that these two elements should be able to form the borane BH_3, with the structure (**22**). However, this structure

(22)
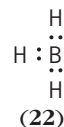
would clearly constitute a violation of the octet rule for the boron atom. Accordingly, two of the entities combine to form the the molecule diborane, which in resonance symbolism can be represented as (**23**). This molecule differs from the apparently

(23)
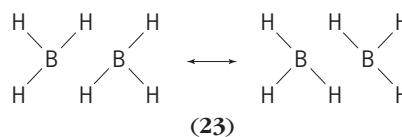
analogous ethane (**6**). In diborane, there are four hydrogen atoms bound to boron by bonds of order 1, and two that are bound to boron by bonds of order 0.5; in ethane, all six are bound to carbon by simple single bonds. Both contributing structures show atoms bound together but without a bond or electron pair—hence the name, no-bond resonance. The best-known examples occur in the boron hydrides and in carbocations such as the *tert*-butyl cation (**24**),


(24)
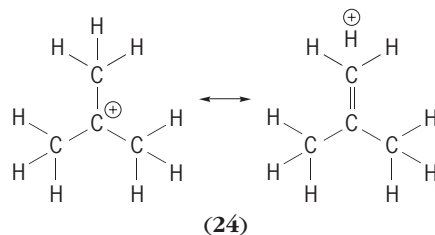
where any of the other eight hydrogen atoms could be chosen as the nonbonded one. *See* REACTIVE INTERMEDIATES.

Equilibration versus resonance. The distinction between equilibrating structures and resonance structures is not always as clearly in favor of the latter as it is in benzene, the resonance energy of which is about 36 kcal/mole. The cyclooctatetraene structures (**25**) and (**26**) furnish a clear example of the absence of

(25)

(26)
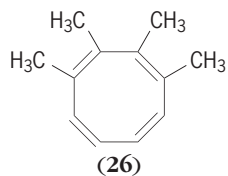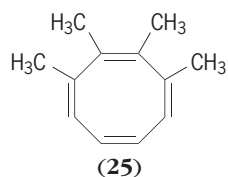
resonance. These so-called bond-shift isomers correspond to different, stable, and isolable compounds that equilibrate only slowly, even at elevated temperatures; the energy barrier between them is about 30 kcal/mole. Between such extremes as benzene and cyclooctatetraene, there are many instances of degenerate equilibration (reactions that give a product identical with the reactant) in which the rates are extremely fast; cyclobutadiene **(27)**

(27)
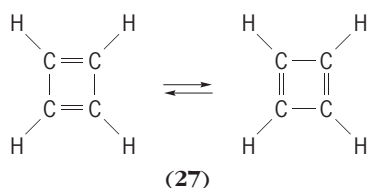
is an example of a species that equilibrates many times per second, even at temperatures as low as a few kelvins. On the other hand, there are also compounds that are hybrids of identical structures but that are not highly stablized; some members of the thiothiophthene family **(28)** exemplify this class.

(28)
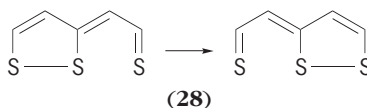
The question of resonance versus rapid equilibration has in fact not been settled in many instances; the so-called nonclassical carbocations such as **(29)**, the semibullvalenes such as **(30)**, and the mixed valence complexes **(31)** are among them, and this general question has become a flourishing research

(29)

(30)

(31)
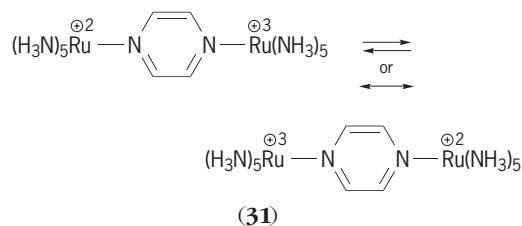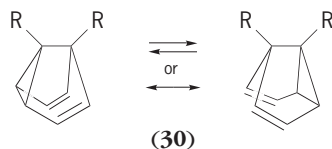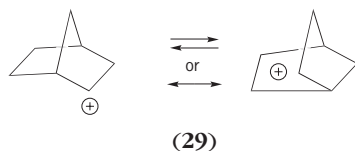
area. *See* ALICYCLIC HYDROCARBON; FLUXIONAL COMPOUNDS.

Limitations. The theory of resonance owed its initial success primarily to the fact that molecules could continue to be described in terms of the same structures that chemists had used for a century. However, many organic chemists have looked askance at no-bond structures. Moreover, use of the concept led to new questions that eventually began to undermine its popularity. Thus, its simple language cannot explain why cyclobutadiene is extremely unstable: it resisted all attempts at synthesis for many years, and it readily converts into other substances at temperatures as low as 20 K ($-424°$F). It is not obvious why cyclobutadiene is apparently not subject to resonance stabilization just as benzene is. The molecular orbital treatment can readily account for this observation; it also provides a much simpler account for the spectral properties of organic compounds. While it represented a more radical departure from the traditional structures and language of classical chemistry, it has been accepted. Although the molecular orbital approach has largely supplanted the valence-bond method, the resonance language remains so convenient that it is still used. *See* CHEMICAL BONDING; CONJUGATION AND HYPERCONJUGATION; MOLECULAR STRUCTURE AND SPECTRA; ORGANIC CHEMISTRY; QUANTUM CHEMISTRY. William J. le Noble

Resonance (quantum mechanics)

An enhanced coupling between quantum states with the same energy. The concept of resonance in quantum mechanics is closely related to resonances in classical physics.

Frequency and energy matching. A child's swing provides a simple example of resonances. The swing is a mechanical pendulum with a natural frequency that depends only on the length of the suspending rope. Only rhythmic pushing with the same frequency as the natural frequency of the swing, that is, at resonance, is effective in building up a swinging motion of a reasonable amplitude. Other examples of resonances, in acoustics and in alternating-current circuits (tuning a radio or television to a particular broadcasting channel's frequency), display this same feature. *See* PENDULUM; RESONANCE (ACOUSTICS AND MECHANICS); RESONANCE (ALTERNATING-CURRENT CIRCUITS).

The matching of frequencies is central to the concept of resonance. An example is provided by waves, acoustic or electromagnetic, of a spectrum of
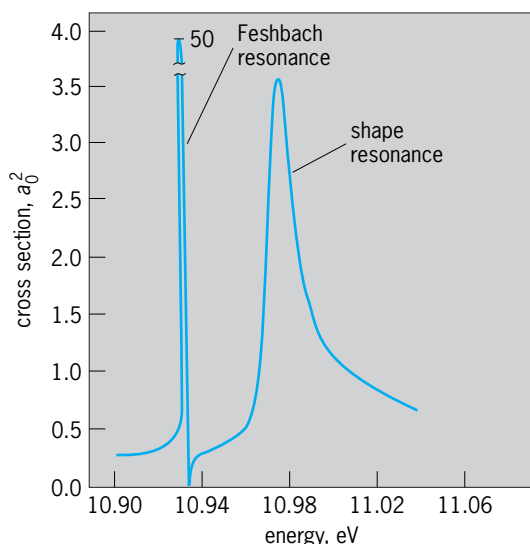
Fig. 1. Two resonances of the electron + hydrogen atom complex seen when the H⁻ ion absorbs light of energy shown. Cross section is measured in terms of the square of the Bohr radius, $a_0 \cong 0.5 \times 10^{-10}$ m. (*After H. C. Bryant et al., Observation of resonances near 11 eV in the photodetachment cross section of the H⁻ ion, Phys. Rev. Lett., 38:228–230, 1977*)

frequencies propagating down a tube or waveguide. If a closed side tube is attached, its characteristic natural frequencies will couple and resonate with waves of those same frequencies propagating down the main tube. This simple illustration provides a description of all resonances, including those in quantum mechanics. The propagation of all quantum entities, whether electrons, nucleons, or other elementary particles, is represented through wave functions and thus is subject to resonant effects. Indeed, nanometer-size channels for electron flow in semiconductors, fabricated by modern lithographic techniques, directly realize the above example. The electron current exhibits resonances. *See* ACOUSTIC RESONATOR; CAVITY RESONATOR; HARMONIC (PERIODIC PHENOMENA); WAVEGUIDE.

An important allied element of quantum mechanics lies in its correspondence between frequency and energy. Instead of frequencies, differences between allowed energy levels of a system are considered. In the presence of degeneracy, that is, of different states of the system with the same energy, even the slightest influence results in the system resonating back and forth between the degenerate states. These states may differ in their internal motions or in divisions of the system into subsystems. The above example of wave flow suggests the terminology of channels, each channel being a family of energy levels similar in other respects. These energies are discretely distributed for a closed channel, whereas a continuum of energy levels occurs in open channels whose subsystems can separate to infinity. The energy of quantum systems replaces frequency in the classical counterpart. If all channels are closed, that is, within the realm of bound states, resonance between degenerate states leads to a theme of central importance to quantum chemistry, namely,

stabilization by resonance and the resulting formation of resonant bonds. *See* DEGENERACY (QUANTUM MECHANICS); ENERGY LEVEL (QUANTUM MECHANICS); RESONANCE (MOLECULAR STRUCTURE).

Resonances in scattering. Resonances occur in scattering when at least one channel is closed and one open. Typically, a system is divided into two parts: projectile + target, for example, electron + atom, or nucleon + nucleus. One channel consists of continuum states with their two parts separated to infinity. The other, closed channel, consists of bound states. In the atomic example, a bound state of the full system would be a state of the negative ion and, in the nuclear example, a state of the larger nucleus formed by incorporating one extra nucleon in the target nucleus. Degeneracy implies that the bound state is an excited state of the complex. Thus, in the
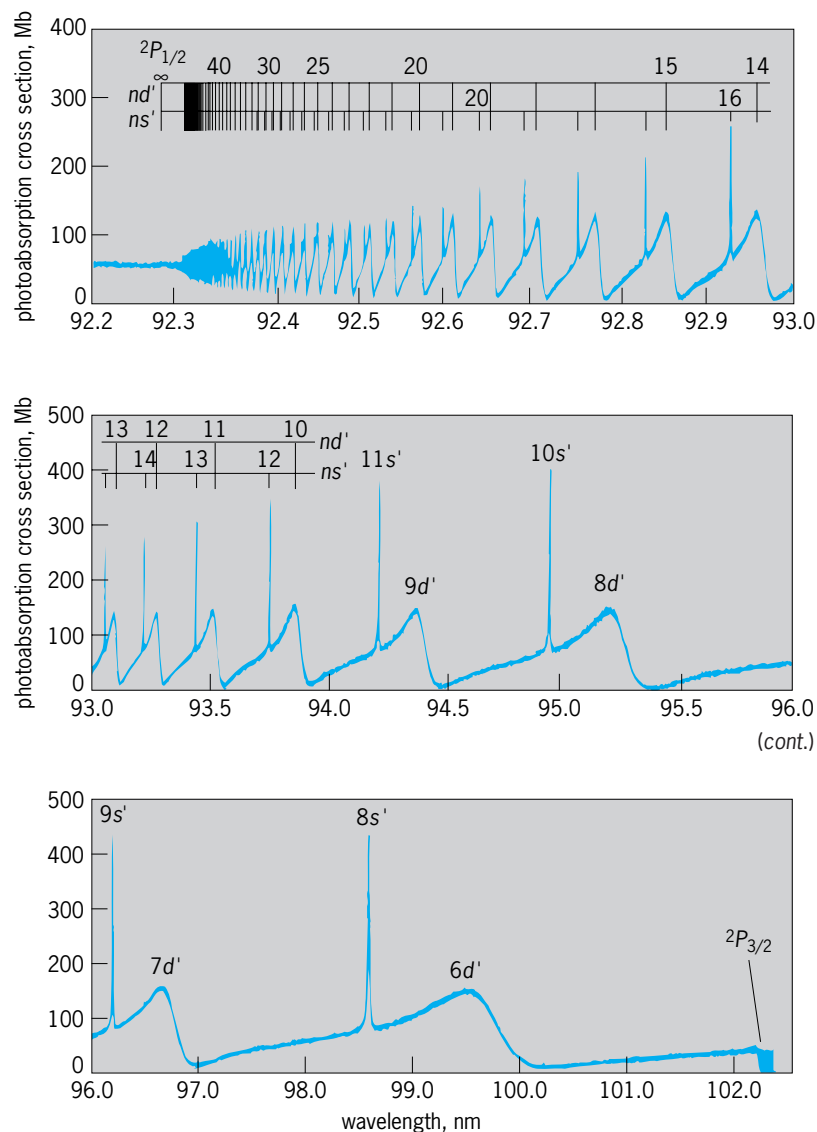

(*cont.*)

Fig. 2. Two prominent sequences of resonances seen in photoabsorption by the krypton atom between its two lowest ionization limits, marked as $^2P_{3/2}$ and $^2P_{1/2}$. Resonances are labeled ns' or nd', $n = 1, 2, 3, \ldots$. Decreasing photon wavelength translates to increasing energy. 1 Mb $= 10^{-22}$ m². (*After K. Maeda, K. Ueda, and K. Ito, High-resolution measurement for photoabsorption cross sections in the autoionization regions of Ar, Kr, and Xe, J. Phys. B: At. Mol. Opt. Phys., 26:1541–1555, 1993*)

example of electron + hydrogen atom, the complex's state is denoted by H^{-**}, the two asterisks showing that both electrons are excited and have principal quantum numbers n larger than 1. For the lowest such H^{-**}, the electron is bound to excited hydrogen with $n = 2$, its energy lying slightly below the threshold energy with hydrogen in $n = 2$ and an electron at infinity. This state lies in the $n = 2$ channel. At that same energy lies a continuum state of the $n = 1$ channel, of elastic electron + hydrogen scattering, with hydrogen in the ground state ($n = 1$) and an electron with asymptotic kinetic energy of $\simeq$10 eV. *See* ATOMIC STRUCTURE AND SPECTRA; QUANTUM NUMBERS.

The complete and correct description of the state of the system is as a superposition of the bound and the continuum states. It is neither a pure bound state nor a pure continuum state, and different labelings are possible. Thus, starting from the bound-state end, a quasibound state is referred to. The mixing of the continuum state indicates that the system initially prepared as H^{-**} falls apart (decays), leaving behind a hydrogenation ($n = 1$), with a 10-eV electron escaping to infinity. A characteristic lifetime describes this decay, called autodetachment. Starting from the other end of electron + hydrogen ($n = 1$) scattering, at the energy corresponding to the bound state, the system resonates from continuum state into bound state where the electron is trapped in the H^{-**} state. This trapping is only temporary because of the continuum state character within the superposition, the electron returning eventually into the continuum channel. In the elastic scattering of electrons from hydrogen ($n = 1$), the temporary trapping when the energy approximates that of the bound state leads to an anomaly in the scattering cross section, the common manifestation of the resonance. *See* SUPERPOSITION PRINCIPLE.

A resonance may also be observed in quantities other than the cross section, such as the phase shift which typically rises through π radians. The energy range encompassing this rise (that is, the range covered by the cross-section anomaly) is called the resonance width, which is inversely related to the lifetime for decay of the quasibound state in the complementary description. A closely related parameter that measures the rate of rise of the phase shift is called the time delay, namely, the duration of the electron's temporary trapping in the negative-ion state as it scatters from the atom. Besides the energy position and the width, the shape of the cross section is characterized by a so-called profile index. *See* SCATTERING EXPERIMENTS (ATOMS AND MOLECULES); SCATTERING EXPERIMENTS (NUCLEI).

Feshbach and shape resonances. Such a resonance in the electron + hydrogen system appears as a sharp feature lying just below the hydrogen ($n = 2$) threshold at 10.95 eV (**Fig. 1**). This feature is called a Feshbach resonance, after the discussion by H. Feshbach of analogous features in nuclear physics. A second type of resonance, referred to as a shape resonance, appears as a second, broader feature. Its name arises because the bound and continuum aspects are now associated with a subdivision of the real three-dimensional space of motion reflecting the shape of the potential between projectile and target. This shape consists typically of a barrier separating an inner deep well from a shallow, asymptotically vanishing potential at large separation distances. Atomic examples arise from combinations of centrifugal and Coulomb (or dipole) potentials, whereas a nuclear example results in alpha-particle radioactivity from
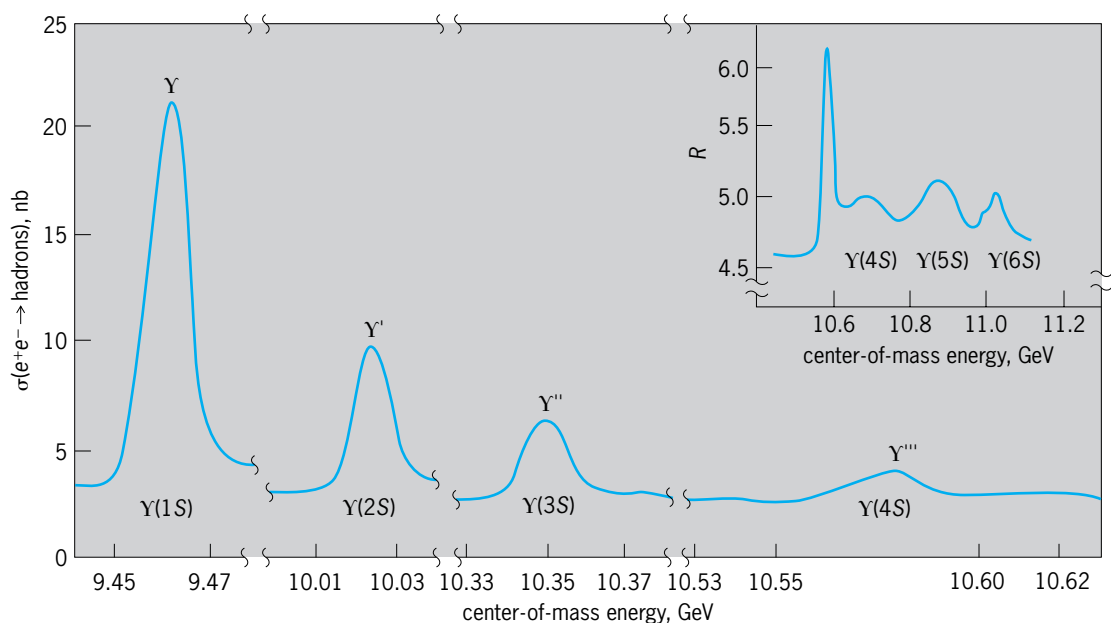

Fig. 3. Quasibound quark-antiquark states called upsilon, seen as resonances in the collision cross section (σ) of electrons and positrons. Data from two different detectors, one of which is shown in an inset as the ratio $R = \sigma(e^+e^- \rightarrow \text{hadrons})/\sigma(e^+e^- \rightarrow \text{muons})$, show six resonances. 1 nb $= 10^{-37}$ m^2. (*After D. Besson et al., Observation of new structure in the e^+e^- cross section above the $\Upsilon(4S)$, Phys. Rev. Lett., 54:381–384, 1985*)

the combination of attractive nuclear and repulsive Coulomb forces between the alpha particle and the residual daughter nucleus. When the inner well supports a bound state above the zero of the energy, this state appears as a quasibound resonance because the alpha particle or the electron can tunnel quantum-mechanically through the barrier and escape to infinity. The distinguishing characteristic of a shape resonance is, therefore, that it lies just above the threshold to which it couples, in contrast to a Feshbach resonance, which lies below it. *See* ALPHA PARTICLES; RADIOACTIVITY.

Multiple resonances. Whole sequences of resonances are often observed because a closed channel typically supports many quasibound energy levels (**Figs. 2** and **3**).

Another direction for generalization involves many coupled channels, some open and others closed. Energy levels in any closed channel resonate with degenerate energy states of all the open channels and show up as cross-section anomalies, elastic or inelastic. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.   A. R. P. Rau

Bibliography. A. de Shalit and H. Feshbach, *Theoretical Nuclear Physics*, vol. 1, 1974, reprint 1992; U. Fano and A. R. P. Rau, *Atomic Collisions and Spectra*, 1986; E. Merzbacher, *Quantum Mechanics*, 3d ed., 1997; R. G. Newton, *Scattering Theory of Waves and Particles*, 2d ed., 1982; D. H. Perkins, *Introduction to High Energy Physics*, 4th ed., 2000.

Resonance ionization spectroscopy

A form of atomic and molecular spectroscopy in which wavelength-tunable lasers are used to remove electrons from (ionize) a given kind of atom or molecule. Laser-based resonance ionization spectroscopy (RIS) methods have been developed and used with ionization detectors, such as proportional counters, to detect single atoms. Resonance ionization spectroscopy can be combined with mass spectrometers to provide analytical systems for a wide range of applications, including physics, chemistry, materials sciences, medicine, and the environmental sciences.

Theory. When an atom or molecule is irradiated with a light source of frequency ν, photons at this selected frequency are absorbed only when the energy $h\nu$ (h is Planck's constant) is almost exactly the same as the difference in energy between some excited state (or intermediate state) and the ground state of the atom or molecule. If a laser source is tuned to a very narrow bandwidth (energy width or frequency width) at a frequency that excites a given kind of atom (**Fig. 1**), it is highly unlikely that any other kind of atom will be excited. An atom in an excited state can be ionized by photons of the specified frequency, ν, provided that $2h\nu$ is greater than the ionization potential of the atom. Even though the final ionization step can occur with any energy above a threshold, the entire process of ionization is a resonance one, because the intermediate state must

first be excited in a resonance photon absorption. In sharp contrast to all other ionization means, such as x-rays, energetic charged particles, or radioactive sources, resonance ionization spectroscopy is a selective process in which only those atoms that are in resonance with the light source are ionized. Modern pulsed lasers can provide a sufficient number of photons in a single pulse to remove one electron from each atom of any selected type. The first transition from the ground state to the intermediate state is easily saturated as these states come into equilibrium with a moderate photon flux. To saturate the final transition from the intermediate state to the ionization continuum (to remove one electron from each atom) requires laser pulses of considerable energy per pulse and is generally difficult. These frequency-tunable (or wavelength-tunable) lasers have made resonance ionization spectroscopy a potential method for the sensitive (and highly selective) detection of nearly every type of atom in the periodic table. *See* ATOMIC STRUCTURE AND SPECTRA; IONIZATION POTENTIAL; LASER; LASER SPECTROSCOPY; PHOTOIONIZATION; RESONANCE (QUANTUM MECHANICS).

Laser schemes. The simple type of resonance ionization spectroscopy discussed above (Fig. 1) was used to illustrate the resonance ionization process; more elaborate schemes are required in practice. Fortunately, many others are possible. An atom already in an excited state can absorb other photons of another wavelength to create a more highly excited state. Each kind of atom has its own unique set of possible excited states, much like a ladder with steps that are widely spaced near the ground and are more closely spaced as the steps converge to the ionization continuum. (An energy equal to the ionization potential is required to ionize a given atom; the liberated electron is no longer bound, and is thus in the ionization continuum.) After a few of these steps have been taken, another photon may be used to complete the resonance ionization process by the ejection of an electron with a small kinetic energy into the ionization continuum. Alternatively, if the final bound state
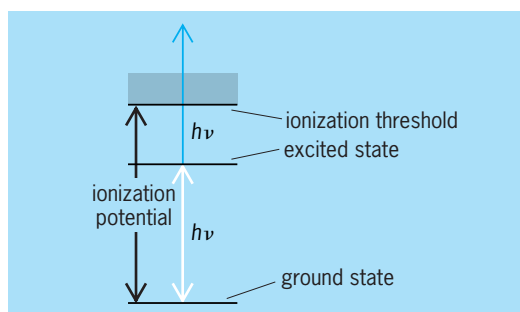

Fig. 1. Basic laser scheme for resonance ionization spectroscopy (RIS). The atom or molecule is irradiated by a light source with frequency ν, or photons of energy $h\nu$, where h is Planck's constant. The shading shows the beginning of the ionization continuum, which starts at several electronvolts above the ground state, depending on the type of atom. An electron from the ground state of the atom or molecule is being excited by two photons, thus resonantly ionizing the atom or molecule.

is very close to the edge of the continuum, a pulsed electrical field can be applied to eject the electron. Thus, a rich variety of laser schemes are in use, so that any atom in the periodic table can be detected with resonance ionization spectroscopy. However, those atoms with high ionization potentials, such as the inert gases, require the more difficult resonance ionization schemes. *See* ENERGY LEVEL (QUANTUM MECHANICS).

Ultrasensitive detection. Both high selectivity and the ultrasensitive detection of atoms have been demonstrated by pulsing a laser directly through a proportional counter. Proportional counters, like Geiger-Muller counters, will detect a single electron. Therefore, if a laser beam is directed through a proportional counter to remove one electron from each of the atoms of a selected type, selective detection of a single atom is possible. In the original demonstration of one-atom detection, it was proved that one atom of cesium could be selected out of 10^{19} atoms of the counting gas (argon and methane). *See* GEIGER-MÜLLER COUNTER; IONIZATION CHAMBER.

Applications. Resonance ionization spectroscopy is used in a wide variety of applications such as the trace analysis of low levels of elements in extremely pure materials, for example, semiconductors in the electronics industry. For these analyses, sputter-initiated resonance ionization spectroscopy (SIRIS) is achieved with the apparatus shown in **Fig. 2**. An argon ion beam sputters a tiny cloud of

Fig. 2. **Sputter-initiated resonance ionization spectroscopy (SIRIS).** (*After H. F. Arlinghaus, M. T. Spaar, and N. Thonnard, Quantitative and sensitive profiling of dopants and impurities in semiconductors using sputter-initiated resonance ionization spectroscopy, J. Vac. Sci. Technol. A, 11:2317–2323, 1993, updated by T. Whitaker, Atom Sciences, Inc.*)

atoms from a sample placed in a high-vacuum system and a pulsed laser is tuned to detect the specified impurity atom. These lasers create positive ions (using the resonance ionization process) of a given atomic number within the cloud of neutral atoms released from the solid material. Ions are then further selected for atomic mass by using a mass spectrometer. The combination is doubly selective and eliminates isomers (atoms having the same mass but different atomic numbers) as well as certain backgrounds resulting from fragmentation of molecules by electron impact in the usual mass spectrometer. Trace elements present at the level of parts per billion and, in some cases, parts per trillion can be detected. Further, imaging capability provides maps of these impurities on the surface of materials, while depth profiling gives the concentration as a function of depth into the materials. This kind of information is of increasing importance to the development of very large scale integrated devices. *See* INTEGRATED CIRCUITS; MASS SPECTROMETRY; MASS SPECTROSCOPE; SPUTTERING.

The sputter-initiated resonance ionization spectroscopy method is also used for chemical and materials research, geophysical research and explorations, medical diagnostics, biological research, and environment analysis. When imaging capability is not needed, a simpler technique known as thermal-atomization resonance ionization spectroscopy (TARIS) may be used (instead of SIRIS) for the bulk analysis of materials. By simply using resonance ionization spectroscopy with ionization chambers or proportional counters, gas-phase work can be done to study the diffusion of atoms, measure chemical reaction rates, and investigate the statistical behavior of atoms and molecules. *See* CHEMICAL DYNAMICS; DIFFUSION.

Resonance ionization spectroscopy is used in sophisticated nuclear physics studies involving high-energy accelerators. The spacing of atomic energy levels in atoms is governed mainly by the protons in the atom's nucleus and the orbiting electrons. However, the spin of electrons causes a perturbation of these levels known as the fine structure. Further, there is a minute effect due to the spin of the nucleus, which gives rise to a hyperfine structure. Resonance ionization spectroscopy is used as an on-line detector to record the hyperfine structure of nuclei with short lifetimes and hence to determine several nuclear properties such as nuclear spin and the shape of nuclei. *See* FINE STRUCTURE (SPECTRAL LINES); HYPERFINE STRUCTURE; NUCLEAR STRUCTURE.

In the usual ion source, even with magnetic separation, the ion beam is contaminated with other isobars (atoms with the same mass number but different atomic numbers). The use of resonance ionization spectroscopy in the ion source (**Fig. 3**) essentially eliminates the isobars and leads to much more definitive determination of nuclear parameters. Thus, resonance ionization spectroscopy is used at both ends of major accelerators for nuclear physics research. *See* ION SOURCES; ISOBAR (NUCLEAR PHYSICS).
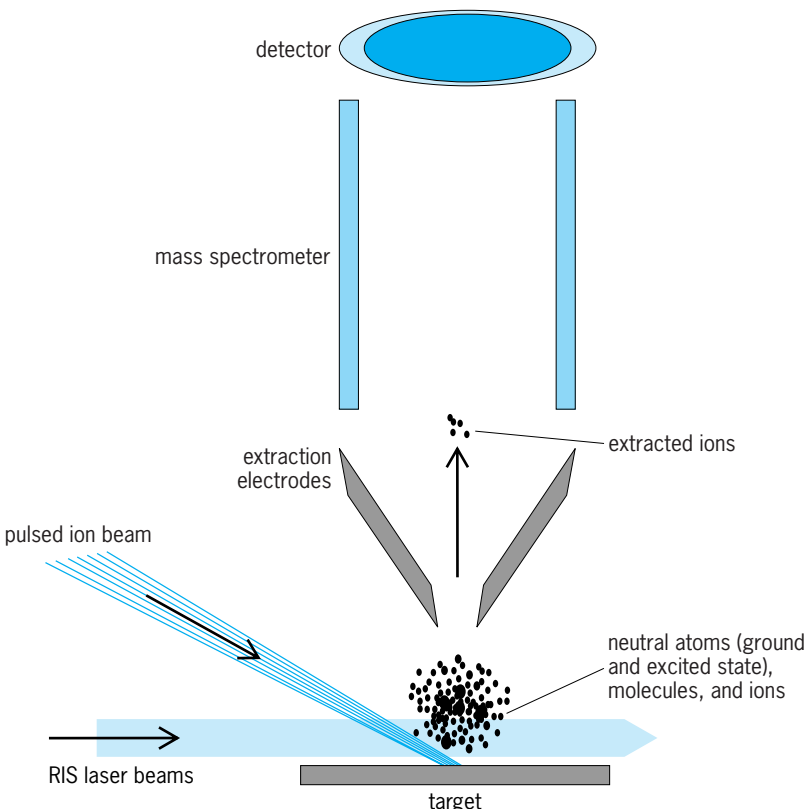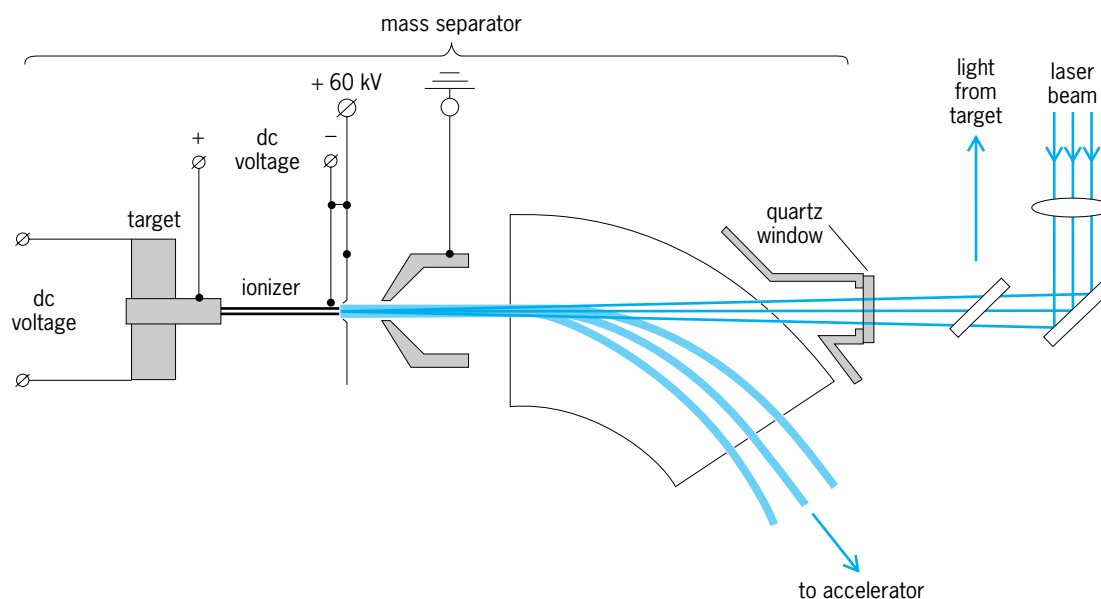
An example of the potential of resonance

Fig. 3. Selective ion source using resonance ionization spectroscopy. The laser beam comes to a focus at the ionizer after being reflected by the right diagonal mirror. The left diagonal mirror is used for precision timing of the laser pulse. (*After V. I. Mishin et al., Chemically selective laser ion-source for the CERN-ISOLDE on-line mass separator facility, Nucl. Instrum. Meth. Phys. Res., B73:550–560, 1993*)

ionization spectroscopy for weak-interaction physics is the detection of krypton-81, which results from neutrino capture in bromine-81. A detector using large tanks of a bromine solution could be placed underground (reducing cosmic-ray backgrounds) to study neutrinos from the Sun. Resonance ionization spectroscopy could be used to count the few atoms of krypton-81 directly, before radioactive decay. Decay counting of isotopes with long half-lives (for krypton-81 it is about 200,000 years), for samples of 100–1000 atoms, is obviously impossible. *See* NEUTRINO; SOLAR NEUTRINOS; WEAK NUCLEAR INTERACTIONS.

Resonance ionization spectroscopy is used for measurements of krypton-81 in the natural environment to determine the ages of polar ice caps and old ground-water deposits. Oceanic circulation and the mixing of oceans could also be studied by measuring the concentrations of noble-gas isotopes by resonance ionization spectroscopy. *See* DATING METHODS; RADIOISOTOPE (GEOCHEMISTRY).

G. Samuel Hurst

Bibliography. G. S. Hurst and M. G. Payne, *Principles and Applications of Resonance Ionization Spectroscopy*, 1988; G. S. Hurst et al., Resonance ionization spectroscopy and one-atom detection, *Rev. Mod. Phys.*, 51:767–819, 1979; V. S. Letokhov, *Laser Photoionization Spectroscopy*, 1987.

Respiration

Commonly, the processes by which living organisms take up oxygen and eliminate carbon dioxide in order to provide energy. A more formal, comprehensive definition is the various processes associated with the biochemical transformation of the energy available in the organic substrates derived from foodstuffs to energy usable for synthetic and transport processes, external work, and, eventually, heat. This transformation, generally identified as metabolism, most commonly requires the presence of oxygen and involves the complete oxidation of organic substrates to carbon dioxide and water (aerobic respiration). If the oxidation is incomplete, resulting in organic compounds as end products, oxygen is typically not involved, and the process is then identified as anaerobic respiration. The oxygen required in aerobic respiration is ultimately derived from the atmosphere or from oxygen dissolved in water. The process of the exchange and transport of oxygen from an animal's environment to the intracellular sites involved in metabolism, and the reverse transport and exchange of the end product carbon dioxide, is often called respiration, although it is only a part of the overall process.

The term external respiration is more appropriate for describing the exchange of oxygen and carbon dioxide between the organism and its environment. In most multicellular organisms, and nearly all vertebrates (with the exception of a few salamanders lacking both lungs and gills), external respiration takes place in specialized structures termed respiratory organs, such as gills and lungs.

Diffusion and convection. The ultimate physical process causing movement of gases across living tissues is simple passive diffusion. Diffusion is the random movement of molecules from regions of high concentration (activity) to regions of low concentration. Fick's law describing this process states that the rate of transfer of a gas is directly proportional to its concentration gradient and to the surface area. The concentration gradient is the difference in the concentrations at the two ends of the diffusion path,
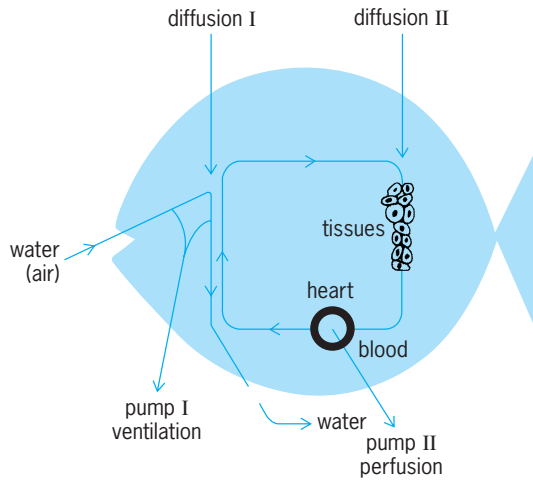
Fig. 1. Transport path for respiratory gases in a vertebrate. At diffusion site I, water or air exchanges gases with blood; at diffusion site II blood exchanges gases with metabolizing tissues. The rate of diffusion is increased by bulk convective transport of air or water and blood effected by a ventilatory pump (pump I) and a perfusion pump (pump II). The presence of hemoglobin in blood greatly increases its gas-transporting capacity and also facilitates the rate of diffusion across diffusion sites I and II. (*After A. G. Kluge, ed., Chordate Structure and Function, 2d ed., Macmillan, 1977*)

divided by the length of the path. Thus, the rate of gas transfer is inversely proportional to the tissue thickness (diffusion distance). Moreover, the rate of diffusion of gases through tissues is much slower than through water or air.

Respiratory gas exchange also depends on two convective fluid movements. The first is the bulk transport of the external medium, air or water, to
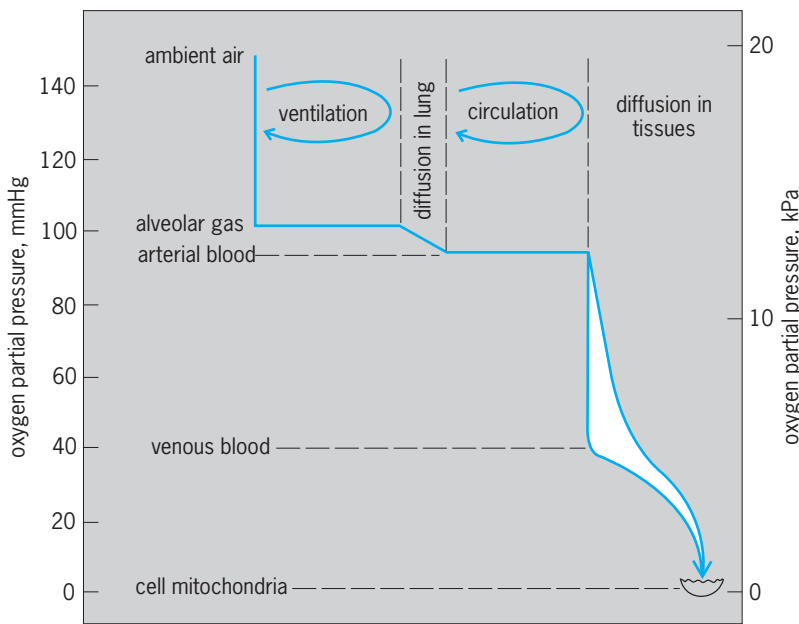

Fig. 2. Oxygen cascade from the ambient medium to the cellular metabolic sites in the mitochondria for a lung breathing animal. The oxygen partial pressure gradient is determined by the ventilation and circulation and by diffusion in the lung and tissues. 1 mmHg = 0.04 in. Hg. (*After S. C. Wood and C. Lenfant, eds., Evolution of Respiratory Processes, vol. 13 in Lung Biology in Health and Disease, Marcel Dekker, 1979*)

and across the external respiratory exchange surfaces. The second is the transport of coelomic fluid or blood across the internal surfaces of the respiratory organ. These two convective transports are referred to as ventilation and circulation (or perfusion), respectively. They are active processes, powered by muscular or ciliary pumps. In addition to providing bulk transport of the respiratory gases, the convection pumps are important in maximizing the concentration gradients for diffusion across the gas-exchanging tissue surfaces. Aquatic organisms living in air-saturated water, and supplied with oxygen by diffusion alone, cannot have a radius greater than about 0.02 in. (0.5 mm). Consequently, larger organisms that depend on aerobic energy production must be equipped with external and internal convective pumps.

In almost all vertebrates and many invertebrates, the circulating internal medium (blood, hemolymph, or coelomic fluid) contains a respiratory pigment, for example, hemoglobin or hemocyanin, which binds reversibly with oxygen, carbon dioxide, and protons. Respiratory pigments augment respiratory gas exchange, both by increasing the capacity for bulk transport of the gases and by influencing gas partial pressure (concentration) gradients across tissue exchange surfaces. *See* HEMOGLOBIN; RESPIRATORY PIGMENTS (INVERTEBRATE).

Respiration in a vertebrate illustrates well the transport chain that moves respiratory gases between the external medium and the intracellular metabolic sites (**Fig. 1**). The links in this chain are the two bulk flows, ventilation and perfusion, with their convective pumps; diffusion across the specialized exchange surfaces in the respiratory organs (lungs or gills) connecting ventilation and perfusion; and diffusion across the tissue capillaries to and from the sites of metabolism. The delivery, by this transport chain, of the molecular oxygen needed for aerobic metabolism is often referred to as the oxygen cascade (**Fig. 2**). In addition to the steps in the oxygen cascade, needed chemical reactions may be rate-limiting in respiratory gas exchange, particularly at the lower body temperatures of many poikilotherms. Two of these are the reaction of oxygen with hemoglobin and the dehydration of carbonic acid (H_2CO_3), the latter depending on the enzyme carbonic anhydrase.

The physiological adjustment of organisms to variations in their need for aerobic energy production involves regulated changes in the exchange and transport of respiratory gases. The adjustments are effected by rapid alterations in the ventilatory and circulatory pumps and by longer-term modifications in the respiratory properties of blood.

Respiratory media. Air and water differ qualitatively as respiratory media, and the differences are reflected in the structural and functional qualities of the external gas exchangers (respiratory organs) of aerial and aquatic breathers.

The primary quality of a respiratory medium is its capacitance for oxygen and carbon dioxide. The capacitance coefficient is defined as the increment in

concentration per increment in partial pressure. The capacitance coefficient for oxygen in water (equal to the physical solubility) is very low compared to air and depends on temperature and salinity. Due to the low capacitance of water for oxygen, aquatic breathers must ventilate 20–30 times more medium than air breathers to obtain the same amount of oxygen. The carbon dioxide capacitance of water and internal fluids, like blood, exceeds that for oxygen. Therefore, an aquatic breather will increase carbon dioxide tension in the ambient medium far less than it decreases oxygen tension, assuming that the changes are caused only by the animal's respiratory gas exchange. Similarly, the high ventilatory rate characteristic of aquatic gas exchange will be accompanied by very small differences between inspired and expired carbon dioxide tensions, and will cause low internal carbon dioxide tensions in body fluids and tissues. In seawater, and in some special freshwater habitats, carbonate and bicarbonate, or other buffer substances, increase the carbon dioxide capacitance by varying in concentration with changes in the carbon dioxide tension. Similarly, the chemical binding of the respiratory gases, oxygen and carbon dioxide, in blood containing a respiratory pigment will greatly increase the capacitance coefficient of both gases. Due to a number of properties of respiratory pigments, the capacitance coefficient for oxygen in blood will depend on the oxygen tension. Similarly, the capacitance coefficient for carbon dioxide will depend nonlinearly on the buffering properties of blood, and on the carbon dioxide tension. *See* DIVING.

Oxygen consumption. Most animals satisfy their energy requirements by oxidation of organic molecules obtained in food. Thus, the rate of oxygen uptake reflects the energy metabolism, or the metabolic rate, of an aerobic organism. The energy made available and transformed to heat when food stuffs are completely oxidized corresponds to 5.0 kilocalories (20.9 kilojoules) per liter oxygen for carbohydrate, 4.5 kcal (18.8 kJ) for protein, and 4.8 kcal (20.1 kJ) for fat. The respiratory quotient (RQ), which is calculated as the ratio between carbon dioxide produced and oxygen consumed in metabolism, varies between 0.7 for fat and 1.0 for carbohydrate. The RQ value is therefore an index to the foodstuffs being used by an organism. In the context of dynamics of external gas exchange, the RQ is often called the respiratory exchange ratio (RE). In these circumstances it implies how oxygen uptake and carbon dioxide output are divided between respiratory organs in animals depending on multiple modes of external gas exchange (such as amphibians).

The oxygen consumption of organisms at rest varies with their phylogenetic position, with their size, and with oxygen availability in their immediate environment. In addition, oxygen consumption increases with their food intake and during their physical activity. Homeothermic vertebrates, including birds and mammals, have much higher oxygen uptake rates than poikilothermic vertebrates, even when the comparison is made at similar body temperatures (98.6–104°F or 37–40°C). *See* HIBERNATION AND ESTIVATION.

Organisms with stable oxygen uptake rates over wide ranges of external oxygen concentration are classified as oxygen regulators, or as being oxygen-independent. In contrast, an oxygen uptake more or less linearly related to external oxygen availability is termed oxygen dependence, and the animal an oxygen conformer. Whereas homeotherm vertebrates are usually oxygen-independent, many aquatic invertebrates show total oxygen dependence. The critical oxygen tension is the environmental partial pressure at which an organism shifts from oxygen independence to oxygen dependence. For poikilotherm organisms, the critical oxygen tension varies with temperature.

For nearly all animal groups, the metabolic rate per unit body weight, or the weight-specific metabolic rate, decreases with increasing body weight. This implies that the metabolic requirements for oxygen and organic substrates in food are relatively much higher for smaller organisms than for larger ones. Resting metabolic rate (M) is defined by the equation $M = aW^b$, where W is body weight, and a and b are constants, with b varying between 0.6 and 0.9. The weight-specific metabolic rate is therefore expressed by $M/W = aW^{b-1}$.

The aerobic scope for activity is defined as the difference between maximal oxygen uptake and resting oxygen uptake. For poikilotherms, the scope for activity is greatly increased by temperature, up to a certain level. Homeotherms have aerobic scopes for activity many times greater than poikilotherms.

Respiratory organs. Nearly all respiratory organs of water-breathing animals are characterized by a unidirectional passage of water across the respiratory exchange surfaces. This design is basically different from that of the respiratory organs of air breathers where air typically enters and exits through the same channels as a tidal or bidirectional flow. The significance of this difference is that anatomical dead space is virtually eliminated in water breathers, and much higher oxygen tensions occur on the external respiratory surfaces than is possible in a tidal system where the conducting airways constitute dead space.

The gas exchange in the respiratory organs of postembryonic vertebrates and many invertebrates can conveniently be classified according to the four exchange types illustrated in **Fig. 3**.

The unidirectional ventilatory current in gills of fishes and many invertebrates offers a structural basis for a countercurrent exchange of gases between water and blood. Such a system is highly efficient and allows arterial oxygen partial pressure (Fig. 3*a*) to exceed the expired oxygen partial pressure (*e*) and, in theory, to approach the oxygen partial pressure in external water (*i*). The respiratory organs of birds are unique in design among vertebrates; the system includes air sacs that serve to ventilate the gas exchange units of the lung, the parabronchi. *See* COUNTERCURRENT EXCHANGE (BIOLOGY).
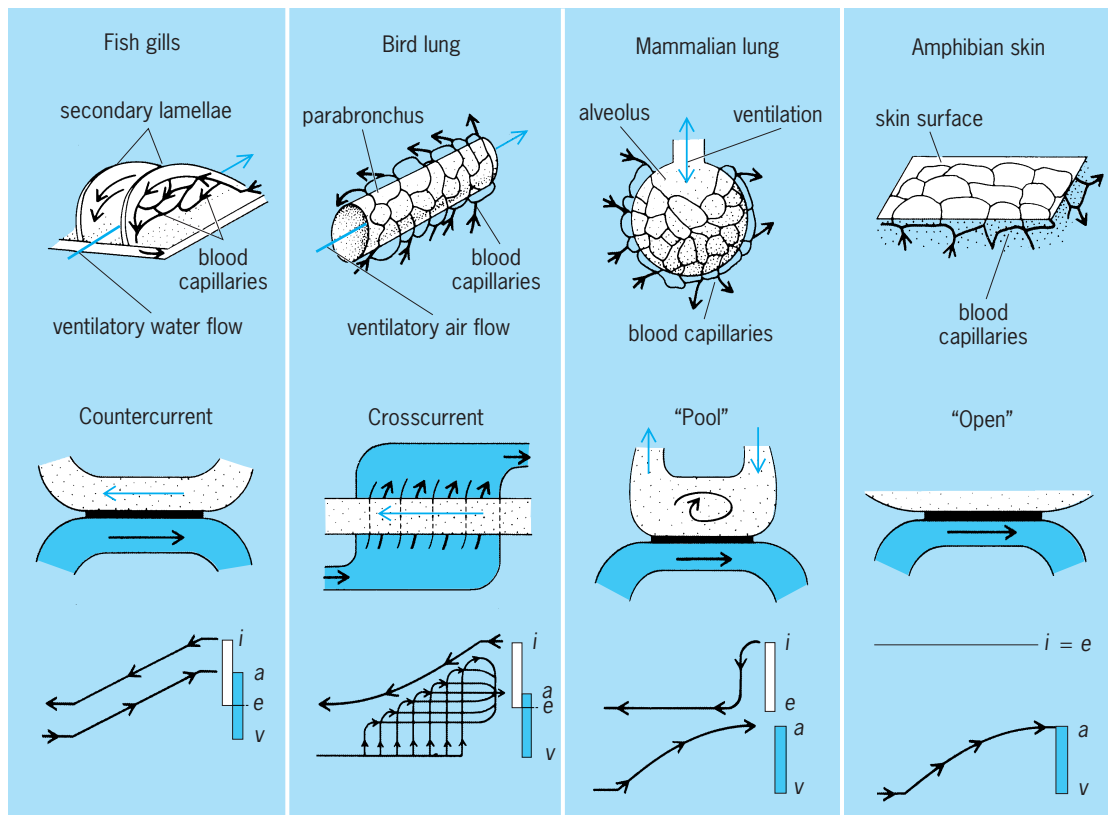
Fig. 3. Most common external gas exchangers with models of their exchange types. At the bottom, typical curves of oxygen tension changes in the external medium and blood along the exchange areas *i*, inspired; *e*, expired; *a*, arterial; *v* venous. (*After J. Piiper and P. Scheid, Comparative physiology of respiration: Functional analysis of gas exchange organs in vertebrates, Int. Rev. Physiol., vol. 14, 1977*)

Gas exchange in the parabronchi of the bird lung can best be analyzed as a crosscurrent exchange system (Fig. 3). In the parabronchi, as in fish gills, ventilation occurs in a flow-through, one-way manner. Gas exchange occurs by diffusion along blind-ending air capillaries radiating from the parabronchial lumen. Blood capillaries are profusely intertwined with the air capillaries. The partial-pressure profile for gas exchange in the parabronchial lung documents a high gas exchange efficiency, with larger arterial oxygen partial pressures than at the end of the parabronchus.

Mammalian lungs are referred to as alveolar lungs, and are homologous with the lungs in lungfishes, amphibians, and reptiles. These lungs have bidirectional conducting airways that, in the most advanced mammalian design, terminate in the blind-ending alveoli after 20–25 subdivisions (branchings). The lungs of humans may contain about 3×10^8 alveoli, representing a surface area of more than 860 ft^2 (80 m^2). Gas in the alveoli, and blood inside the profuse network of capillaries investing the alveolar walls (Fig. 3), are separated by less than 0.5 micrometer of intervening tissue. In lungfishes, amphibians, and reptiles an ascending degree of internal septation of the lung is the basis for the corresponding surface expansion of the alveolar surfaces. A comparison of vertebrate lungs shows lung volume varying in direct proportion to body mass, and lung surface

area positively correlated with metabolic rate, physical activity, and environmental factors. Increased lung surface area results when the number of alveoli increases and their individual size diminishes. The gas-to-blood diffusion distance also gets shorter in smaller or more active animals, correlating with their higher weight-specific metabolism. The bidirectional airways in lung breathers create a nonfunctional air space, called the anatomical dead space. It causes oxygen partial pressures in alveoli to be lower, and carbon dioxide partial pressures higher, than in outside air. Compared to fish gills, the gas exchange surfaces of lungs may be subject to smaller, more stable gas partial pressure gradients. The gas exchange in alveolar lungs can be modeled as a ventilated pool, and the oxygen partial pressures in the blood leaving a lung can never, as in fish gills or bird lungs, be higher than that in the medium leaving the gas exchange unit (Fig. 3). *See* LUNG.

The muscular basis for ventilation in lungfish and amphibians is a buccal force pump bringing air into the lungs under positive pressure; in reptiles, birds, and mammals the lungs are ventilated by aspiration or suction.

Breathing through the general body surface is phylogenetically the most primitive method, and it remains an important complement to external gas exchange for most aquatic invertebrates, fishes, and amphibians. In some amphibians, the skin is the sole site

of gas exchange. Skin gas exchange can be modeled as an infinite or open pool, typically with no organized convective movement of the external medium. The capillary network in the skin is not as well developed for exchanging gases with the external medium as are the capillaries in lungs, or in the lamellar blood spaces of fish gills. Moreover, the diffusion distance between the blood and the external medium in skin is longer than that in gills or lungs. Since skin is an unspecialized gas exchanger, it also lacks the vascular connections that are used for optimizing internal gas transport and that are characteristic of specialized organs. Of the four gas exchange models in Fig. 3, the exchange efficiency declines in the order: countercurrent exchange (fish gills), crosscurrent exchange (bird lungs), ventilated pool (alveolar lungs), and infinite pool (general body surface).

Tracheal respiration. In insects, the largest and most diversified order of terrestrial animals, gas exchange depends largely on diffusion. The basic design is a highly branching internal system of air-filled tubes (trachea) connected to the ambient air through controllable openings (spiracles). The smallest air tubes (tracheoles) may be only 1 μm in diameter; they ramify extensively, and even penetrate individual muscle fibers. In smaller insects, oxygen is transported through the tracheal system solely by diffusion; larger insects depend on convective air movement produced by ventilation. Insect flight muscles are strictly aerobic and during flight may consume more oxygen than any other animal tissue; the oxygen uptake increases 400–fold when activity begins. In insects, internal transport by the hemolymph is of no importance for respiratory gas exchange, but the organic substrates that are needed to support the high and variable aerobic metabolism must be delivered to the cells by hemolymph circulation. *See* INSECT PHYSIOLOGY.

Smaller insects may have an air-filled internal tracheal system with no external openings, so gas exchange must occur by diffusion through the body wall. In some aquatic insects the closed-off air-filled tracheal system may project into gill-like structures. These tracheal gills expand the general body surface and thereby augment the diffusive exchange of gas between the surrounding water and internal tracheal air. In the absence of an effective microvascular circulation of hemolymph, the usefulness of the system of closed air tubes for internal gas transport lies in the 250,000-times-more-rapid diffusion of gases in air than in water.

Other aquatic insects with an air-filled tracheal system make regular excursions to the surface to capture a bubble of air that is connected to the tracheal system and held in place by a specialized structure (plastron). When the insect dives and the air bubble becomes compressed, gases will exchange between the insect and the bubble. This exchange, and the large difference in the solubilities of oxygen and carbon dioxide in water, will set up partial pressure differences between the bubble and the surrounding water, favoring efflux of carbon dioxide over influx of oxygen. This system will function for a long time as a diffusion lung, but since carbon dioxide escapes the bubble more easily than oxygen that enters it from the water, the bubble will shrink, and its N_2 pressure will rise. The bubble will therefore gradually lose N_2 to the water, and the insect must return to the surface, to renew its N_2 supply and not its oxygen supply.

Blood circulation and gas transport. In animals using blood circulation for exchange and transport of respiratory gases, the efficiency of the transport function is highest if the vascular system conveys the most deoxygenated, carbon dioxide–rich blood to the respiratory organs, gills or lungs, and if the flow path of oxygen-rich blood, from the respiratory organ to the metabolizing cells, is direct and with no admixture of venous blood. The carbon dioxide–rich blood increases the oxygen affinity of hemoglobin and therefore assists loading of oxygen in the lung, while the carbon dioxide–depleted blood with its reduced oxygen affinity enhances oxygen delivery in peripheral tissues. Among vertebrates, this scheme is present in water-breathing fish, birds, and mammals. Fish have a single circulation, with the gill perfusion in direct series with the perfusion that serves metabolism (systemic perfusion). Birds and mammals have a double circulation, with the pulmonary and systemic circulations connected in series by the two cardiac pumps, that is, the right and left ventricles. In vertebrate phylogeny, the transition from a single to a double circulation arose concomitantly with the transition from aquatic to aerial breathing. *See* CIRCULATION.

The hearts and vascular systems of lungfishes, amphibians, and reptiles are incompletely divided, providing for selective passage of blood in the pulmonary and systemic circulations. The regulated shunting of blood between these two circuits is associated with the phenomenon of periodic breathing. During breathing a nearly complete selective passage of blood occurs between the pulmonary and systemic circuits; but during the often long periods of breath-holding (apnea), blood is increasingly shunted from the pulmonary to the systemic circuit, and the circulation to the lungs is gradually bypassed. This pattern of alternating breathing and shunting ensures that blood circulation is adjusted to the oxygen availability at the gas-exchanging surfaces of the lungs. Such matching of blood perfusion and ventilation (that is, oxygen availability) is important for gas exchange efficiency in all animals in which the exchange depends on active ventilation and circulation.

The higher metabolic rate in birds and mammals accords with a higher pulmonary ventilation that stabilizes alveolar oxygen availability. A complete separation of the pulmonary and systemic circulation is therefore optimal for these animals. *See* CARDIOVASCULAR SYSTEM.

Diffusing capacity. Since surface areas and diffusion distances cannot readily be measured in live animals, gas exchange efficiency is often expressed by the diffusing capacity of the individual gases. For aquatic animals, diffusing capacity is often referred

to as the transfer factor. Both terms are defined as the amount of a gas (such as oxygen) transferred across the respiratory exchange surfaces per unit time and per unit of pressure difference across the gas exchange surfaces. The diffusion capacity of both gills and lungs are positively related to the rates of oxygen uptake of an animal. And this relationship holds whether the oxygen uptake level is set by the genetic or phylogenetic status of the animal, or is responding to behavioral patterns (such as physical activity) or to environmental factors (such as oxygen availability). *See* METABOLISM; RESPIRATORY SYSTEM.

Kjell Johansen; John B. West

Bibliography.   R. G. Crystal et al. (eds.), *The Lung: Scientific Foundations*, 2d ed. (2 vols.), Raven Press, 1997; E. R. Weibel, *The Pathway for Oxygen*, Harvard University Press, Cambridge, 1984; J. B. West, *Respiratory Physiology: The Essentials*, 7th ed., Lippincott Williams & Wilkins, Philadelphia, 2004.

Respirator

A device designed to protect the wearer from noxious gases, vapors, and aerosols or to supply oxygen or doses of medication to the wearer. Respirators are used widely in industry to protect workers against harmful atmospheres, and in the military to protect personnel against chemical or biological warfare agents. **Figure 1** shows the general classification of respirators according to whether they are atmosphere-supplying or air-purifying.

The hood, helmet, and facepiece provide the essential interface between the respirator and the wearer. Quarter-mask facepieces cover the wearer's nose and mouth above the chin; half-mask facepieces cover the nose and mouth and extend below the chin; and full facepieces cover the nose, mouth, and eyes. Facepieces made of elastomeric material or flexible fabric are equipped with a head harness or straps designed to secure the device to the wearer.

Important to the industrial application of respirators is a program that includes the monitoring of the respiratory hazard, physiological and psychological limitations of the wearers, selection of properly approved equipment, training of personnel, and equipment maintenance. Facepiece selection for a satisfactory fit is necessary to guarantee low face-seal leakage for negative-pressure respirators.

Atmosphere-supplying respirators. Atmosphere-supplying respirators are used in atmospheres deficient in oxygen or extremely hazardous to the health of the wearer. Such atmospheres can occur in unventilated cellars, wells, mines, burning buildings, and enclosures containing inert gas. In the United States the minimum legal requirement for such a respirator is a supply of oxygen of 19.5% or greater by volume to the wearer, compared with 21% existing naturally in pure air.

Self-contained breathing apparatus (SCBA). This is a completely self-contained unit with the air supply or the oxygen-generating material being carried by the wearer. It may be of the open-circuit (**Fig. 2**) or closed-circuit type. A closed-circuit SCBA (rebreather) involves recycling the exhaled air through sodium peroxide to remove the exhaled carbon dioxide, and adding oxygen to the air from an oxygen bottle.

Air-supplied respirators. Equipped with the same variety of facepieces as the SCBA, these respirators can have the air supplied to the facepiece by means of a hose and a blower—the hose mask—or from a compressed-air source equipped with proper air-flow and pressure-regulating equipment—the air-line mask. Helmet or hood air-line respirators can afford temperature control with an unencumbered wide-angle view. A highly specialized version of an air-line respirator is the medical respirator used for artificial respiration, or for the delivery of oxygen and medication. The aviator mask is another example.
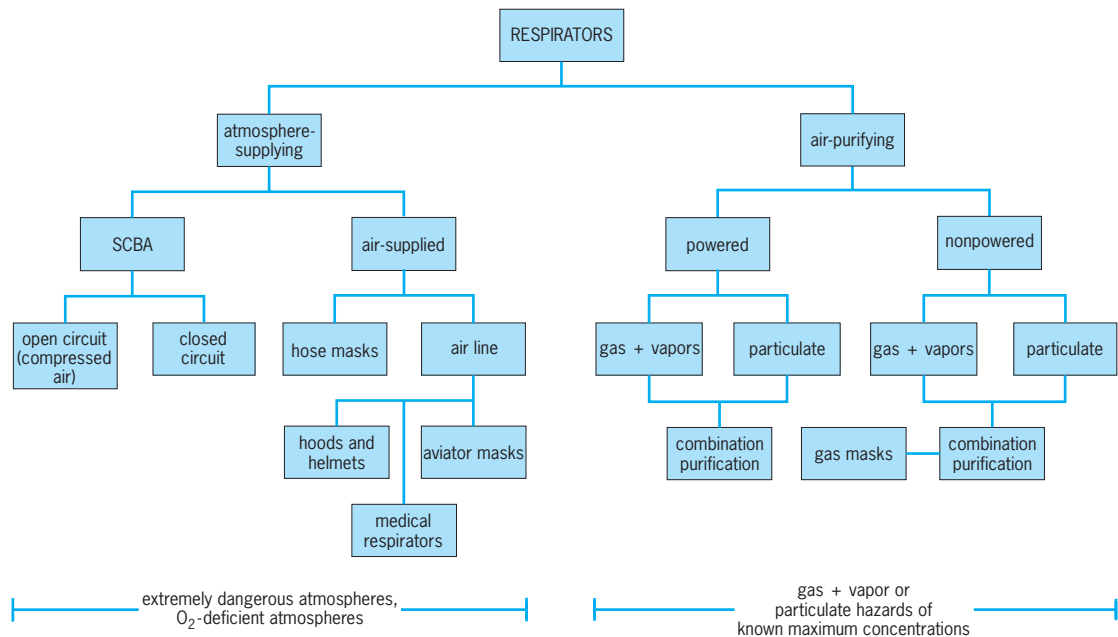

Fig. 1. **General classification of respirators according to design and application.**

Air-purifying respirators. In an air-purifying respirator, ambient air is passed through a purifying medium to remove the contaminants. However, these devices do not provide oxygen or protect against oxygen-deficient atmospheres.

Nonpowered. A widely used air-purifying respirator is the nonpowered, or negative-pressure, respirator (**Fig. 3**). Ambient air is inhaled through the purifying medium in the replaceable cartridges and exhaled through an exhaust valve. Sometimes the air is exhaled back through the medium if an exhaust valve is not provided. Because of the negative air pressure created inside the facepiece during inspiration, proper fit to the wearer is important to prevent air leakage through the face seal.

An increasingly popular form of nonpowered, air-purifying respirator is the disposable mask (**Fig. 4**). Here, the entire half-mask facepiece is made of the respirator filter material. It is light in weight and can be discarded after use.

Powered. In the case of the powered air-purifying respirator, an external blower, usually powered by a belt or helmet-mounted battery pack, forces air through the purifying medium and supplies it to the wearer under positive pressure, thus minimizing the problem of face-seal leakage. The purified air can be supplied to the wearer through either a hood or a helmet-face shield, which provides ample room for beards, eyewear, and hair styles that may otherwise cause face-seal leak problems in facepiece devices. While the design of powered respirators eliminates the inhalation airflow resistance of nonpowered respirators for easy breathing, powered respirators are larger and more complicated to use. Belt or helmet-mounted models provide added mobility over atmosphere-supplying respirators.

Particle removal. The filter medium used in a respirator for particle removal is usually selected according to the type of hazard against which the respirator is designed to protect.

Particulate hazards. These include dusts, mists and sprays, and fumes.

Dusts are solid particles that are mechanically produced by activities such as grinding, drilling, or sanding. Dusts range in size from submicrometer to tens of micrometers and can cause lung disease such as fibrosis (silica), chronic irritation (acids) and allergies (pollen, spices), and diseases from systemic poisons (cadmium) and carcinogens (asbestos, radioactive particles).

Mists and sprays are liquid particles produced by condensation (mists) or mechanical means (sprays). They may range in size from submicrometer to a few micrometers, and can cause respiratory illness such as major chemical irritation (smoke containing acidic condensate), systemic poisoning (spray paint), and chronic respiratory ailments (cigarette smoke, smog).

Fumes are solid particles produced by condensation, generally of metal or metal oxides with diameters in the submicrometer range. They can cause systemic poisoning (lead-welding fumes) or chills followed by fever (zinc or copper fumes).

Fig. 2. Open-circuit self-contained breathing apparatus (SCBA) atmosphere-supplying respirator.

Filters. Respirator air filters for particle removal fall into three general classes. Filters for dust or mist removal generally consist of a filter medium of moderate efficiency. They are made of fiber mats designed to provide particulate protection while giving low breathing resistance. A common dust/mist filter material is wool felt mechanically impregnated with modified phenolic resin to produce electrostatic capture sites. Another material is fine-fiber (about 2 μm in diameter) polyolefin web formed over a supporting facepiece shell.

Filters intended to remove dust, mist, and fume require a filter medium of moderate to high efficiency. Cartridge fume filters are made of fine-fiber matting or paper pleated to give a large extended filter area, since the loading of the filter with fume particles can cause rapid increases in breathing resistance. Noncartridge dust/mist/fume filters (Fig. 4) utilize
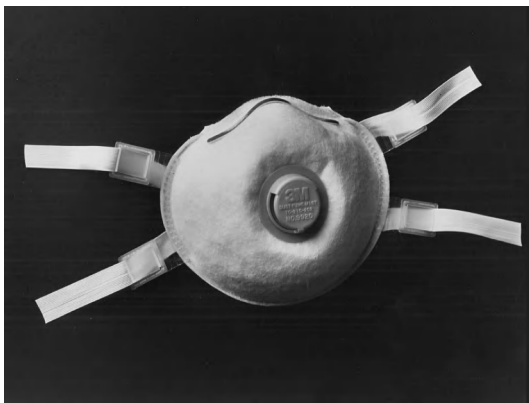

Fig. 3. Negative-pressure air-purifying respirator.

Fig. 4. Disposable negative-pressure air-purifying respirator.

a combination of very fine fibers and electrostatics in a very open matrix to provide high efficiency with low breathing resistance during particle loading on the filter.

High-efficiency particulate air (HEPA) filters are made of a special filter medium intended to remove nearly all airborne particles. The filters are generally made of very fine fibers, and the filter paper is pleated to provide a large extended surface area while giving an acceptable but moderately high breathing resistance. Such filters capture virtually all airborne particles and are used in nuclear installations or other critical applications and in the presence of extreme hazard. *See* AIR FILTER.

Gas and vapor removal. Media for gas and vapor removal are selected according to their service life performance against the specific type of gas or vapor hazard present. Gas and vapor hazards generally fall into the following major categories: (1) acidic (substances that are acids or react with water to produce acids, for example hydrogen chloride, fluorine, and hydrogen sulfide); (2) alkaline (substances that are alkalies or react with water to produce alkalies, for example ammonia, amines, and phosphenes); (3) organic (compounds of carbon such as hydrocarbons, alcohols, halides, isocyanates, and aromatics); (4) organometallic (metals chemically bonded to organic groups, for example tetraethyllead and organic phosphates); and (5) hydride (compounds in which hydrogen is chemically bonded to metals or metalloids, for example diborane and tetraborane).

Activated carbon in a granular form is used widely as an adsorbent to remove organic vapors. Vapors of high-molecular-weight substances are readily adsorbed onto the carbon surface, while low-molecular-weight gases or vapors, such as carbon monoxide, must be decomposed to be rendered harmless.

For the removal of acidic or alkaline gases, activated carbon is often modified to provide sites for a chemical reaction to occur with the gas. Metal oxides or caustic impregnation of the carbon can make it effective for removing acidic gases, while acidic impregnation of the carbon can make it effective against alkaline gases. Sometimes, activated carbon is simply mixed with a chemical adsorbent in a granular form to provide the necessary protection against the gas.

Gases or vapors that are not readily adsorbed either physically or chemically, such as carbon monoxide or formaldehyde, must be decomposed to render them harmless. This can be achieved by impregnating carbon with a catalyst such as a metallic salt or oxide. *See* ACTIVATED CARBON; ADSORPTION.

Cartridges and combination filters. To provide protection against both vapor and particles, activated carbon is often combined with a particulate filter to form a cartridge which can then be attached to the facepiece. Specially designed nonwoven facepiece respirators are available which use particulate filter media that are layered or loaded with gas- and vapor-adsorbent material for both vapor and particle removal.

Military gas masks. These are specialized devices which protect the wearer from the chemical, biological, and radioactive challenges possible in warfare. Some designs provide the wearer with tubes and other means to eat and drink without doffing the respirator. Included with the military mask is a storage bag or carrier that keeps the respirator and adsorbent media from contamination when not in use.   Benjamin Y. H. Liu; Daniel A. Japuntich

Bibliography.   American National Standards Institute, *Practice for Respiratory Protection*, ANSI788.2-1980, 1980; D. S. Blackwell and G. S. Rajhans, *Practical Guide to Respirator Usage in Industry*, 1988; Los Alamos Scientific Laboratory, *Guide to Industrial Respiratory Protection*, 1977; U.S. Nuclear Regulatory Commission, *Manual of Respiratory Protection Against Airborne Radioactive Materials*, 1976.

Respiratory pigments (invertebrate)

Colored, metal-containing proteins that combine reversibly with oxygen, found in the body fluids or tissues of multicellular invertebrate animals and microorganisms. The role of these pigments is primarily to aid in the transport of molecular oxygen. Thus they are distinguished from respiratory enzymes, which are concerned with the metabolic consumption of oxygen. Four distinctly colored groups of respiratory pigments exist among invertebrates: hemoglobins (purple, become orange-red with oxygen), chlorocruorins (green, become red with oxygen), hemocyanins (colorless, become blue with oxygen), and hemerythrins (colorless, become red with oxygen). Formerly, invertebrate hemoglobins were called erythrocruorins to distinguish them from functionally similar yet structurally distinct pigments of vertebrate bloods. Those hemoglobins confined to muscle cells are called myoglobins. *See* HEMOGLOBIN.

Each pigment is composed of two parts, a large protein molecule to which is bound one or more small moieties called prosthetic groups, each of which is or contains a metal. The metal binds the oxygen, and this binding imparts the characteristic

color to the pigment. In hemoglobins the prosthetic group is an iron porphyrin compound called heme. Chlorocruorin contains a similar iron porphyrin which differs from heme only in that a vinyl group in the molecule is replaced by formyl. The prosthetic group of hemerythrin consists of two adjacent iron atoms which bind an oxygen molecule between them. The prosthetic group of hemocyanin is analogous and consists of two adjacent copper atoms. Pigments containing vanadium have been found in tunicates, but these substances do not combine reversibly with oxygen and so cannot be considered respiratory pigments.

The protein part of the pigment confers reversibility upon the combination of the metal with oxygen. In the absence of protein, the prosthetic groups lose their capacity to combine with oxygen reversibly. Instead, the metals are irreversibly oxidized: Electrons are transferred from metal to oxygen. The bonds between metal and protein so alter the electronic energy levels of the metal that this transfer, if it occurs, is reversible. For this reason, the combination of hemoglobin with oxygen is described as oxygenation rather than oxidation. The protein is also responsible for certain physiological adaptations of the pigment to the environment. Thus the affinity of the pigment for oxygen is often highest in animals that inhabit environments with the lowest oxygen content.

Hemoglobins. Hemoglobin molecules of invertebrates are usually relatively small, with molecular weights of 65,000 or less, and enclosed in cells, or else they are extracellular and very large with molecular weights ranging from 400,000 (in *Daphnia*) to 8–12×10^6 (in *Cardita*, a marine bivalve mollusk). The size of the minimum functional subunit of many of these pigments has not been accurately determined, but the hemoglobin from the perienteric fluid of *Ascaris* has a molecular weight of about 41,000 and has only one heme. The very large hemoglobins are composed of a large number of subunits and may contain as many as 200 or more heme-containing protein subunits. The stability of these assemblages often depends on the presence of divalent cations such as calcium. The confining of the smaller molecules to cells may prevent their loss and may also provide a local ionic environment favorable to their stability. It is also possible that the large hemoglobin molecules in free solution in the body fluids of various marine animals help to maintain an osmotic balance of salts. *See* ANNELIDA.

Distribution and function. Hemoglobins are widely distributed among invertebrates as well as among almost all vertebrates. Those pigments found in the body fluids, whether in cells or in free solution in the plasma, function primarily as agents that transport molecular oxygen to the tissues. Two barriers exist to this transport: One is at the surface of the animal (skin and gill membranes) and the other is within the tissue cells, where the oxygen must move from the cell surface to the respiratory enzymes of the mitochondria that consume the oxygen. Transport across these barriers depends on the physical process of diffusion, which is extremely slow and often is entirely inadequate by itself to supply the metabolic needs for oxygen. However, if hemoglobin is present within these barriers, diffusion of oxygen is greatly facilitated because the total capacity of the cell for oxygen is much greater. Hemoglobins that function in this way usually have high affinities compared to the affinities of blood or fluid hemoglobins. Myoglobins apparently work in this way: They pick up oxygen at the muscle cell surface and facilitate its diffusion to the mitochondria adjacent to the contractile muscle fibers. Cellular hemoglobin is found in the brain of the predatory nemertean worm *Cerebratulus*, where it acts as an oxygen store to enable the worm to invade anoxic marine muds for short periods of time in search of other invertebrates. *See* MITOCHONDRIA.

Oxygen-carrying capacity. The oxygen-carrying capacity of blood is greatly increased if it contains hemoglobin. In the absence of any pigment the maximum amount of oxygen which 100 cm³ of blood can physically dissolve is only about 0.3–0.5 cm³. The blood of the annelid worm *Arenicola* has a capacity of up to 10 cm³ of oxygen per 100 cm³ of blood. Thus the hemoglobin accounts for at least 95% of the oxygen carried by the blood of this animal. If a blood pigment is to be efficient as a transport agent, it must not only bind oxygen but be able to release it at sufficiently high pressures to meet the needs of the animal. The higher the oxygen pressure at which the pigment unloads its oxygen, the better will be the supply of oxygen in the tissue capillaries. The more active a particular species of animal, the greater, in general, is the oxygen pressure at which its blood unloads its oxygen. An increased temperature lowers the affinity with which all blood pigments bind oxygen. Thus an increased temperature not only accelerates metabolism generally but also tends to make more oxygen available by raising the pressure at which the hemoglobin delivers oxygen to the tissues.

Many hemoglobins aid in the transport of carbon dioxide as well as oxygen. Vertebrate blood hemoglobins have a lower oxygen affinity in the presence of carbon dioxide than in its absence. This phenomenon, found in many invertebrates, is called the Bohr effect after its discoverer, the physiologist, Christian Bohr, the father of the atomic physicist Niels Bohr. The effect facilitates the unloading of oxygen in the tissue capillaries, and the discharge of carbon dioxide at the gills or other gas-exchanging surfaces. A pronounced Bohr effect is absent from most high-affinity hemoglobins such as myoglobin, which appear to function by facilitating diffusion.

The distribution of hemoglobins among invertebrate animals does not fit any phylogenetic scheme. It appears as if hemoglobins have arisen many times in the course of evolution. In some species the hemoglobins appear to be essential for the animal's survival; in others the advantages appear marginal at best, and perhaps their occurrence is the fortuitous result of the evolution of an enzyme whose primary function is quite different from the more familiar hemoglobins. This supposition is strengthened

by the discovery that hemoglobin-like proteins have now been found in virtually all cells. Furthermore, hemoglobin-like molecules occur that have spectra resembling those of certain cytochromes (b_5). Strong evidence for this possibility comes from the discovery that cytochrome b_5 (calf liver) is homologous in its primary structure (sequence of amino acids) to that of vertebrate hemoglobin.

Multicellular organisms. Hemoglobins are distributed in every major phylum of invertebrates except those of the sponges and cnidarians. The pigments occur in the parenchyma of the flatworm *Phaeno-cora* and in the blood cells of a few nemertean worms. Hemoglobins occur in several parasitic nematodes, including *Ascaris, Nippostrongylus*, and *Strongylus*. Each pigment possesses an extremely high affinity for oxygen and dissociates its oxygen very slowly. Thus, it takes about 150 s to deoxygenate the hemoglobin of the perienteric fluid of *Ascaris*; under the same conditions the dissociation time of sheep hemoglobin is 0.008 s. *Strongylus* hemoglobin binds oxygen so tightly that the worms die of anoxia when placed in an anaerobic environment before their hemoglobin loses its oxygen. This fact suggests that the function of the hemoglobin might be catalytic and not to transport oxygen. In contrast, the hemoglobin in the body wall of *Ascaris* may serve a transport or diffusion function, because all body motion stops when the deoxygenation of the pigment starts. *See* NEMATA; PLATYHELMINTHES.

Among mollusks, hemoglobin occurs chiefly in the gastropods, but it has been found in a few bivalved mollusks. It occurs in the radular muscle of the whelk *Busycon*, whose blood pigment is hemocyanin. This muscle controls the radula, an organ in the buccal cavity that is used for scraping and gathering of food. The hemoglobin probably facilitates oxygen supply to this muscle. *See* MOLLUSCA.

Hemoglobin is found in the ctenidia (gills) of the bivalve mollusk *Phacoides*, which lives in the mud of mangrove swamps. The pigment may facilitate transport of oxygen across the gill to the associated blood vessels. It has a very high oxygen affinity; an oxygen pressure of only 0.007 in. Hg (0.19 mmHg or 25 pascals) is sufficient to half-saturate the pigment with oxygen. In contrast, the hemoglobin of the blood of another bivalve, *Anadara*, requires an oxygen pressure of 10 mmHg to achieve the same degree of oxygen binding. This 50-fold difference in oxygen affinity may be related to the fact that *Phacoides* dwells deep in the hypoxic mud, whereas *Anadara* lives on the surface. Hemoglobin occurs in the blood of many annelid worms. In those worms with a closed circulation the hemoglobin is usually found dissolved in the plasma; in those with less well-developed circulatory systems the hemoglobin is usually in cells confined to the coelomic fluid. Hemoglobin apparently occurs both in the plasma and in the cells of the coelomic fluid of two annelids, *Terebella* and *Travisia*.

Hemoglobins occur in the blood plasma of several crustaceans, including certain species of the water flea, *Daphnia*. Its presence in *Daphnia* depends on the oxygen content of the water: the lower the oxygen content, the greater the quantity of hemoglobin synthesized. The mechanism of this relationship is not known but may depend on the fact that one of the steps in the synthesis of heme from δ-aminolevulinic acid is strongly dependent upon oxygen, and that the synthesis of hemoglobin protein is apparently controlled in part by the quantity of heme available. The sea cucumber *Thyone*, an echinoderm, contains hemoglobin in distinct blood cells. *See* CRUSTACEA.

Hemoglobin has been found in only a few adult insects, all of which are aquatic, and in two larval insects, each of which develops in an low-oxygen aquatic environment. The tracheal tubes of the back swimmers *Buenoa* and *Anisops* are surrounded by clusters of cells that contain hemoglobin. It occurs in the tracheal cells of the larva of the botfly *Gastrophilus*, which is found in the stomach of horses. It also occurs in the water boatman *Macrocorixa* and in the body fluid of the larva of the midge *Chironomus*. No hemoglobin has been discovered in any protochordate. *See* INSECTA.

Curiously, hemoglobins have been found in the nerve tissues of a wide variety of invertebrate animals, including nematode and annelid worms, mollusks, and crustaceans. Indeed, the nerve ganglia of the annelid *Aphrodite* is bright crimson. When the hemoglobins in the tissues of certain worms and mollusks are deprived of oxygen, they exhibit a peculiar optical absorption spectrum quite different from that of deoxyhemoglobin. These spectra suggest the possible presence of ferryl hemoglobin, in which the oxidation state of the iron is 4+. The hemoglobins may serve a catalytic function, as has been suggested for similar pigments in microorganisms. A hemoglobin-like pigment occurs in the ganglia of the mollusk *Spisula*, which when deoxygenated exhibits an absorption spectrum characteristic of cytochrome b_5; the absorption spectrum of the oxygenated pigment resembles that of oxyhemoglobin. *See* HEMOGLOBIN.

Microorganisms. Hemoglobins and hemoglobin-like pigments are also found in microorganisms, being widely distributed in Protozoa, Fungi, and bacteria. The root nodules that fix molecular nitrogen in leguminous plants are pink as a result of the presence of hemoglobin. Both the fixation process and the occurrence of hemoglobin require the presence of the symbiotic bacterium *Rhizobium*. The hemoglobin is present only during periods of active fixation, when it appears to sequester oxygen for delivery to mitochondria and to protect the nitrogen fixation enzymes from the toxic effects of oxygen. Yeast and many bacteria contain flavohemoglobins, which have both heme- and flavin-containing parts. These proteins are enzymes that catalyze the oxygen-dependent conversion of nitric oxide to nitrate. *See* NITROGEN CYCLE.

Hemocyanins. Hemocyanins are found only in mollusks and in arthropods other than insects. However, they are not the only copper proteins that combine reversibly with oxygen. Ascorbic acid oxidase, an enzyme found in plants, has about the same copper content. In the presence of oxygen this enzyme is blue,
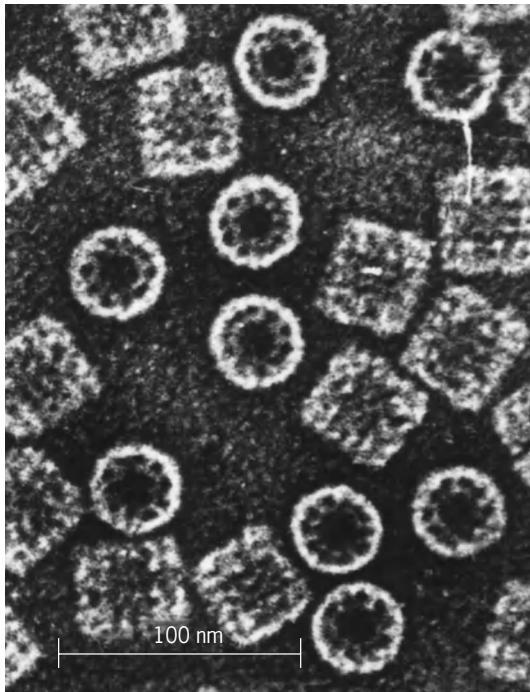
Electron micrograph of hemocyanin molecules from the snail *Helix pomatia*, showing "pillbox" subunit structure. Circular forms show subunits in a tenfold symmetrical array. There are six parallel rows of subunits in the rectangular forms.

as is hemocyanin, and when oxygen is removed, the pigment becomes colorless, as does hemocyanin. Hemocyanin itself has no oxidase activity. Hemocyanins always occur in the plasma, never in cells. The molecular weights are very high, ranging up to 7×10^6. They are sufficiently large to be seen readily under the electron microscope, where they appear as pillbox-like aggregates of many subunits (see **illustration**).

This pigment is found in many of the higher crustaceans such as crabs and lobsters, in the horseshoe crab *Limulus*, and in scorpions. Among the mollusks it occurs both in the cephalopods (squid and octopus) and in certain gastropods. The squid is an active carnivore and requires well-aerated seawater for survival. The blood of both the squid and the octopus has about the same oxygen capacity; yet the two hemocyanins have very different properties, the hemocyanin of the squid being half-saturated at an oxygen pressure of only 3 mmHg (400 Pa). This large difference reflects the fact that the squid is a more active animal than the octopus. About 92% of the oxygen of the arterial blood of the squid is normally removed during circulation through the tissues; in humans no more than about 30% is so removed. Thus the squid is poorly adapted to survive even very short periods in an environment with little oxygen. *Limulus*, on the other hand, survives for weeks with all of its hemocyanin removed.

Hemocyanins are of two distinct classes. Those from crustaceans have a minimum functional subunit of 68,000–70,000 (the unit combining with one molecule of oxygen and containing two atoms of copper). In contrast, the analytical data for molluscan hemocyanins indicate a minimum unit of about 50,000, although no functional subunit of this size has yet been prepared. Divalent cations strongly influence both the aggregation of subunits and the oxygen equilibrium; their removal from crayfish hemocyanin results in a reversible decrease in oxygen affinity. Lobster (*Homarus*) hemocyanin has a molecular weight of about 825,000 and probably consists of 12 subunits.

Chlorocruorin. Chlorocruorin has properties very similar to those of hemoglobin. It is restricted to certain sessile marine annelids. One species of the genus *Spirorbis* contains chlorocruorin in its blood, while another has hemoglobin. A third has no pigment at all. Another closely related worm, *Serpula*, has both chlorocruorin and hemoglobin dissolved in the blood. This is the only animal known to have two blood pigments. No functional reason for these differences has been suggested; each of these worms lives in a similar environment. The oxygen affinity of chlorocruorins is much lower than that of most hemoglobins.

Hemerythrin. Hemerythrin is similarly restricted and is found in a few annelid and sipunculid worms and in the brachiopod *Lingula*. It is found only in cells. The hemerythrin of the sipunculid worm *Golfingia gouldii* consists of eight identical subunits that form a molecule of 107,000. Each subunit is composed of 112 amino acids and contains two atoms of iron. The oxygen is believed to be bound between each pair of iron atoms. The subunits are held together by noncovalent bonds. *See* RESPIRATION.

Austen F. Riggs

Bibliography.   P. W. Hochachka and G. N. Somero, *Biochemical Adaptation*, 2002; C. L. Prosser, *Adaptional Biology: Molecules to Organisms*, 1986; E. E. Ruppert and R. D. Barnes, *Invertebrate Zoology*, 1993; R. E. Weber and S. N. Vinogradov, *Physiolog. Rev.*, 81:569–628, 2001.

Respiratory syncytial virus

A virus belonging to the Paramyxoviridae, genus *Pneumovirus*. This virus, although unrelated to any other known respiratory disease agent and differing from the parainfluenza viruses in a number of important characteristics, has been associated with a large proportion of respiratory illnesses in very young children, particularly bronchiolitis and pneumonia. It appears to be one of the major causes of these serious illnesses of infants. It is the only respiratory virus that occurs with its greatest frequency in infants in their first 6 months of life. In older infants and children, a milder illness is produced. Respiratory syncytial virus infection in adult volunteers gave rise to a coldlike illness; adult infection took place readily, even though the adults had moderate to high levels of antibody.

The respiratory syncytial virus is 90–140 nanometers in size. It shares with measles, mumps, and parainfluenza 2 the property of producing a

characteristic syncytial effect (pseudogiant cell) in cell cultures.

The clinical disease in young infants may be the result of an antigen-antibody reaction that occurs when the infecting virus meets antibody transmitted from the mother. For this reason respiratory syncytial vaccines that stimulate production of antibodies in the serum but not in the nasal secretions may do more harm than good. Efforts to develop an attenuated vaccine that infects subclinically and produces nasal antibody are encouraging. *See* ANIMAL VIRUS; VIRUS CLASSIFICATION.   Joseph L. Melnick; M. E. Reichmann

Bibliography. R. M. Chanock, M. A. Mufson, and K. M. Johnson (eds.), Comparative biology and ecology of human virus and mycoplasma respiratory pathogens, *Progr. Med. Virol.*, 7:208–272, 1965; A. S. Evans (ed.), *Viral Infections of Humans*: *Epidemiology and Control*, 4th ed., 1997.

Respiratory system

The system of organs involved in the acquisition of oxygen and the elimination of carbon dioxide by an organism. The lungs and gills are the two most important structures of vertebrates involved in the phase known as external respiration, or gaseous exchanges, between the blood and environment. Internal respiration refers to the gaseous exchanges which occur between the blood and cells. Certain other structures in some species of vertebrates serve as respiratory organs; among these are the integument or skin of fishes and amphibians. The moist, highly vascular skin of anuran amphibians is important in respiration. Certain species of fishes have a vascular rectum which is utilized as a respiratory structure, water being taken in and ejected regularly by the animal. Saclike cloacal structures occur in some aquatic species of turtles. These are vascular and are intermittently filled with, and emptied of, water. It is thought that they may function in respiration. During embryonic life the yolk sac and allantois are important respiratory organs in certain vertebrates.

Structurally, respiratory organs usually present a vascular surface that is sufficiently extensive to provide an adequate area of absorption for gaseous exchange. This surface is moist and thin enough to allow for the passage of gases. This article treats the embryology and physiology of the gills and embryology, anatomy, and histology of the vertebrate lung and vertebrate respiratory physiology. *See* ALLANTOIS; HEMOGLOBIN; LUNG; YOLK SAC.

Gill Development and Anatomy

In the most primitive chordates, the Acrania, the gills are really part of the digestive, rather than the respiratory, system. However, in all these animals, except one genus, *Rhabdopleura*, there is a pharyngeal region in which there are one or more (usually many) paired visceral clefts with delicate intervening strands of tissue which constitute the primary bars or arches; these are greatly augmented in some cases by the down growth of tongues of tissue between them, which are termed secondary bars. There are also transverse connecting strands termed synapticulae in some species, such as *Branchiostoma*. The bars and synapticulae usually contain supporting gelatinous or chitinoid rods; in *Branchiostoma* the primary bars contain prolongations of the dorsal coelom. Most important, blood circulates through the bars and synapticulae during respiration and is aerated by water drawn in through the oral opening by the ciliated endoderm and passed out through the interstices of the basketlike structure. This arrangement acts primarily as a strainer to retain food particles (**Fig. 1**).

The gills of chordates are located in the pharyngeal region. They are always related to the visceral arches and clefts. The latter structures are so defined because, in vertebrates that respire entirely with lungs, the arches and clefts, or incipient clefts, lack gills, which exist only in the embryo and disappear or are highly modified in the adult. They always arise by the outpushing of endodermal pouches, which are usually dorsoventrally elongated, through the mesoderm, which contact corresponding inpushings of ectoderm. Where these meet and break through, clefts are formed, and the intervening concentrations of mesoderm form the arches, which are covered internally by endoderm and externally by ectoderm. The number of clefts, the eventual content of the arches, and the extent to which the latter develop gills of one type or another are indicated in the following discussion.

Visceral arches in craniates. In this group the typical number of paired visceral arches and pouches or clefts is six, including the mandibular and hyoid arches with the intervening spiracle. However, the number may be greater in some elasmobranchs and cyclostomes, and less in some animals where these structures occur only in the embryos. In the latter cases the number of pouches which actually become clefts is usually very limited; for example, there are three pairs in the chick, one pair in the cow, and
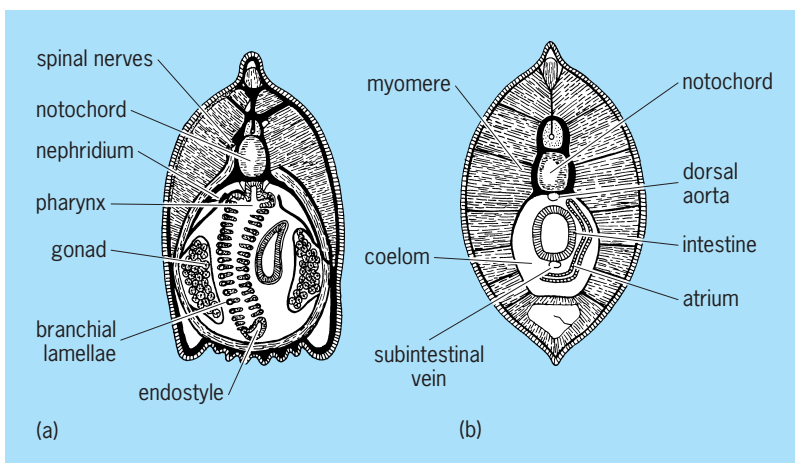

Fig. 1. *Branchiostoma lanceolatus*. (*a*) Transverse section of the pharyngeal region. (*b*) Transverse section of the intestinal region. (*After T. J. Parker and W. A. Haswell, A Textbook of Zoology, 6th ed., Macmillan, 1956*)
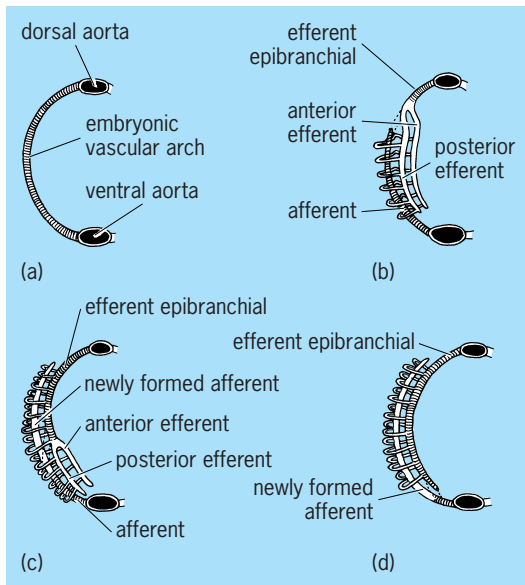
Fig. 2. Diagrams of the development of adult branchial vessels in various fishes. (*a*) Original continuous embryonic arch cross-lined. (*b*) Selachian. (*c*) Intermediate form, such as *Acipenser*. (*d*) Teleost: newly developed vessels, white. In *b*, *c*, and *d* the original arch is interrupted. (*After E. S. Goodrich, Studies on the Structure and Development of Vertebrates, Macmillan, 1930*)

normally none in the pig and human, in which the number of arches is reduced to five, with the last vestigial. The blood vessels found in the arches in animals other than fish are of considerable phylogenetic interest. Thus the vessels in the third visceral arches become the roots of the internal carotids; those of the fourth form a pair of aortae in amphibians and reptiles, or one of this pair an aorta in birds and mammals; and those in the fifth pair give rise to the pulmonary arteries. The pulmonary arteries are really the sixth pair of aortic vessels because a vestigial fifth pair occurs, together with the sixth, in the fifth visceral arches.

In fishes the aorta in each of the arches which bear gills, commonly the posterior four pairs, becomes altered and augmented to form the afferent and efferent vessels of the gills of that arch. The precise method of alteration and final condition vary somewhat in different groups of fishes.

In the elasmobranchs the ventral two-thirds of the original aortic arch persists as the afferent vessel. This becomes disconnected from the dorsal third, from which two vessels grow ventrally. These and the dorsal third thus become the efferent aortic arteries and are connected with the afferent vessel through gill capillaries in a manner to be described below (**Fig. 2***b*). The more anterior of the two growing vessels drains the posterior half of the gill pouch preceding it, while the posterior half drains the anterior half of the gill pouch following. In each arch the two efferent vessels have transverse connections midway between them, and all the efferent vessels on a side are eventually united by dorsal and ventral vessels. The teleost condition represents essentially the reverse of what occurs in the elasmobranchs with respect to the origin of the afferent and efferent vessels. Thus

in the teleost it is a new upgrowth from the base of the original arch which forms the afferent vessel, whereas most of the original aortic arch forms the efferent vessel. Teleosts also have only one efferent vessel in each visceral arch (Fig. 2*d*). An intermediate form (Fig. 2*c*) with respect to the origins and character of these vessels occurs in *Acipenser*, *Amia*, and *Lepisosteus*, although in detail these forms are not precisely as shown in the diagram.

Branchial gills.   The term branchial gills is used because there are other gills and structures functioning as gills, which are not in the pharyngeal region. There are generally considered to be two kinds of branchial gills, external and internal. Because the actinopterygian internal gill is apparently derived from the elasmobranch type, the elasmobranch gill is discussed first.

Elasmobranch gill.   In these fishes, there develops within each branchial arch a delicate cartilaginous bar slightly proximal to the vessels, the median afferent and two lateral efferent vessels, which have already been described. The arches also become supplied by branches of the ninth and tenth cranial nerves (the hyoid by the seventh) and acquire muscle fibers derived from the walls of coelomic extensions by which the arches are temporarily invaded. From the outer side of each arch there grows along its dorsoventral extent a thickish sheet of tissue which develops along its border two indentations or clefts, tending to divide the single sheet into anterior, median, and posterior layers. The indentations never become very deep, however, so that these layers are essentially one. The middle layer or part of this sheet becomes the septum, consisting of connective tissue and a little muscle and covered, where it is free, with epithelium. The anterior and posterior layers on each side of the septum become half gills or hemibranchs, and the whole structure a holobranch (**Fig. 3**).
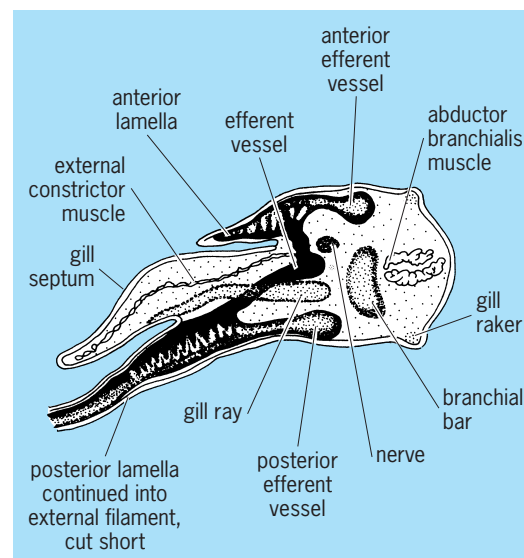

Fig. 3.   Section across gill-bar of *Scyllium canicula* late embryo 1.3 in. (32 mm) long, showing blood supply to lamellae. (*After E. S. Goodrich, Studies on the Structure and Development of Vertebrates, Macmillan, 1930*)
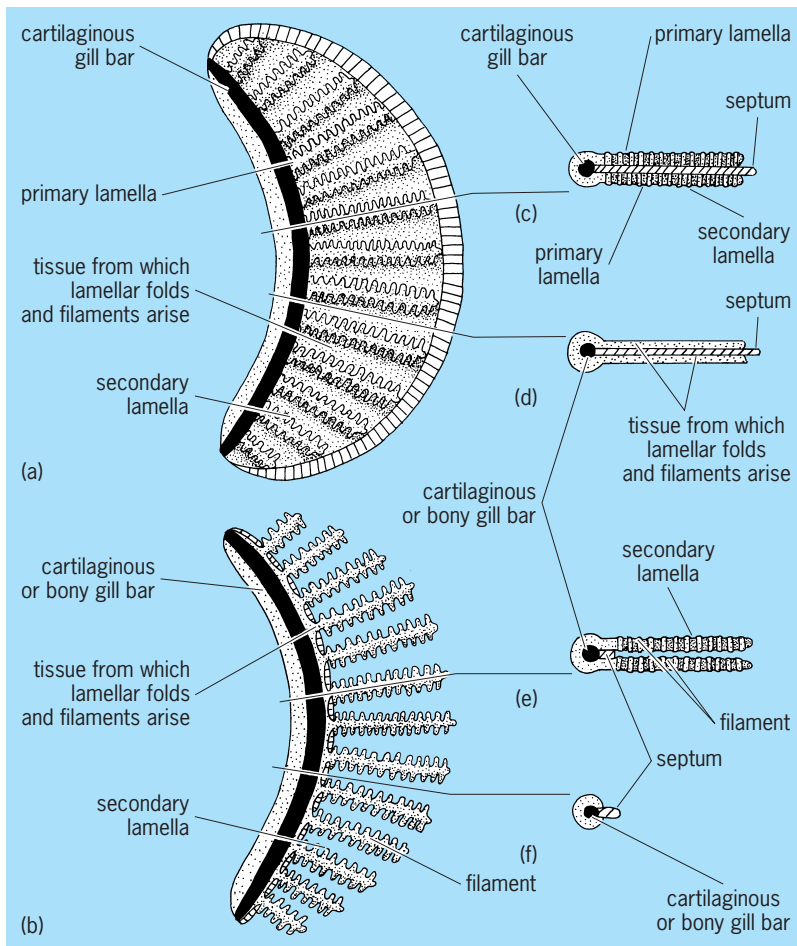
Fig. 4. Diagrams of elasmobranch and teleost gill structure. (*a*) Gill arch and gill (hemibranch) from the right side of the head of an elasmobranch; posterior view. (*b*) Same view of teleost arch and gill. (*c*) Horizontal section along the length of a pair of primary lamellae and the septum of the holobranch of an elasmobranch. (*d*) Horizontal section through three tissue layers between primary lamellae of an elasmobranch. (*e*) Horizontal section along the length of a pair of filaments of a teleost holobranch. (*f*) Horizontal section between the filaments of a teleost holobranch, hence through the arch and remains of the septum only.

branchs the primary lamellae of the posterior hemibranch grow outward as filaments beyond the external openings of the gill clefts (Fig. 3). These temporary filaments are therefore essentially external gills, although in both origin and structure they are unlike the "true" external gills, described below. Because they occur only in the unhatched fishes, where they float in the albuminous fluid of the egg, they are probably as much for absorption of food as respiration. The septum of each holobranch grows out beyond the gill in these fishes and turns posteriorly beneath the outer skin to the posterior edge of the following cleft, thus forming for the latter an individual cover analogous to the operculum. Each branchial arch develops an anterior and posterior row of papillae, sometimes covered with enamel, on its internal border. These are the gill rakers, which act as strainers to prevent the escape of food (Fig. 3).

Holocephalan gills. In the subclass Holocephali the gills are similar to those in the elasmobranchs, although arranged slightly differently with respect to the arches, and are covered by an operculum.

Cyclostome gill. Although this class is considered the most primitive of all the craniates, its consideration has been delayed because the gills present a highly specialized condition most easily described as a variant of the elasmobranch type. The early development is similar to that in elasmobranchs. The septa are soon drawn out, however, so that the clefts become elongated tubes that sometimes open separately, as in *Bdellostoma*, or unite to open through a single orifice, as in *Myxine*. The gills proper have primary and secondary lamellae, as in the elasmobranchs, and occur as hemibranchs on the sides of the septa near the pharynx where the tubes become enlarged into spherical pouches. The pouches open directly either to the pharynx or through very short ducts. The pharyngeal region becomes divided during development in the lampreys into a dorsal

Along the surface of each anterior and posterior layer there arises a series of transverse folds of tissue extending outward from the arch (**Fig. 4**). These are the primary lamellae covered by an epithelium, flat for the most part but with some columnar or cuboidal secretory cells. Between the epithelial cells there is connective tissue which contains, within each lamella, a loop from the afferent and adjacent efferent vessel of the arch. Another series of folds transverse to these lamellae develops next on the dorsal and ventral surface of each primary lamella. They are the secondary lamellae, whose internal surfaces are connected by numerous columnar cells which send out delicate protoplasmic strands beneath the epithelium of the folds. These cells are termed pilaster cells, and a network of capillaries runs among them connecting the afferent and efferent loops in the respective lamella (**Fig. 5**). A cartilaginous ray extends outward along the base of each primary lamella at the boundary between the posterior layer or hemibranch and the septum and helps to support the entire gill (Figs. 3 and 4*c*). In unhatched elasmo-
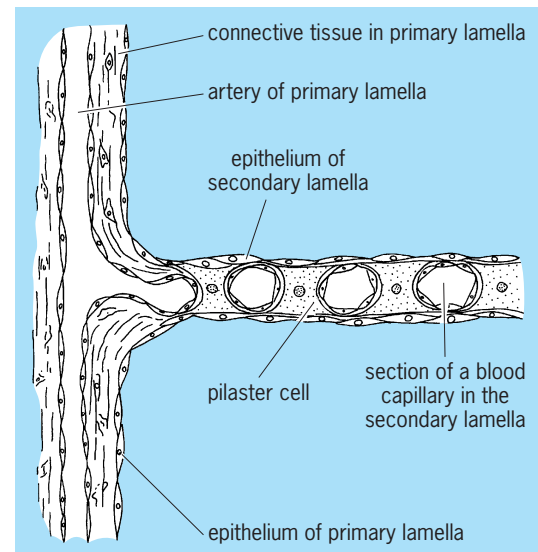

Fig. 5. Diagram of a section through a part of a secondary lamella of the gill of an elasmobranch fish, showing connection of the secondary lamella with a primary lamella.
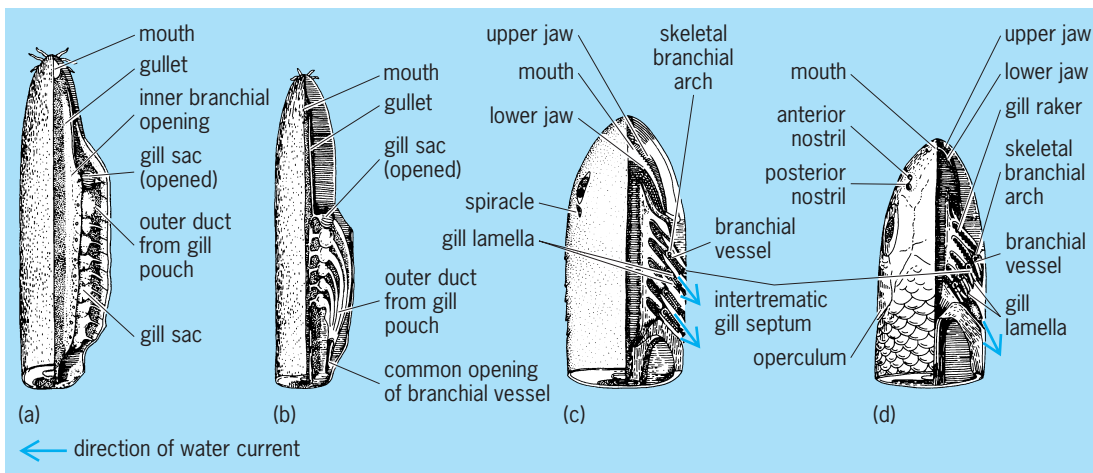
Fig. 6. Diagrams of head and gill region in (*a*) *Bdellostoma*, (*b*) *Myxine*, (*c*) a selachian, and (*d*) a teleost. Dorsal view partly dissected to show arrangement of gills. (*After E. S. Goodrich, Studies on the Structure and Development of Vertebrates, Macmillan, 1930*)

esophageal portion and a ventral respiratory tube with a blind posterior ending. It is this ventral portion which gives rise to the gill slits. The number of slits, eventually pouches and ducts, varies greatly in this class, ranging from 6 to 14 in different species of *Bdellostoma* (**Fig. 6**).

Actinopterygian gill. With a few exceptions to be noted below, it is possible to describe the gill situation in most bony fishes by indicating the ways in which it differs from that of the elasmobranchs. Although the holobranchs start to develop from the four branchial arches, as in the latter group, actinopterygians generally lack an open spiracle and the hemibranch related to it; instead, there is often a so-called pseudobranch which is probably glandular. Usually no temporary filamentous external gills occur. The septum grows out only a very short distance, and the anterior and posterior hemibranchs continue to grow out, not as sheets, but as numerous free filaments, corresponding to the primary lamellar folds (Fig. 4b, e, f). The secondary lamellae develop from these in the same way and with the same histological characters, as in elasmobranchs. Although there is only a single efferent artery in the arches of most bony fishes (Fig. 2d), the circulation in the filaments and secondary lamellae is similar to that in the primary and secondary lamellae of the former group. Another difference is that instead of a single ray extending between each primary lamella of the posterior hemibranch and the septum there is a ray in each pair of filaments of a holobranch (Fig. 4b, e). Lastly there is a single operculum on each side, attached anteriorly to the hyoid arch and covering all the gills. Each operculum is a sheet of tissue in which are embedded three flat bones, the operculars. Gill rakers are present in one or more (often two) rows, sometimes supported by ossifications (Fig. 6d).

There are a few exceptions to the situation just described. The gills of *Acipenser* are about halfway between those of elasmobranchs and typical actinopterygians, and the gill filaments of the lophobranchs consist of tufted processes.

Dipnoan gill. The Dipnoi may be considered as a sort of connecting link between the fishes and the amphibians. In Dipnoi there are both internal and apparently true external gills. These latter gills exist in the embryos of all the Dipnoi, and vestiges of them persist in the adult where they are attached to the last three pairs of arches. The internal gills in this group are reduced in correlation with the accessory respiration furnished by the lung or lungs. The septa are somewhat diminished, causing the primary lamellae to become partly filamentous, but without rays. There may be hemibranchs on the hyoids, on the last pair of branchial arches, or on both; there are thus two to four pairs of holobranchs.

Other respiratory devices of fishes. In addition to pharyngeal gills of the types indicated, there are other aquatic respiratory mechanisms, which sometimes occur in quite different locations. One such organ, still associated with the pharynx, is found in *Anabas*, the climbing perch; in this fish some of the pharyngeal bones are developed into folded plates covered with vascular epithelium, through which respiration can occur. This arrangement is covered by the operculum and can be kept moist for extended periods out of water. In one of the phyostomes, *Amphinous*, another pharyngeal derivative acting as a gill consists of vascular sacs opening out through the spiracles. Another peculiar structure used for respiration is the tail fin of an Indian Ocean fish which is immersed while the fish basks out of water. One of the Dipnoi, *Lepidosiren*, despite its lungs, apparently develops respiratory filaments on the paired fins. One of the siluroids employs rectal respiration by sucking in and expelling water from the anus.

Amphibian gills. Almost all amphibians have gills at some time during development and may retain them as adults. At least part of the time these are of the true external type as compared with the filamentous variety previously mentioned. Although some authorities consider that the differences between the internal gills and what are termed true external gills are not very significant, it seems that in this class of
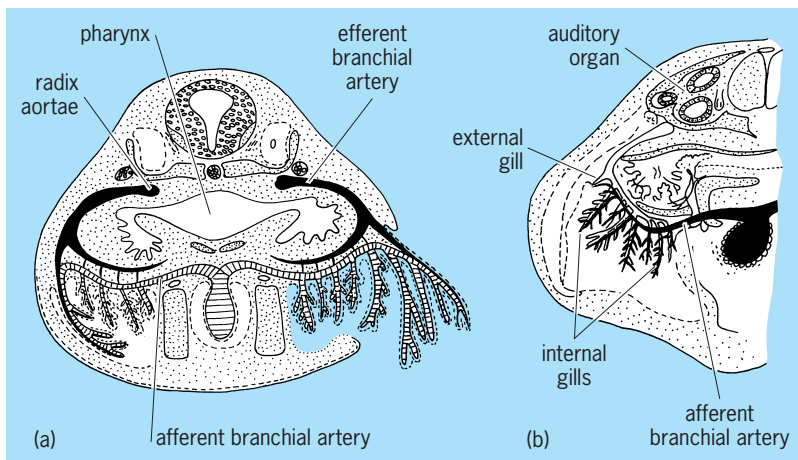
Fig. 7. External gills. (*a*) Relations with internal gills; transverse section. (*b*) External gill development. (*After E. S. Goodrich, Studies on the Structure and Development of Vertebrates, Macmillan, 1930*)

animals these differences are sufficient to demand some notice. Generally speaking, true external gills of amphibians may be distinguished from the gills, internal or external, already discussed in that they arise from the outer borders of the gill arches rather than from the sides and differ in certain details of structure. They are sometimes covered by an operculum.

Each gill of the type under consideration is composed of a rather heavy main stem, or rachis, which arises first. This is much thicker than the filaments of real internal gills and has been compared to the modified septum of such gills, which is otherwise entirely lacking in the variety being described. In those amphibians in which external gills are entirely larval, this rachis usually gives rise to rather blunt, short, fingerlike processes. The epithelium of such gills is ciliated, and within the rachis, at least, are muscle fibers so that the gills can be moved. The rachis also contains an arterial loop with extensions into each branch (**Fig. 7**). Gills of this type, which are typically found in anurans, such as the frog, spring from the upper parts of the first three pairs of posthyoid branchial arches; those from the first pair are the most prominent, overlapping and concealing the other two pairs.

These external gills in the anurans and many urodeles soon begin to be absorbed and are covered by back-growths of tissue from the hyoid arches which are continuous across the ventral side of the throat. This tissue contains no cartilaginous or bony plates and constitutes an operculum which fuses with the body wall everywhere posterior to the gill slits, except for one or two small openings termed spiracles through which water from the clefts leaves the branchial chamber. These spiracles are not homologous with those in fish previously mentioned.

Meanwhile, as the operculum is completed, the original external gills are absorbed, and in the Anura double rows of outgrowths on the borders of the first three pairs of branchial arches and a single row on the anterior border of the fourth are developed in their stead. These processes are quite similar to the processes which arose from the main rachis of the external gills, except they are shorter. They are sometimes termed internal gills, chiefly because they are covered by an operculum, but they are still not highly filamentous and lack any part of a septum. Hence, they may be regarded as reduced external gills which are covered (Fig. 7). There are, nevertheless, parts connected with these gills which appear to be closely related to similar parts of the internal gills of fish. The inner margins of the gill arches bear double rows of papillae which correspond, at least in function (that is, acting as strainers), to the gill rakers of the former group.

In certain of the urodeles, such as *Necturus*, the original gills are never covered or absorbed, but persist throughout life as actual external structures. In these cases, however, they differ somewhat from the larval external gills just described. The rachis is larger and heavier, and the secondary branches give rise to numerous fine, short, tertiary outgrowths, giving the whole structure a bushy effect (**Fig. 8**). Although apparently unnecessary, some of the clefts between these sets of gills remain open, for example, one pair in *Pseudobranchus*, two pairs in *Necturus*, and three in *Siren*. This retention of the gills and clefts in such forms has been regarded as an example of neoteny, that is, the persistence of larval characters in sexually mature animals. *See* NEOTENY.

Other amphibian respiratory devices. In the amphibians as in fish there are various peculiar structures which act as gills. Thus in the marsupial frog the two anterior pairs of gills are transformed into vascular wrappings which surround the body, whereas the balancers of some salamanders have been said to have a partly respiratory effect, although this is doubtful. The skin in amphibians certainly has a respiratory function, and that of *Cryptobranchus* is thrown into vascularized folds which are waved about. In the "hairy" frog the skin of the breeding male develops filamentous processes which aid in respiration.

Gill ectoderm and endoderm. It has been claimed that the epithelium of internal gills is endodermal, but because the boundary between endoderm and ectoderm is indistinguishable by the time the gills arise, this is difficult to prove. It has also been stated that the ectoderm grows inward and covers the
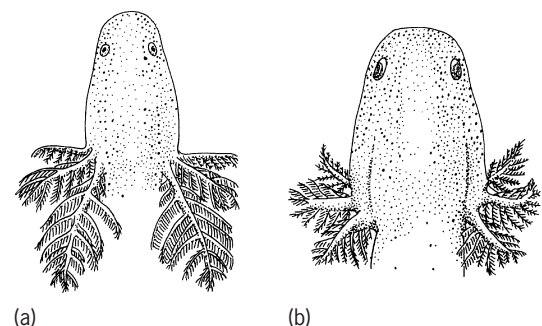

Fig. 8. External gill form in adult urodeles. (*a*) *Pseudobranchus striatus*. (*b*) *Siren lacertina*. Drawn from living specimens about 6 in. (15 cm) total length. (*After G. K. Noble, The Biology of the Amphibia, Dover, 1931*)

areas from which the gill lamellae and rakers develop, thus making the epithelium of these parts ectodermal. This would at least account for occasional rakers with enamel, a substance supposed to be derived only from ectoderm. The point of origin of gills designated as truly external marks their epithelium as clearly ectodermal; this is also probably true of the later internal gills of those amphibians which have them.   Robert S. McEwen; Thomas S. Parsons

Physiology of Gills

Gills are involved in many functions, including the exchange of oxygen, carbon dioxide, water, acid, inorganic ions, ammonia, and urea between the body and the surrounding water. Much of the transfer of these compounds is passive, but some ions are moved against an electrochemical gradient requiring the expenditure of energy. As a result, the gills are a metabolically active tissue; they consume up to 10–25% of the oxygen uptake of a resting fish.

Functionally, the gills can be considered as a large but thin, epithelial surface layer separating the inside (blood) of the animal from the water. In mammals and other air-breathing vertebrates, oxygen and carbon dioxide are transferred across a lung epithelium. The lung has a very different gross structure than gills and is ventilated in a different way. This is because gills are specialized for exchange of material in water whereas lungs have evolved for gas exchange in air. Water is about 1000 times more dense and viscous than air and contains only 1/30 as much molecular oxygen; oxygen diffuses 10,000 times more slowly in water than air. To ensure adequate oxygen transfer from water into the animal, most aquatic animals maintain a continual, unidirectional flow of water over the gills, with only very small diffusion distances. The gills are essentially a fine sieve placed in the path of the water flow so that a thin layer of water passes over each region of the gill epithelium.

The exchange units of the gills of fishes are the lamellae, which are arranged in rows on filaments attached to gill arches (**Fig. 9**). Blood flows through the lamellae in the opposite direction to water flow,
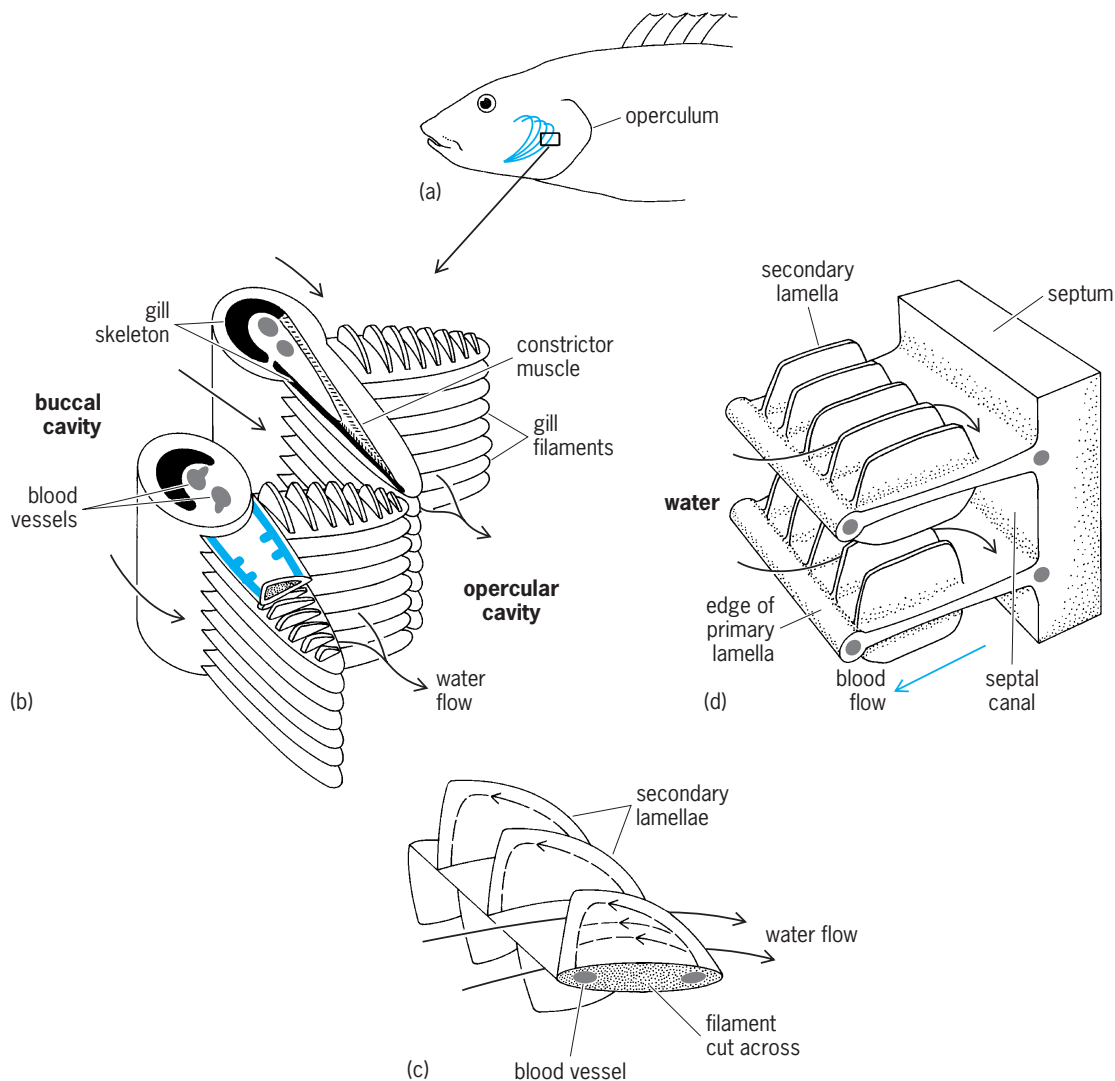

Fig. 9. Structure of fish gills. (*a–c*) Structure of a bony fish, with increasing magnification. (*d*) Elasmobranch gill at the same magnification as *c*. Black arrows show path and direction of water flow. Compare with Fig. 4.
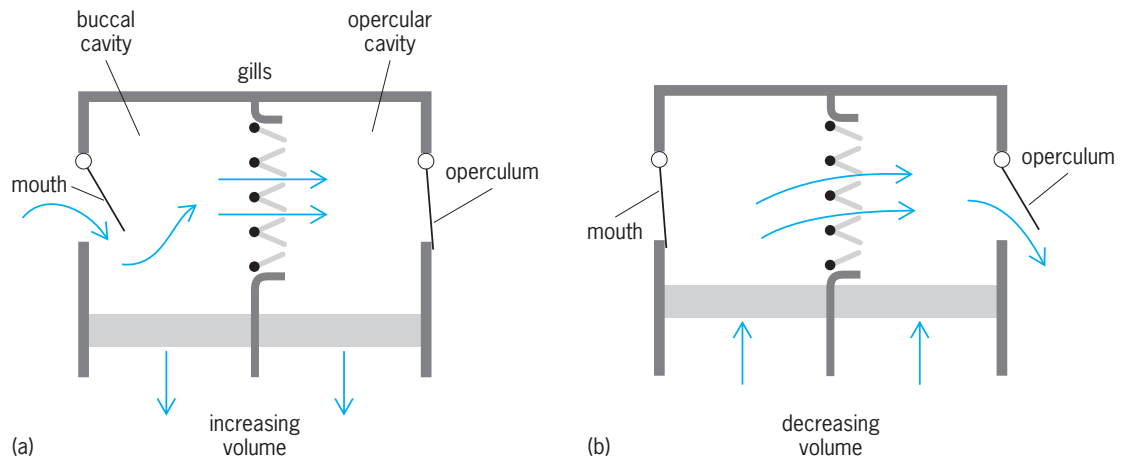
Fig. 10. Diagram of the mechanism of generating water flow over the gills of a fish (*a*) with mouth open and (*b*) with mouth closed.

creating a countercurrent exchange system between blood and water across the gill epithelium. Oxygen is delivered to the gill surface by water flow, diffuses across the gill epithelium, and is carried away in the blood bound to hemoglobin. Oxygenated blood, because of the presence of hemoglobin, contains much more oxygen per unit volume than water. The flow of water over the gills is much greater than the flow of blood through the gills, but the volume flow times oxygen content of blood leaving and water entering the gills is about the same.

Gill ventilation. Water flow over the gills of fishes is maintained by the action of muscular pumps and unidirectional flap valves situated at the mouth and the opercular cleft in bony fishes, and the mouth and gill slits of elasmobranches. The valves are skin folds that permit flow in only one direction, into the mouth and out via the gill slit (**Fig. 10**). The mouth (buccal) cavity expands, drawing water into the mouth; at the same time the postgill or opercular cavity expands with the outlet valve closed, drawing water over the gills; the opercular and buccal cavities then contract, forcing water out via the gill slit. The buccal force and opercular suction pumps operate at the same frequency but slightly out of phase so that, although water velocity over the gills varies with each breathing cycle, there is nearly always some water flow over the gills, with little or no backflow. The relative importance of the buccal versus the opercular pump differs among species.

Fast-swimming fishes, for example mackerel and tuna, swim forward with an open mouth so that water is forced over the gills by the forward motion of the fishes. This is referred to as ram ventilation: clearly there is no requirement for, nor are there any, buccal or opercular movements. If, however, the animal stops swimming, then the buccal and opercular pumps are used to generate gill water flow. Some fishes, for example tuna, swim all their lives and maintain water flow by ram ventilation. The actual water flow rate depends on the swimming speed and is adjusted at any one speed by changing the size of the mouth gap.

Some amphibians have external gills, that is, the gills are not enclosed in a buccal cavity. These gills are not ventilated but are waved around in the water to ensure adequate oxygen delivery. Some other amphibians have neither gills nor lungs but obtain oxygen from the water simply by diffusion across the general body surface.

Low oxygen levels often occur in rivers and lakes and even in the sea. Under these conditions, many fish supplement their oxygen supply by breathing air. The air-breathing organ is usually the swimbladder or a buccal or gut cavity and not the gills, which are retained for gas transfer in water. Some fish are obligate air breathers and die if denied access to air. In this case, the gills are retained for excreting carbon dioxide into water and maintaining ionic balance. The air-breathing organ is for oxygen uptake. This is associated with changes in the circulatory system (**Fig. 11**) culminating in the completely divided circulation of mammals and birds.

Blood supply. The fish heart is situated ventral to the gills. Blood, pumped through the gills by the heart, spreads into a thin sheet within the lamellae. Blood pressure within the gill lamellae is high, but the lamellae do not expand because the two sides of the structure are held parallel to each other by a number of cells that contain collagen fibers, called pillar cells. Their collagen fibers are extracellular but contained in lateral folds of the pillar cell and extend along the lateral face of the lamellae, enclosing and supporting the blood space and retaining its thin, sheetlike nature in the face of a high blood pressure. Such high pressures are required to drive blood around the body after it has left the gills.

Level of oxygen transfer. If oxygen levels in the water are low, fishes attempt to maintain oxygen supply by increasing water flow across the gills. If the oxygen requirements of the fishes increase (for example, during exercise), then both water flow and blood flow are increased. The blood is more evenly distributed throughout the gills, and the epithlium is stretched and thinned, enhancing the area and reducing the thickness of the gill epithelium. This
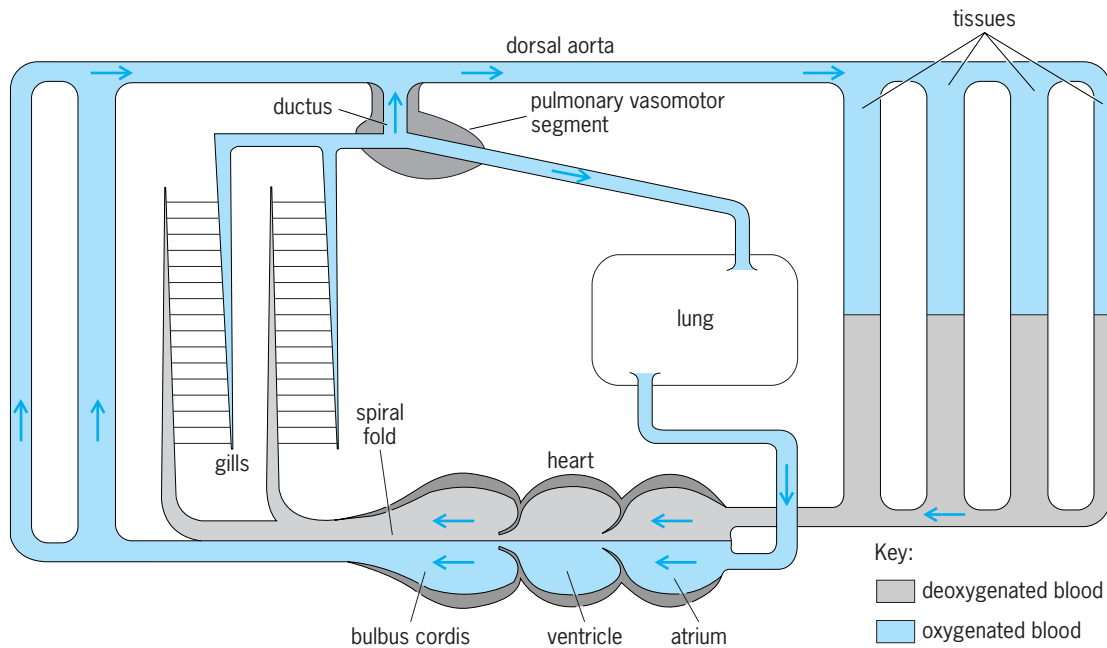
Fig. 11. Circulation of the African lungfish, *Protopterus*. Separation of oxygenated blood and deoxygenated blood is achieved by a septum which divides the atrial and ventricular chambers. Blood flows directly to the systemic circulation via the dorsal aorta. The blood is directed to the dorsal aorta or lungs depending on whether the fish is breathing in air or water. (*After D. Randall et al., eds., Eckert Animal Physiology, 4th ed., W. H. Freeman, New York, 1997*)

increases the capacity of the gills to transfer gases. The fish can increase oxygen transfer across the gills by 10–15 times and so maintain high levels of aerobic exercise. There are oxygen receptors on the gills that respond to reduced oxygen levels in the water and in the blood. These receptors when activated result in an increase in gill water flow. They may also cause a slowing of the heart rate, but cardiac output is not much affected because stroke volume increases. It seems that fishes adjust water and blood flow in the gills to meet the oxygen requirements of the animal rather than to regulate the carbon dioxide excretion of the fishes.

CO_2 delivery to gills. The tissues of animals produce about the same amount (or a little less) of CO_2 as the amount of O_2 consumed. The exact ratio of CO_2 produced to oxygen utilized (the respiratory quotient) depends on whether fats, proteins, or carbohydrates are used as the metabolic substrate. Carbon dioxide is transferred more rapidly than oxygen in tissues and water, so if the respiratory system is set to deliver adequate supplies of oxygen from water to the tissues, there will be no problem excreting carbon dioxide in the reverse direction. The only problem concerns the fact that CO_2 reacts with water to form carbonic acid which dissociates to form bicarbonate. At body pH most of the CO_2 exists as bicarbonate, but membranes are much more permeable to CO_2 than bicarbonate. Thus bicarbonate in tissues is converted to CO_2, diffuses into blood, forms bicarbonate, and is carried to the gills, where it is converted to CO_2 to diffuse across the gills into water. The conversion of CO_2 to bicarbonate and the reverse reactions occur slowly unless catalyzed by the enzyme carbonic anhydrase. Teleost fish red blood cells and fish gills have high levels of this enzyme but no activity is found in the plasma. Bicarbonate in plasma enters the red blood cell and is converted to CO_2 before diffusing across the gills. Carbonic anhydrase in the gill epithelium appears not to play an important role in CO_2 excretion but rather couples CO_2 excretion to ion transfer (**Fig. 5**). Elasmobranchs have carbonic anhydrase activity in the plasma as well as the red blood cells and the gills.

Ion transfer. Animals in fresh water lose NaCl across their body surfaces. This loss is counterbalanced by an uptake of NaCl in the gills. Influx of Na^+ is coupled to H^+ or NH_4^+ excretion, whereas Cl^- is coupled to HCO_3^- excretion in a separate mechanism. The HCO_3^- excreted via this pathway represents only a small portion (perhaps 5–10%) of the total CO_2 excretion. Sodium uptake via a sodium channel is maintained by the activity of a proton ATPase on the apical membrane of the gill epithelium. The excretion of protons generates a potential across the apical membrane, the inside negative. This creates an electrochemical gradient for sodium and draws sodium into the cell (**Fig. 12**). The sodium is then transferred into the blood via a Na^+/K^+ ATPase on the basolateral border of the epithelium. Carbonic anhydrase catalyzes CO_2 hydration and supplies protons for the proton pump. The associated bicarbonate production is exchanged for chloride. Seawater fishes have evolved specialized chloride cells which remove excess NaCl from within the fishes (**Fig. 13**). *See* OSMOREGULATORY MECHANISMS.

Fishes like the salmon that can tolerate a wide range of salinity increase the number of chloride cells within the gill epithelium as they migrate from rivers to the sea. An associated accessory cell develops
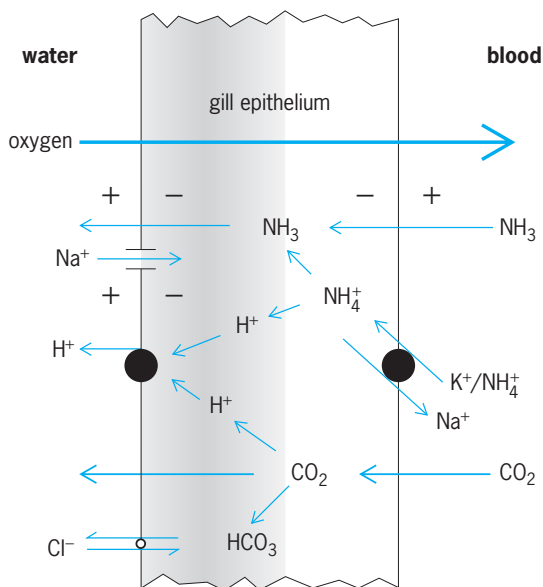
Fig. 12. Model of gas, sodium, and proton transfer across the gill epithelium of a fresh-water teleost. Shading indicates carbonic anhydrase activity. (*After H. Lin and D. Randall, Proton pumps in fish gills, in C. M. Wood and T. J. Shuttleworth, eds., Cellular and molecular approaches to fish ionic regulation, Fish Physiology, vol. 14, pp. 229–255, Academic Press, New York, 1995*)

alongside each chloride cell with loose connections forming a gap between the cells (paracellular channel). The chloride cell is much larger than the epithelial cells and extends from the outside across the gill epithelium. The epithelial cells and the chloride cells are tightly cemented together; the only open channel is between the chloride and accessory cell. The basal (blood) border of the chloride cell is infolded, and the membrane is associated with high levels of the enzyme Na^+/K^+ ATPase. This enzyme is associated with Na^+ removal from the chloride cell into the blood. This creates a Na^+ gradient from blood into
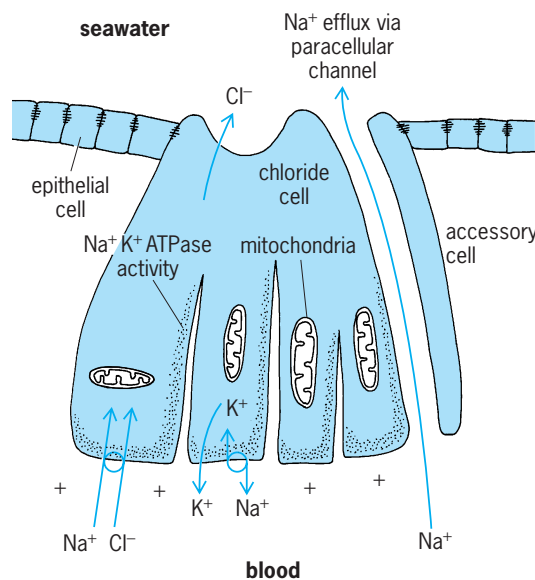

Fig. 13. Sodium chloride (NaCl) excretion in seawater fish occurs via the chloride cell found within the gill epithelium.

the chloride cell. Na^+ diffuses back into the chloride cell, involving a carrier that couples chloride movement to Na^+ movement. In this way chloride builds up in the cell and, because the cell is negatively charged, there is a large electrochemical gradient from the chloride cell into seawater and chloride is excreted from the cell. This net movement of chloride from blood to seawater results in the development of a positive potential across the gill, creating an electrochemical gradient for sodium, which diffuses from blood to seawater mainly via the paracellular channels. If a fish returns to fresh water, this paracellular channel closes, the accessory cells disappear, sodium loss is reduced, and chloride efflux also declines. Thus the gills of fishes are very much involved in maintaining internal NaCl levels.

Ammonia excretion. The main excretory end product of protein metabolism in fishes is ammonia (NH_3), but some urea is also produced. Ammonia and urea are both lost by passive diffusion across the gills. In elasmobranchs the body surfaces are not so permeable to urea, which is retained at high levels along with trimethylamine oxide to create conditions in the body fluid which are iso-osmotic with seawater.

Most of the ammonia excreted can be accounted for by passive diffusion of ammonia gas. Membranes are very permeable to NH_3 but less so to ammonium ion, which is sometimes moved by a carrier molecule through membranes. Ammonia is toxic, and if levels in the water increase as a result, for example, of the deposition of biological waste, then fishes cannot excrete ammonia and it builds up in their bodies, and they die. Under some circumstances fishes can excrete ammonium ions in exchange for sodium via a carrier-mediated process which may occur against a net ammonia gradient. Most ammonia excreted across the gills is produced in the liver, but some may be produced in the gills, which have the capacity to deaminate adenylates. The gill tissue also has the capacity to produce glutamine, but there is no evidence that ammonia is ever excreted as glutamine.

pH regulation. Fish have difficulty excreting ammonia into alkaline waters or when air-exposed. Some fish have a complete ornithine urea cycle and can convert ammonia to urea for excretion. For example, the Lake Magadi (Kenya) tilapia, living in very alkaline conditions, can produce urea via the ornithine urea cycle at about the same rate as a rat. Other fish store ammonia as glutamine or other amino acids during air exposure and then convert the amino acids back to ammonia for excretion after returning to water. The gill epithelium also appears to transport hydrogen ions. Thus any acid produced by the fish could result in the movement of H^+ across the gills into the water. The environment can be considered, therefore, as a very large H^+ sink, ameliorating pH changes within the animal. When, however, the water is acidic, proton excretion is inhibited, and this in turn inhibits sodium uptake and sodium levels in the fish and the animal dies. As a result, acid waters are a danger to many fish. Calcium levels in the

water decrease the permeability of the gill epithelium to ions, including H^+, so acid conditions in the water are better tolerated if calcium levels in the water are high. Industrial pollution in the form of acid rain quickly reduces the pH of soft-water rivers and lakes. These waters are poorly buffered and also have low Ca^{2+} levels; as a result the fishes in these waters are very susceptible to acidification. Large numbers of lakes in North America and Europe no longer contain fishes because of acidification due to industrial pollution. *See* ACID RAIN; PH REGULATION (BIOLOGY).

David J. Randall

Anatomy of the Lung

The shape and volume of the lung, because of its pliability, conforms almost completely to that of its cavity. The lungs are conical; each has an apex and a base, two surfaces, two borders, and a hilum.

Thoracic cavity. The apex extends into the superior limit of the thoracic cavity. The base is the diaphragmatic surface. The costal surface may show bulgings into the intercostal spaces. The medial surface has a part lying in the space beside the vertebral column and a part imprinted by the form of structures bulging outward beneath the mediastinal pleura (**Fig. 14**). The hilum and pulmonary ligament descending from it are notable. The cardiac impression is deeper on the left lung because of the position of the heart. The aorta arches over the left hilum, and the azygos vein over the right. Joining the right cardiac impression are the groove for the superior vena cava in front of the hilum and that for the inferior vena cava in front of the pulmonary ligament. Other impressions shown in the figures are those of the esophagus and trachea, the left subclavian artery, and brachiocephalic or innominate vein. The borders of the lung are pinched extensions between the pericardium and body of the sternum (anterior border) and between the diaphragm and body wall (inferior border). The inferior limit of the thoracic cavity on both sides is related to ribs 8, 10, and 12, whereas the lower borders of the lungs stop at ribs 6, 8, and 10, and the unoccupied space is the costodiaphragmatic recess on both sides. The cardiac notch is the absence of lung because of pressure from the heart. A similar, but smaller, notch is made in the underlying pleura, which gives rise to the left costomediastinal recess.

The oblique fissure cuts through the costal, diaphragmatic, and medial surfaces to the root of the lung. In the right lung the horizontal fissure runs backward from the anterior border and meets the oblique fissure in the midlateral line. Thus the right lung has three lobes, superior, middle, and inferior, whereas the left lung has two, the superior and inferior. The lingula, or anteroinferior part of the left upper lobe, and the cardiac notch above it correspond to the middle lobe of the right lung.

Bronchopulmonary segments. For convenience in exploration and study of the lung, it may be divided into anatomical areas. The bronchial tree branches mainly by dichotomy (**Fig. 15**). The ultimate generations, that is, the respiratory bronchioles, alve-
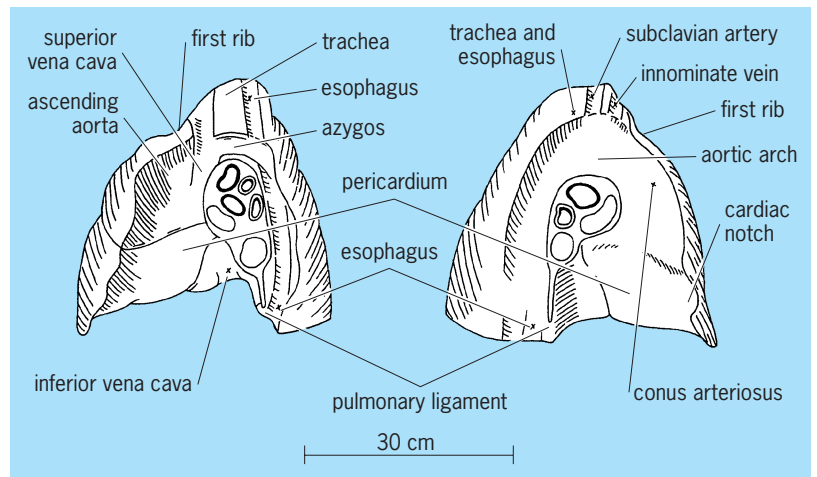

Fig. 14. **Impressions on the mediastinal surfaces of the lungs.**

olar ducts, and alveoli constitute all of the respiratory portion of the lung. This respiratory portion (**Fig. 16**) consists of 10 segments in the right lung and 8 in the left, each of which is supplied by a tertiary branch of the bronchial tree (Fig. 15). Two or more of the segments make up a lung lobe. Further, the primary bronchi (from the trachea) divide as secondary bronchi, three on the right side and two on the left, corresponding to the lobes of the lung. Because the upper left lobe results from the fusion of two lobes, the prospective left upper and middle lobar bronchi become partially fused. As a result, the four segmental bronchi of the upper lobe are not tertiary, but bronchi of the fourth division of the tree. However, there is great similarity in the structure of the right and left bronchi.

The trachea and extrapulmonary bronchi are kept open by C-shaped bars of hyaline cartilage. Within the lung the bars are arcs staggered in the bronchial walls at different levels. When in their branching the bronchi and bronchioles are reduced to a diameter of 0.04 in. (1 mm) or less, they are then free of
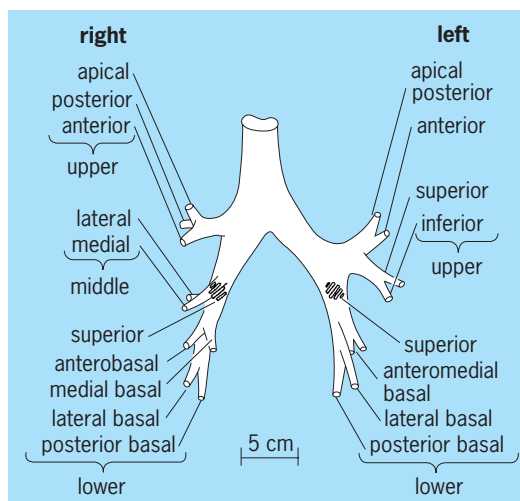

Fig. 15. **Dichotomous branching of bronchial tree, showing 10 right and 8 left segmental bronchi.**
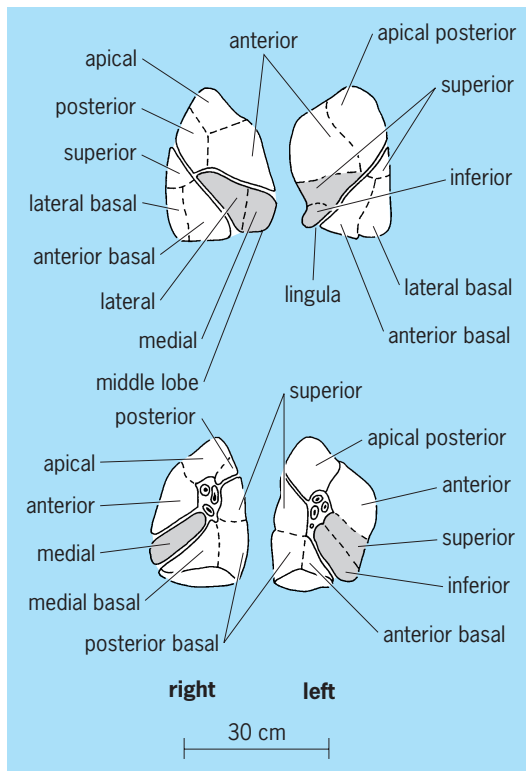
Fig. 16. The 10 right and 8 left bronchopulmonary segments.

pulmonary veins, except that to the largest bronchi, which is returned by the bronchial veins to the azygos veins.

The pulmonary veins have 10 branches; their main stems run in the medial or inferior sides of the bronchi. Because their tributaries run intersegmentally and drain adjacent segments, and because the arteries may cross intersegmental boundaries, a bronchopulmonary segment would not be a morphologic bronchovascular unit. The upper and lower lobes of the left lung send an upper and a lower pulmonary vein, respectively, into the left side of the left atrium. The upper and middle right lobes provide the upper vein to the right side of the atrium.

Lymphatic channels are not found in interalveolar partitions but occur everywhere else in the lung, except in cartilage. There are two main sets of lung lymphatics, those in the pleura and those within the lung. The latter begin at the alveolar ducts and follow the bronchi and pulmonary vessels to the lymph nodes at the hilum. Those in the pleura form secondary and primary networks in the lobular septa and drain into the hilar lymph nodes.

Nerve supply. The nerves which supply the lung are branches of the vagus and of the thoracic sympathetic ganglia 2, 3, and 4. Efferent vagal fibers are bronchoconstrictor and secretory, whereas the afferents are part of the arc for the breathing reflex. Efferent sympathetic fibers are bronchodilators; hence, the use of adrenalin for relief of bronchial spasm resulting from asthma.

cartilage and are called terminal bronchioles. One of the terminal bronchioles enters the apex of a secondary lobule of the lung. These secondary lobules are anatomic units of the lung, whose hexagonal bases, 0.4–0.8 in. (10–20 mm) in diameter, rest on the pleura (Fig. 16) or next to a bronchiole or blood vessel, and whose apices point toward the hilum. Finer lines divide the bases of the secondary lobules into smaller areas (**Fig. 17**). These are the bases of primary lobules, each served by a respiratory bronchiole.

Blood supply. The blood supply to the lung is provided by the pulmonary and the bronchial arteries. The right pulmonary artery runs dorsally between the bronchi of the right upper and middle lobes so that the upper bronchus arises 1 in. (25 mm) from the trachea. The left pulmonary artery and the arch of the aorta pass dorsally between the trachea and the bronchus of the left upper lobe bronchus, so that the latter is 2 in. (50 mm) from the trachea. Each pulmonary artery divides into 10 branches, which follow closely the posterosuperior walls of the segmental bronchi. They take the names of the 10 right segmental bronchi.

The bronchial arteries arise on the left side from the aorta, and on the right from either an intercostal or the left bronchial artery. They supply the walls of the bronchi, pulmonary vessels, and lymph nodes. They pass with the radicles of the pulmonary vein through the interlobular septa and supply the pulmonary pleura. Blood delivered to the lung by the bronchial arteries is returned by the radicles of the
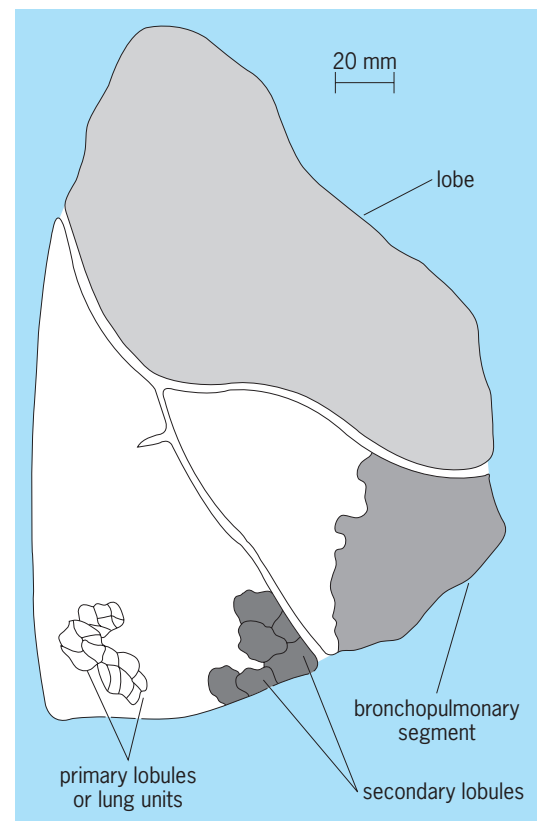

Fig. 17. Subdivisions of the lung.

Histology of the Lung

Essential to lung function, histologically, are the respiratory passages, pleura, blood vessels, lymphatics, and respiratory surfaces proper.

Epithelium. Between the cartilage in the walls and the surface epithelium of the air passages are serous and mucous glands, which continuously pour their secretion through ducts to the surface of the epithelium. Also between the cartilage and epithelium below the trachea are smooth muscles which increase relatively in amount with decrease in size of the bronchioles. In the terminal bronchioles, which are free of cartilage, smooth muscle is the main constituent and may close the bronchiole, as in asthma. The epithelium is the respiratory type; that is, the cells are tall and capped by motile cilia. The cilia wave toward the throat, carrying along the secretions of the glands laden with inspired dust.

Pleura. The pleura lines the outer surface of the lung and the inner surface of the pleural cavity. It consists of a layer of flattened mesothelial cells and an underlying layer of collagenous and elastic fibers which support numerous blood and lymphatic capillaries, fibroblasts, and macrophages. The pleura continuously pours out a mucoid exudate, which lubricates the opposed surfaces. The pleura is pink in the newborn child.

Respiratory portion. This portion includes the respiratory bronchioles, which branch from the terminal bronchioles, the respiratory ducts, alveolar sacs, and alveoli. The respiratory ducts are similar to terminal bronchioles, except that they are smaller and have a few scattered alveoli protruding from their walls. Alveolar ducts, the next order of branching, have their walls studded with contiguous alveoli so that the openings of the alveoli occupy the greater area of the wall. Alveolar sacs are terminal dilations of the alveolar ducts. Alveoli are the smallest functional units of the lung. Their walls are made up of one or two capillaries, a few elastic fibers, and more reticular fibers, among which are a few fibroblasts, dust cells or lung macrophages, and septal cells. The septal cells give rise to some of the macrophages, and others are carried in by the blood. Macrophages from both sources wander freely in the alveolar spaces. Two alveoli frequently have a single wall in common. The epithelial lining of the alveoli is so thin that it cannot be demonstrated with ordinary techniques. The greater volume of the lung is taken up by air space of the respiratory portion. Thus the lung feels spongy and, when seen in sections, looks like fine lace.

Embryology of the Lung

The lung is an organ adapted to respiratory exchange of gases between air and the blood of vertebrates. It ranges from the simple swim bladder of dipnoid fish to the large spongy, compound air sac of humans. The story of the lung includes the evolution of certain vital accessories to respiration, the diaphragm and thoracic wall. All lungs correspond in originating as a pocket from the pharynx near the level of
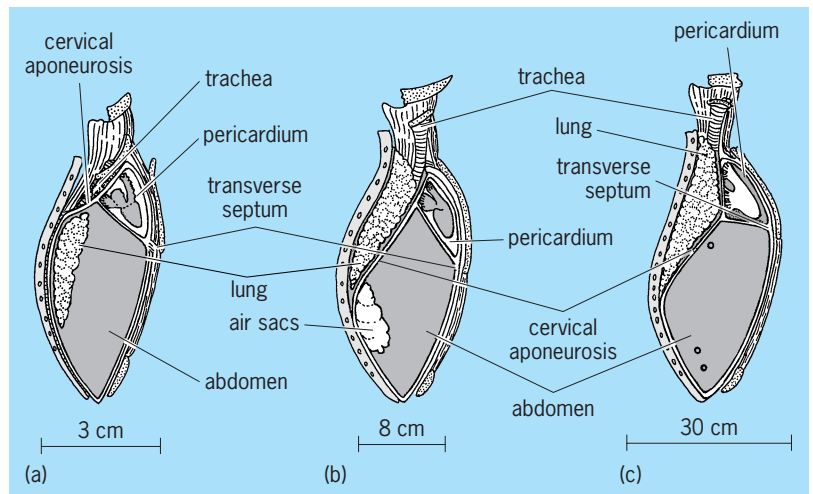
Fig. 18. Position of the lung in vertebrates. (*a*) Fishes and amphibians. (*b*) Birds. (*c*) Mammals.

the sixth aortic arches which furnish the blood supply; in innervation by the vagus nerve; and in their embryonic position in the abdomen. This position is retained in the adult of fish, amphibians, and reptiles and partially retained in birds, but in mammals the lung is segregated from the abdomen in fetal life into a new-formed thoracic cavity (**Fig. 18**). *See* PHARYNX; SWIM BLADDER.

Extrusion of lung. In the extrusion of the lung from the abdominal cavity it first invaginates the cervical

Fig. 19. Lung extrusion in (*a*) amphibians, (*b*) birds, and (*c*) mammals.

aponeurosis of the abdomen and in so doing enters the space being formed between the inner and middle layers of the body wall, the pleural cavity. These events bring the lung to an anterior, thoracic position, and also allow it to develop an apex,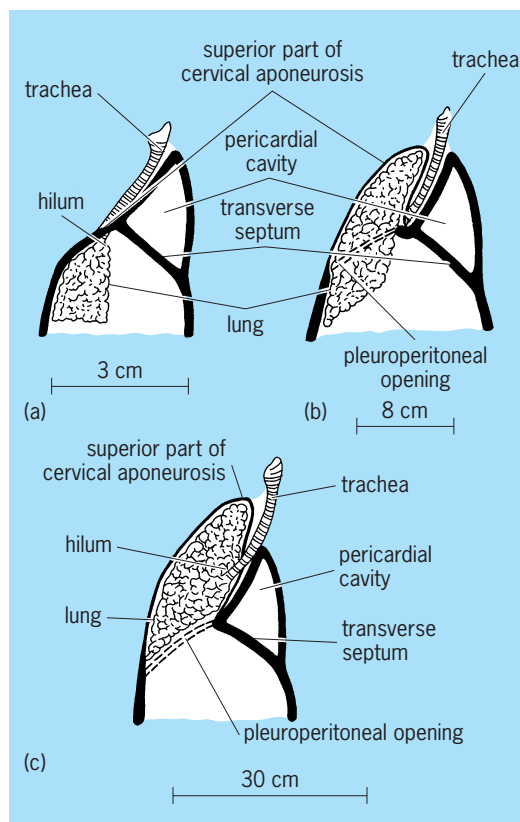 both important in respiration. In amphibians (**Fig. 19***a*) the lung comes to lie in the pleuroperitoneal cavity. It is infrapericardial, with both the root and apex in the same plane. Figure 19*b* represents the condition in birds in which an apex grows cranially beyond the root of the lung. Thus the entire lung remains abdominal, because the apex carries forward the entire cervical aponeurosis. The human lung becomes supradiaphragmatic (Fig. 19*c*). The diaphragm is represented by only a thin layer of the aponeurosis that is carried forward (Gibson's fascia), and the remainder of the aponeurosis serves to close the passageway. Moreover, the root also is carried forward, and in humans it becomes closely bound to the pericardium (**Fig. 20**).

The pericardial bond is of primary importance in humans and the great anthropoids because the pericardium is bound also to the central tendon and diaphragm (Fig. 20). The part of the lung above the pulmonary root is relatively large, and the apex, located at a fixed point at the level of the neck of the first rib, can expand when the diaphragm contracts and pulls the root and heart in a downward and forward direction (**Fig. 21**). Upon expiration they move in the reverse direction. In the typical mammal there
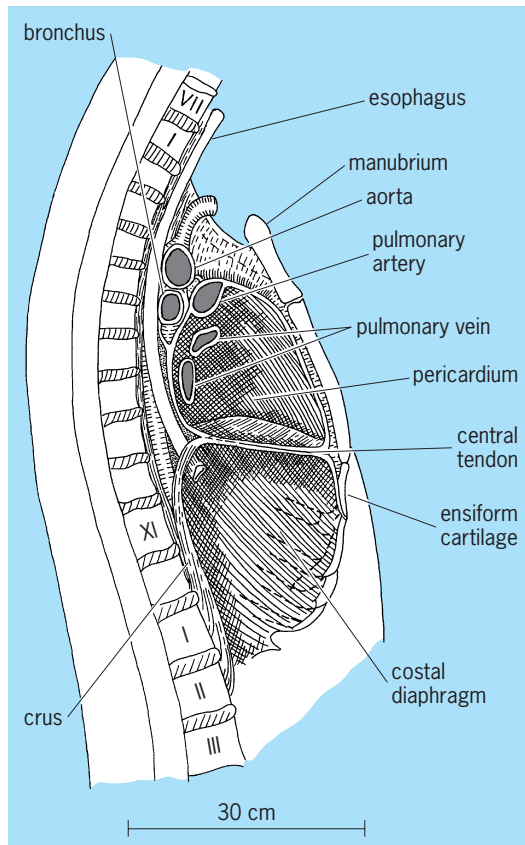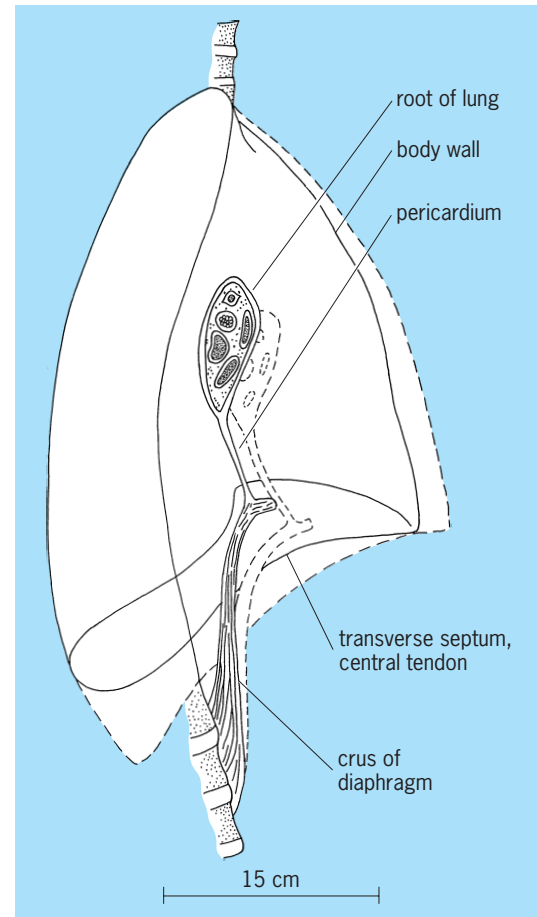

Fig. 21. Diagram showing movements of the diaphragm, pericardium, and root of the lung during respiration in higher mammals. Broken lines show position of thorax and viscera during inspiration.

Fig. 20. Diagram showing closely bound connections of the diaphragm, pericardium, and root of the lung in a human.

is little or no contact of the diaphragm with the pericardium, and the part of the lung above the root is relatively small (**Fig. 22**). Here the descent of the diaphragm has little effect on the movement of the heart and pulmonary roots. Instead, the azygos lobe expands between the esophagus and inferior vena cava into the subpericardial space beside the pericardial mesentery.

The body wall becomes an inspiratory mechanism in reptiles as a result of the ribs, sternum, and intercostal muscles which evolve in the middle layer of the wall. In phylogeny the first intermuscular septum to extend completely to the median ventral line is that behind the seventh segment of the rectus abdominis in the amphibian *Necturus maculatus*. In this septum the first sternal rib develops in typical reptiles and mammals.

Amphibia. The amphibian lung is bifid, and each half projects into the abdomen above the pericardium and liver (**Fig. 23**). The inner surface of the lung presents folds and ridges which increase respiratory surfaces. Air is pumped into the lung by pharyngeal muscles and in the absence of ribs is forced out by abdominal muscles.

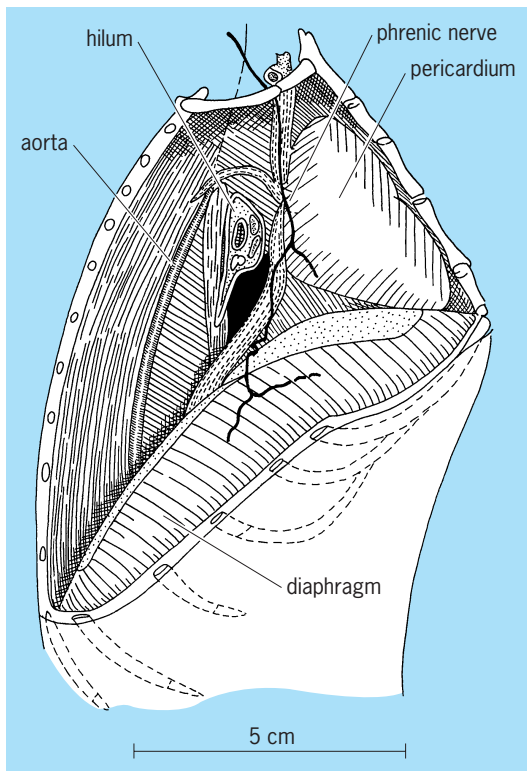**Reptilia.** The collapsed lung of *Iguana* (**Fig. 24**) extends forward to the cervical aponeurosis. The 8th

Fig. 22. Diagram showing discontinuity of the diaphragm with the pericardium and hilum of the lung in the typical mammal (rabbit).

Fig. 23. Diagram showing the lung in an amphibian (Surinam toad) and the muscles from which the diaphragm has evolved.

to 14th ribs are complete, and three cervical ribs can both expand and compress the apex of the lung. The sternum is a strong fulcrum for the costal cartilages. The cervical part of the trunk is increased in length, so that the head and pharynx are carried away from the heart. Whereas in amphibians the swallowing of air tends to force pulmonary blood out of the lungs, in the lizard the negative pressure produced by its costal mechanism draws both air and blood into the lungs.

Birds. Lungs in adult birds are extra-abdominal, except the posterior ends, which are dilated to form abdominal air sacs. However, the avian lungs have grown through the cervical aponeurosis of the abdomen into a new space in the body wall, the pleural cavities (**Fig. 25**). Air sacs develop in this aponeurosis, and the septum is divided by these into two layers, the dorsal or pulmonary and ventral or abdominal (Fig. 25). If these lungs were removed from the thorax and the avian septum were replaced against the dorsal wall, the three-layered body wall of the amphibian would be restored. Although the lungs are in the thorax, the air sacs remain in the abdomen.

Human embryo. The lung in the 5-week human embryo lies in the abdomen. The septum transversum in humans consists of a ventral part (**Fig. 26***b*, *c*) representing the entire septum of amphibians (Fig. 19*a*) and a dorsal part representing a fusion of the amphibian cervical aponeurosis with the dorsal wall of the pericardium. The common cardinal vein (Fig. 26*b*, *c*)
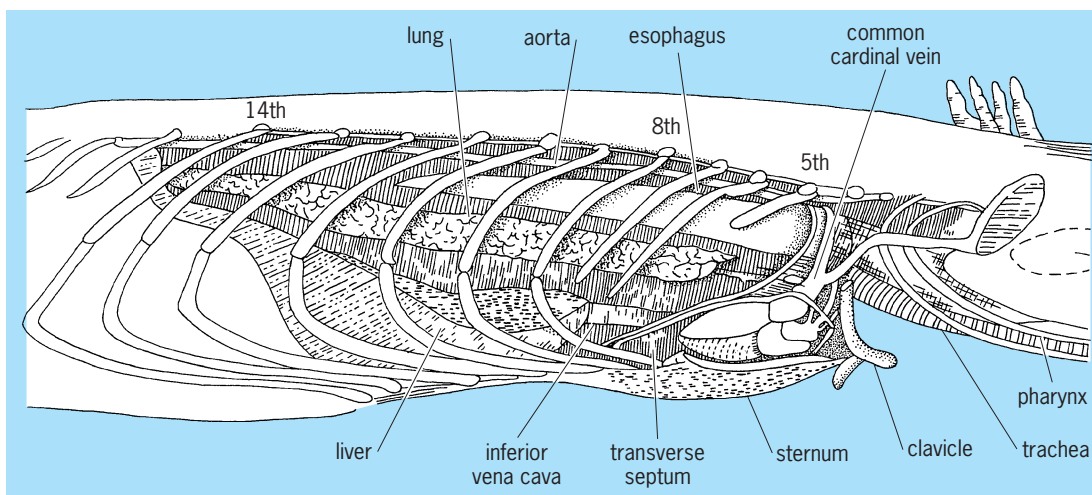
Fig. 24. Diagrammatic sketch of a reptile, *Iguana* sp., showing the collapsed abdominal lung and the rib structures.

Fig. 25. Cross-sectional diagram of thorax (lung), abdomen, and pericardium in pigeon.

runs in the anterior part of the septum next to the wall of the pericardium. The lung lies medial to the Wolffian fold, which is attached to the mesonephros dorsally, the liver ventrally, and the septum in front,

Fig. 26. Human embryo. (*a, b*) Diagram of embryo of 5 weeks showing initial process of lung extrusion. (*c*) Embryo of about 7 weeks showing extrusion into the cervical aponeurosis.
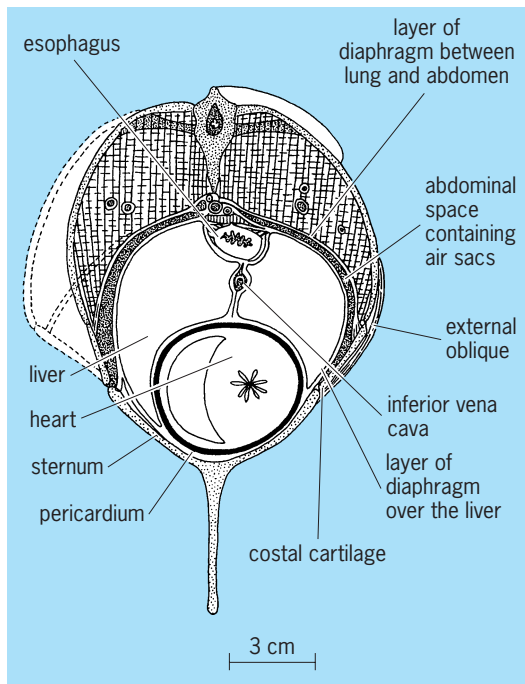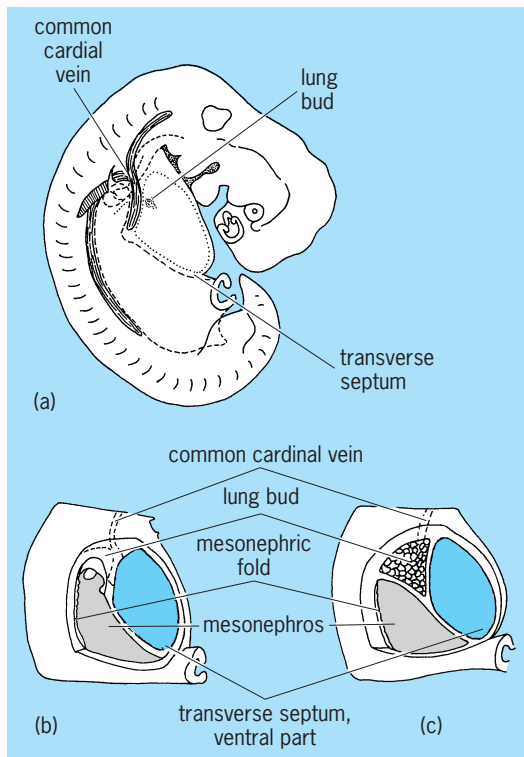
with its free border facing backward. In the amphibians the lungs grow backward in the abdomen and are attached to the mesentery between the aorta and esophagus above and liver and stomach below. In the human embryo they advance medially to the septum (Fig. 26*b, c*), then evaginate laterally into the dorsal part of the septum between its anterior part, which contains the common cardinal vein next to the pericardial roof, and the posterior part, which develops to become the pleuroperitoneal membrane.

After this extrusion of the human lung from the abdominal cavity, the lung lies next to the medial surface of the middle layer of the body wall. Upon removal of the body wall, the lungs are exposed. The lungs can grow in a forward or backward direction and ventrally between the inner and middle layers of the body wall at equal rates, as these open up. The site of extrusion (pleuroperitoneal opening) need only close, and the lung will occupy new, completely enclosed, pleural cavities.

Leo P. Clements

Physiology of the Lung

Simple molecular diffusion of gases underlies respiratory exchange in all animals. The lungs of air-breathing vertebrates require mechanical ventilation to sustain this diffusion. The molecular characteristics of respiratory gases (**Table 1**), the physical properties of atmosphere (**Table 2**) and of body fluids (**Table 3**), the physiological limitations on lung structure and location, and the large requirement of the active animal for O_2 uptake and CO_2 removal (**Table 4**), all restrict the effectiveness of diffusion respiration alone. Diffusion suffices only within a few millimeters of those membranes across which interchange of O_2 and CO_2 can occur. Refinements in ventilation respiration intervene as continuing speciation involves increased O_2 requirements. The functional as well as structural changes follow recognizable evolutionary trends. *See* METABOLISM; RESPIRATION.

Other vital requirements. Besides exchange of O_2 and CO_2, there are other vital requirements for breathing processes in air. Among these are (1) limitation of the evaporation of water in general and control of the dissipation of heat in birds and mammals in particular; (2) ventilation of the olfactory membrane; (3) production of high intrapulmonary pressure for such acts as coughing, defecation, and parturition; and low intrapulmonary pressure for regurgitation; (4) provision for large pulmonary volume in such acts as sneezing, yawning, and phonation; and (5) allowance for control of buoyancy in aquatic animals. All these ancillary processes involve corresponding variations in the respiratory mechanisms. As animals are compared, differences are found in buccopharyngeal, pulmonary interstitial, and diaphragmatic musculature for inducing and directing movement of air in valvular muscle of nares, glottis, and pulmonary ducts or apertures for controlling flow and pressure; in accessory air sacs to increase capacity for ventilation itself, or for phonation; in heart for adequate pulmonary blood supply and in circulation, for efficient perfusion; in brain and peripheral nerve

TABLE 1. Characteristics of the respiratory media

Variable	Water		Atmosphere (N_2)	
	Ocean	Fresh	Sea level	6000-m altitude
Temperature, °C	−2.0 to 30.0	2.0−32.0	0.7−15.7	−28.1 to −15.1
Pressure, total mmHg	760−760,000	760−20,000	760	347.5 to 360.2
Density, g/liter	1027* (20°C)	1000* (4°C)	1.223−1.290	0.649−0.659
pH	7.5−8.4	3.2−10.6		
Concentration, vol %				
H_2O	94−97	97−99	1.00[†]	1.00[†]
N_2	1.03* (15°C)	1.33* (15°C)	78.03 (STP)	78.03 (STP)
CO_2	0.02* (15°C)	0.03* (15°C)	0.03 (STP)	0.03 (STP)
O_2	0.58* (15°C)	0.72* (15°C)	20.99 (STP)	20.99 (STP)
Salts	3.46*	0.18*		
Inert gases	Trace	Trace	0.95 (STP)	0.95 (STP)
Partial pressure (tension), mmHg				
H_2O	12.79 (15°C)	6.10 (4°C)	6.40[‡] (15°C)	0.72[‡] (−15°C)
N_2	593.02 (STP)	593.02 (STP)	593.02 (STP)	281.06 (STP)
CO_2	0.23* (STP)	0.23* (STP)	0.23 (STP)	0.11 (STP)
O_2	159.52* (STP)	159.52* (STP)	159.52 (STP)	75.61 (STP)
Inert gases	7.46 (STP)	7.46 (STP)	7.46 (STP)	3.42 (STP)
Total pressure	760.00	760.00	760.00	360.20
Diffusion coefficient, ml/(min)(cm^2)(cm), at 760 mmHg, 20°C				
H_2O				
N_2		0.000018 (0.53)[§]		
CO_2		0.000785 (23.1)[§]		
O_2		0.000034 (1.0)[§]	11.0	

* Average of many determinations; varies with conditions of measurement.
[†] Varies, but never absent and always of biological significance.
[‡] Calculated for 50% relative humidity.
[§] Values in parentheses are relative coefficients with O_2 as unity.

patterns for initiation, maintenance, and adjustment of ventilatory processes, control of circulation, and regulation of ancillary functions; and in skeletal and connecting structures. A brief review of comparative respiration can only delineate the basic trends in phylogeny of respiration. Certain general features of the breathing process must be considered before comparisons among species can be made. *See* SPEECH.

Breathing. The act of breathing induces mass flow of air in and out of the respiratory organs. This inspiration and expiration mechanically ventilates the entire lung of those animals, such as lungfish, salamander, and frog, in which the organ is a simple sac. Here the inner lining or pulmonary epithelium, where diffusion itself occurs, is directly exposed to the tidal airflow caused by breathing. In these animals, breathing is intermittent or periodic; that is, cycles of inspiration and expiration occur in groups after many minutes of nonbreathing or apnea. The periods of apnea may exceed those of ventilating. The lung is normally closed from the atmosphere by a valve (glottal valve) at the opening into the trachea supplemented by valves (nasal valves) at the nares. The lung, however, does not remain a simple sac. Beginning with toads and climaxing in birds and mammals, more elaborate secondary and tertiary sacculations, the alveoli and the alveolar or air sacs, evolve. This evolution increases the diffusion surfaces tremendously and also removes these surfaces from direct exposure to ventilatory airflow. Thus, a uniform gaseous exchange medium comes

TABLE 2. Characteristics of respiratory molecules*

Type	Weight (O = 16)	Diameter,[†] cm × 10^{-8}	Density, g/liter	Mean free path, cm × 10^{-6} (750 mmHg)	Collision frequency (20°C)	Average velocity, cm/s	Water solubility		Vol % (40°C)
							STP	20°C	
N_2	28.02	3.15−3.53	1.251	8.50	5070	45,400	2.35	1.54	1.18
H_2O	18.02	3.0−5.0	0.0005−0.030[‡]			56,600			
CO_2	44.01	3.34−3.40	1.977	5.56	6120	36,200	171.3	87.8	53.0
O_2	32.00	2.92−2.98	1.429	9.05	4430	42,500	4.89	3.10	2.31

* Unless otherwise indicated, values are for standard conditions (STP) of temperature (0°C) and pressure (760 mmHg).
[†] Range indicates variability with method of measurement (such as viscosity, heat conductivity).
[‡] Water vapor in saturated air, that is, in equilibrium with water at 0°C and 30°C.

TABLE 3. Respiratory exchange characteristics in humans

	VENTILATION					
	Inspired air		Alveolar air[a]		Expired air[a]	
Gas	Composition, vol %[b]	Partial pressure, mmHg[c]	Composition, vol %	Partial pressure mmHg[d]	Composition, vol %[b]	Partial pressure mmHg[d]
H_2O	0.00	5.7	00.0	47	00.0	47
N_2	79.02	596.0	80.4	573	79.2	565
O_2	20.95	158.0	14.0	100	16.3	116
CO_2	0.03	0.3	5.6	40	4.5	32

	TRANSPORT[e]							
	Arterial		Capillary		Tissue fluid		Venous	
Gas	vol %	mmHg	vol %	mmHg	vol %	mmHg	vol %	mmHg
H_2O	83 (81–86)	47	83 (81–86)	47	83 (81–86)	47	83 (81–86)	47
N_2	0.975	573	0.975	573	0.975	573	0.975	573
O_2	19.6 (17.3–22.3)	94	1–22.3[f]	1–94[f]	0.185[f]	30[f]	12.9 (11.0–16.1)[g]	40
CO_2	48.2 (44.6–50.4)	40	44.6–57.7[f]	40–50[f]	3.046[f]	50[f]	54.8 (51.0–57.7)[g]	46

[a] Alveolar air, actually last part of expired samples.
[b] Dry air, partial pressure in mmHg = (vol %)/100 × 760 mmHg (Dalton's law).
[c] Ambient air (slight variations exist), in vol %, = (100 × mmHg)/760 (Dalton's law).
[d] Physiological air, normal temperature (37°C) and standard pressure (760 mmHg).
[e] Values in parentheses are ranges.
[f] Variable, depending on blood flow, tissue activity, and relation of sample to capillary length or field.
[g] Internal jugular.

to prevail at the respiratory membranes of birds and mammals, the breathing cycles become continuous and shallow (eupneic), and the lung remains normally open to the atmosphere. Nasal valves are usually absent in animals above reptiles, and the glottis closes only for such occasional acts as swallowing and coughing.

Breathing in vertebrates. Breathing induces air movement resulting from a mechanically imposed difference in pressure between a compressible cavity and the atmosphere. In mammals and reptiles this cavity is the lung itself. In birds and amphibians, however, secondary cavities are responsible, namely, special air sacs in the former and the mouth (buccopharyngeal cavity) in the latter. These secondary cavities communicate with both the lung and the atmosphere and accommodate the bellowslike action. Two basic processes of air breathing reflect the structural differences. The more primitive type, buccopharyngeal breathing, is found in lungfish and amphibians. It involves the same basic neuromuscular

TABLE 4. Combined measurements of respiratory values in pigeon, duck, and chicken to compare roles of various sacs and lung

Category	Volume, % of total	Composition	
		O_2 %	CO_2 %
Tidal air	10–15		
Inspired	10–15	21.0	0.03–0.04
Expired	10–15	13.5	5.1–6.5
Interclavicular	20–25	14.6	5.0–6.9
Abdominal	50–70	19.0	1.9–2.7
Lung	10–12	20.0*	1.0*

*Estimated for parabronchi.

elements of mouth and throat as water breathing in fishes. The more specialized type, thoracoabdominal breathing (in humans, for example), involves trunk musculature to supply pulmonary ventilation in all reptiles, birds, and mammals. A rather continuous buccopharyngeal ventilation, not always involving pulmonary ventilation itself, is characteristic of amphibians (**Table 5**) and persists in lizards, turtles, and other reptiles. In amphibians, therefore, buccopharyngeal activity subserves both breathing and olfaction; in reptiles it subserves only the olfactory sense (smell).

Buccopharyngeal breathing. Buccopharyngeal breathing is indirect, when compared with thoracoabdominal. It involves two distinct stages: ventilation of the mouth and ventilation of the lungs. The necessary pressure gradients between mouth and atmosphere and between mouth and lungs are generated by muscles which raise and lower the hyoid apparatus and floor of mouth and throat; the same mechanism is used for water breathing in fishes. Pulmonary inspiration in buccopharyngeal breathing is more descriptively an injection or adspiration. The volume of air inspired or expired per breathing cycle is called tidal volume.

In the frog, movement of air between mouth and atmosphere requires only about 3–5 mm of water (H_2O) pressure; that between mouth and lungs reaches 25–35 mm H_2O pressure during the peak of pulmonary inspiration. A gradient is directed toward the lung of about 20 mm on inspiration, because a volume of gas (functional residual volume) remains in the lung from the preceding expiration under about 10 mm H_2O pressure. The glottis closes at the end of inspiration, and a positive intrapulmonary pressure persists during apnea, of about 20 mm H_2O, as a result of elastic recoil and muscular

TABLE 5. Some respiratory values in vertebrates

Animal[*]	Weight, kg	Breathing rate,[†] cycles		Tidal volume, ml	Minimum volume, liter/min	O₂ con-sumption, mm³/(g)(h)	Intrapleural pressure, cm H₂O	Com-pliance, liter/ cm H₂O
		Vent./min	Comp./h					
Frog (*Rana fusca*)		41	17 (26°C)			210.0 (20°C)		0.001
Turtle (*Malaclemys centrata*)		3.7 (24°C)		14.0	0.051	35.0 (24°C)		
Alligator (*Alligator mississippiensis*)						8.9 (22°C)		
Canary (*Serinus canarius*)		108				2900.0		
Chicken (*Gallus domesticus*)		17		45.0		497.0		
Duck (*Anas* sp.)		42		36.5		800.0		
Rat (*Rattus norvegicus*)	0.273	60	26	1.4	0.074	770.0	−2 to −8	0.0012
Dog (*Canis familiaris*)	20	17	5	302.0	5.30	580.0	−5.4 to −13.5	0.09
Horse (*Equus caballus*)	696	12		9060.0	107.0	250.0	−8.0 to −22.0	0.80
Human (*Homo sapiens*)	66	14	3	372.0	5.04	220.0	−3.8 to −9.3	0.20

[*] Adult male at rest.
[†] Vent. = ventilatory, except buccopharyngeal in frog; comp. = complementary, except pulmonary in frog (see text).

tonus in the lung itself and in the body wall. These forces also cause expiration when the glottis opens. Whether buccal ventilation alone or pulmonary ventilation or a combination of these occurs depends upon neuroregulatory processes which determine the relationship of nasal and glottal valves with each other and with the breathing musculature. Thus a frog's lungs and body can be distended greatly beyond normal dimensions by successive inspirations alone.

Vital capacity. The excess capacity of any animal to inspire beyond normal tidal volume is called inspiratory reserve volume, and the total breathing capacity of the lungs as measured by the volume which can be completely expired after maximum filling is the vital capacity. Included in this is an amount, called expiratory reserve volume, which can be expired from the functional residual volume. Total expiration might completely empty the simple lungs of some amphibians. However, as the lungs of animals elaborate with alveolar development, it is not possible to expire all lung contents. The remainder after limit of vital capacity is reached is called residual air, and its presence at metamorphosis, hatching, or birth always indicates that breathing has started. The various volumes and pressures of ventilation have not been measured in most species. Some representative values are given in Table 5.

Dead space. Buccopharyngeal ventilation continues in reptiles, but this mechanism no longer provides for pulmonary ventilation. It remains an important adjunct to breathing, however, because it serves to reduce the dead-space volume. Dead-space average normal volume for an adult human, for example, is about 150 ml and tidal volume about 500 ml, which means an actual ventilatory volume of about 350 ml. In a few air-breathing animals which occupy an aquatic habitat, respiration is apparently supplemented by buccopharyngeal breathing of water. Such a process has been described for a few species of turtles which have an especially vascular pharynx, but whether an important amount of O₂ is thus derived has not yet been ascertained. This ventilation with water may actually subserve olfactory and gustatory senses.

Thoracoabdominal breathing. Pulmonary ventilation in reptiles utilizes a more familiar process than buccopharyngeal breathing; that is, it depends upon an aspiration or sucking inspiration, such as in birds and humans. This involves development of a movable rib basket and elaboration of the intercostal musculature which, by enlargement of the body cavity, produces on inspiration a negative pressure in the lungs (intrapulmonary) with reference to the atmosphere. The abdominal muscles and myoelastic tissue of the lungs and air sacs remain, as in amphibians, important in expiration, but these are augmented by striated muscular membranes which form diaphragms and also ensheath the lungs in some turtles.

Turtles are exceptional among reptiles and air breathers generally because the ribs are fused into a shell which prohibits expansion of the body wall for inspiration. In these animals, muscular membranes which enclose the viscera and others which form diaphragms at the leg pockets in the shell produce expiratory and inspiratory force, respectively.

Reptiles. In reptiles generally, when at rest, breathing cycles occur in groups which are interspersed among long intervals of apnea. During apnea the glottis is closed, the lung air is under a few millimeters of mercury positive pressure, and buccopharyngeal ventilation waxes and wanes to a degree associated at least in part with the extent of sensory disturbance. This buccopharyngeal activity resembles olfactory sniffing, as seen in dogs. The reptilian cycle of pulmonary inspiration and expiration is much simpler than the frog's, because the glottis and nares remain open while breathing movements occur, and air moves freely between lung and atmosphere in direct response to the action of breathing muscles in trunk and viscera. These muscles act alternately to enlarge and reduce the body cavity, exerting changes in pressure through tissue fluids directly on the lung air itself. Inspiration is clearly by suction. This is in contrast to the injection action of the buccopharyngeal muscles in amphibians. Like amphibians, however, the breathing remains periodic, and the glottis is normally closed.

Birds and mammals. Birds and mammals utilize strictly thoracoabdominal breathing to ventilate both nasal

and pharyngeal cavities and lungs. The same structures used in reptiles continue to operate, namely, trunk musculature and diaphragms. The glottis does not close the lung from atmosphere normally, however; nasal valves are absent, and buccopharyngeal movements cease. Because the lung and airway are greatly elaborated over those in previously considered classes of animals, certain central regions develop more critical and specific control than glottal valves over airflow and diffusion. Such control regions involve, in mammals, the conducting bronchi and bronchioles and the valvelike sphincters at alveolar sac openings; and in birds, the corresponding ducts and orifices, including parabronchi and air capillaries. Also, other discrete structures develop as bellows, to change breathing forces into ventilation pressures. These structures are the extrapulmonary air sacs in birds (air sacs also occur in some reptiles) and the distal ducts and alveolar sacs in mammals. The alveolus becomes more strictly a diffusion exchange unit. As a result of all these specializations, great stability of physical conditions at diffusion surfaces is achieved in a system which, at the same time, accommodates high and variable rates of exchange.

The muscles of bronchioles, like those of arterioles in the circulatory system, are strategically situated to modify the distribution of inspired air in lungs by affecting duct caliber. However, not only distribution but also direction of airflow is altered in the bronchi of birds, a fact long obscured by emphasis on the simple, uniform mammalian tidal pattern.

It is now well established for breathing in birds that the tidal flow of air in and out of the conducting airways between atmosphere and air sacs is interrupted in certain bronchi leading to the lung itself; the flow is changed from tidal to unidirectional before reaching most air capillaries. A detailed explanation of this change of flow is still obscure, but the advantages of the change are apparent, namely, (1) continual exposure of respiratory surfaces to atmospheric air and the highest available O_2 concentration; concomitant with (2) optimum conditions for diffusion control of H_2O and CO_2 in pulmonary fluids. Birds thus surpass mammals in physiological access to oxygen.

There are other elements of breathing besides respiratory exchange which require control of the air movement itself as a mass flow along pressure differentials, as distinguished from control of movement in molecular migration along diffusion gradients. For example, whenever panting or phonation affects breathing, or when pulmonary pathology occurs, the amount of air going to different parts of the lung might greatly affect underlying respiratory exchange and disrupt diffusion control. But homeostasis requires a stable gaseous composition within alveoli and a minimal evaporation of water from alveolar surfaces; and diseased or injured lung regions require immobilization. Consequently, panting thermoregulation depends primarily on conduction and evaporation only from the upper air passages, and also from peripheral air sacs in birds. The cooling fluid is provided mainly through controlled secretion from serous membranes or glands, such as salivary

glands in dogs. The membranes of passages and air sacs are not adapted for respiratory exchange, and insignificant amounts of O_2 or CO_2 diffuse across them. Phonation likewise uses the segregated air capacity of the ducts and sacs to move large volumes and produce sounds without affecting gaseous molecular exchange.

The structural differences between the respiratory units of bird and mammal are subordinate to their common functional characteristics. The alveolus of mammals is usually a terminal membranous pouch with a porelike orifice (the postmortem diameter is about 70 micrometers in the rabbit) into an alveolar sac; occasionally they are appendant on a respiratory bronchiole. The corresponding structure in birds, called an air capillary or cylindrical alveolus, is an appendant membranous tubule with an orifice (diameter about 50 μm in the chicken) into a tubular parabronchus. These alveolar structures frequently interconnect, via pores in mammals and anastomoses in birds. They both provide, in the aggregate, a reservoir with tremendous surface area (at least 50 m^2 in humans) which sequesters a mechanically stable intra-alveolar atmosphere based on nitrogen. It is through this medium that large quantities of oxygen, carbon dioxide, and water vapor molecules diffuse according to pressure gradients established, as in all vertebrates, on one hand by ventilation and on the other by pulmonary circulation. This intrapulmonary atmosphere, made up of the slightly varying composition among millions of alveoli, is often called alveolar air. It stays remarkably uniform in total composition despite ten- to twentyfold variations in oxygen and carbon dioxide exchange, such as occur during exercise. Some standard resting values for ventilation and diffusion in respiratory exchange of humans are shown in Tables 3–5.

The microscopic size of alveoli exaggerates a persistent tendency of all such bubblelike or balloon-like structures to collapse. This tendency in alveoli is partly the result of the surface tension of a fluid film which bathes their gas-exposed surface and partly the result of the elasticity of stretched fibers widely distributed within pulmonary tissues. Actual collapse (atelectasis) will occur at once as a result of these forces, for example, if much gas (pneumothorax) or fluid (hydrothorax) accidentally enters the pleural cavity.

A lipoprotein material is secreted onto the alveolar surface, a material characterized as surfactant, which reduces the surface tension at the gas-liquid interphase where respiratory exchange begins. This action lowers the alveolar rate of shrinkage and consequently helps to maintain an effective area of respiratory surface.

Alveolar size and effective surface area are further supported by periodic deep inspirations which are an integral part of the respiratory patterns of mammals, with the possible exception of a few very large representatives like the horse. These are called complementary cycles, or sighs, in humans, and they arise from sensory monitoring and nervous control of the functional residual volume of alveolar air.

These cycles occur about nine times an hour in humans; in other animals they occur more or less frequently, depending on the characteristic size of the adult (Table 5).   F. Harold McCutcheon

Avian respiration. The respiratory system of birds has a unique and complex structure. The lungs are relatively rigid and do not change volume significantly during ventilation; air passes through them and into a group of usually nine air sacs. There is no diaphragm separating the thoracic and abdominal cavities, and both phases of ventilation are active, even in birds at rest.

Structure. The primary bronchus originates from the trachea, runs along the entire length of the lung, and divides into two groups of secondary bronchi (**Fig. 27***a*). As the two groups of secondary bronchi enter the lung, they divide into four branches that spread across the medial and ventral surface of the lung. The primary bronchus also forms branches that extend along the lateral and dorsal surface and join the secondary bronchi by means of straight tubes (tertiary bronchi).

This network system of secondary and tertiary bronchi is known as the paleopulmo system, and it is present in all birds (Fig. 27*a*). In the penguins and emus, it is the only system, whereas in all other birds there is also a tertiary bronchial network between the lateral side of the primary bronchus, the laterobronchi, and the posterior air sacs. This network system is known as the neopulmo system (Fig. 27*b*). A meshwork of fine air capillaries, which intertwine with equally fine blood capillaries, radiates from the lumen of each tertiary bronchus. These blood capillaries arise from arterioles and form collecting venules near the tertiary bronchial lumen. Gas exchange occurs in the air capillaries leading from the tertiary bronchi.

The cervical air sacs arise from the first medioventral bronchus of each lung, and the (single) clavicular air sac is joined to the first, second, and third medioventral bronchi of both lungs, whereas the cranial thoracic sacs connect predominantly to the third medioventral bronchus of each lung. These form the cranial (anterior) group of air sacs. The second lateroventral bronchi of each lung form large connections to the caudal thoracic air sac, and the primary bronchi continue to the caudal edge of each lung and empty into the abdominal air sacs. These latter two pairs of air sacs form the caudal (posterior) group.

Pathway of air. Airflow through the tertiary bronchi of the paleopulmo system is in the same direction during both phases of ventilation (**Fig. 28**). During inspiration, some air travels through the primary and tertiary bronchi of the neopulmo system to the posterior group of air sacs, while the remaining air travels through the mediodorsal secondary bronchi and the tertiary bronchi of the paleopulmo system to the anterior group of air sacs. During expiration, the gas in the posterior air sacs travels again through the tertiary bronchi of the neopulmo system, then into the mediodorsal secondary bronchi, through the tertiary bronchi of the paleopulmo system, and into the medioventral secondary bronchi where, together with gas from the anterior group of air sacs, it enters the primary bronchus and trachea.

There is no evidence of any mechanical valves in the airways, and the unidirectional airflow through the tertiary bronchi of the paleopulmo system is thought to result from the aerodynamic conditions in the lung. During inspiration, the velocity of airflow is important; the greater the velocity, the more likely is the inspired gas to flow pass the medioventral bronchi and continue along the primary bronchus. There are indications that the diameter of the primary bronchus decreases close to its junction with the first medioventral bronchus. This region is called the segmentum accelerans because it causes an increase in the velocity of airflow at that point.

Functional implications. Bulk airflow through the tertiary bronchi of the paleopulmo system is in the same direction during both phases of ventilation, and it has been demonstrated that bulk blood flow through the arterioles in the peritertiary bronchial tissue is at right angles to the bulk airflow. This is known as a cross-current arrangement, and it is more effective at exchanging gases (CO_2, in particular) than is the mammalian lung. It has been suggested that the unique structure of the avian lung is an adaptation to the high oxygen requirements of flight and an explanation of the better ability of birds to tolerate the hypoxia of high altitude compared with mammals. However, the typically mammalian lungs of bats have equally high mass-related oxygen consumptions during flight, and although there are indications that the
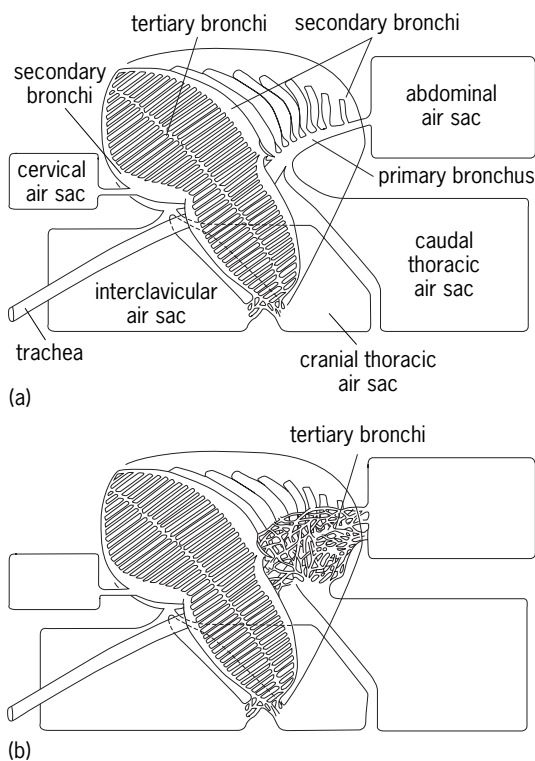
tertiary bronchi   secondary bronchi

secondary bronchi

abdominal air sac

cervical air sac

primary bronchus

caudal thoracic air sac

interclavicular air sac

trachea

cranial thoracic air sac

(a)

tertiary bronchi

(b)

Fig. 27. Overall structure of the lung of birds (*a*) without and (*b*) with the neopulmo system (tertiary bronchi), showing the connections to the air sacs. (*After H.-R. Duncker, Structure of the avian respiratory tract, Respir. Physiol., 22:1–19, 1974*)
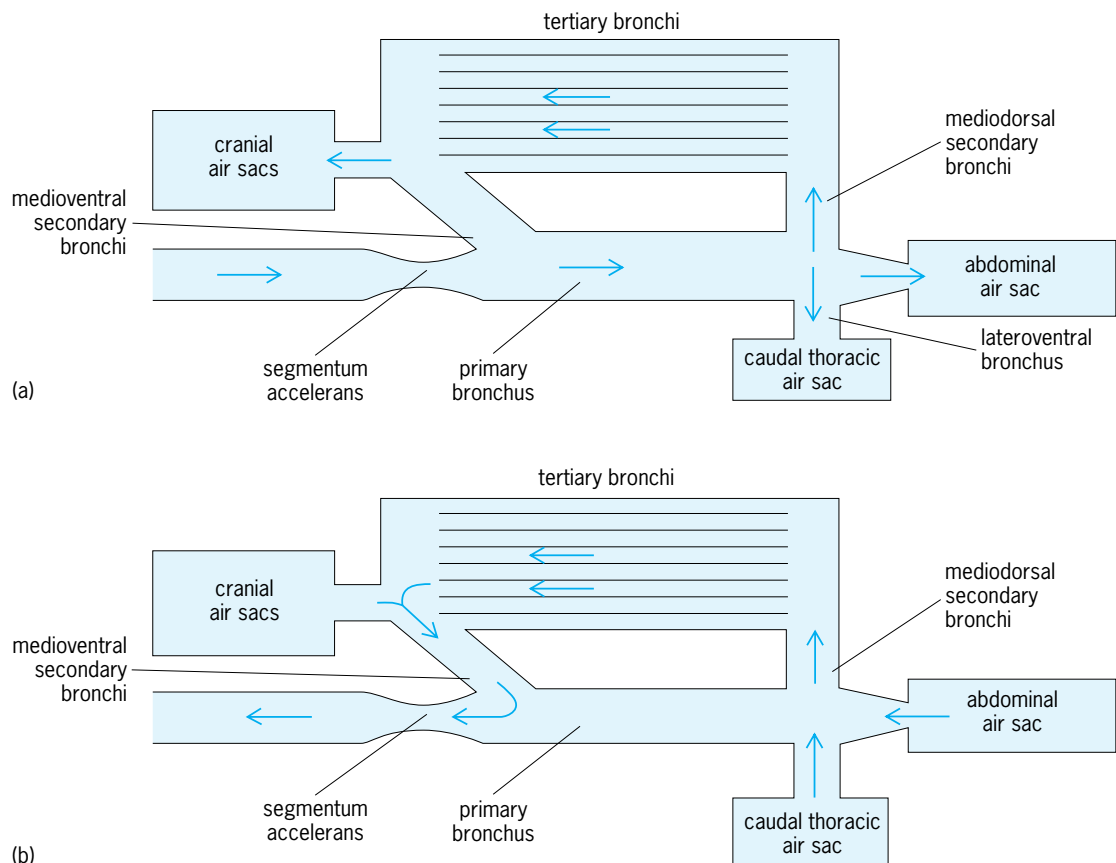
Fig. 28. Pathway of air through the lung of a bird. (*a*) Inspiration. (*b*) Expiration. (*After R. B. Banzett et al., Pressure profiles show features essential to aerodynamic valving in geese, Respir. Physiol., 84:295–309, 1991*)

lungs of birds confer some advantage during hypoxia compared to that of mammals, this decreases at the highest altitudes.   P. J. Butler

Mammalian respiration. Ventilation in mammals is more nearly like that in reptiles than in birds. In mammals, alveoli are terminal sacculations in a serial arrangement of ventilatory ducts and sacs, rather than appendant tubules along a parallel duct arrangement as in birds. The only nondiffusion regions of mammalian lungs are the major airways down through the terminal bronchioles; there are no ventilatory sacs in the pulmonary system. However, the alveolar sacs of mammals may be considered the ventilatory analogs of the air sacs of birds, whereas the alveoli themselves incorporate most of the functionally specialized exchange surface. Alveolar sacs histologically are diffusion structures, unlike avian air sacs; however, the size and structural location relative to the alveoli indicate that they and not the alveoli accommodate much of the lung volume change during eupnea. Thus enlargement of the lung during inspiration brings air down to the mouths of alveoli, including a small remainder of expiratory air in the ducts from the preceding breath. This unavoidable dilution of inspired air is like that found in the neopulmo of birds. But unlike the bird, there is no counterpart of a paleopulmo system to provide undiluted atmosphere to any part of the exchange surface. Consequently, the pulmonary system of mammals has H_2O and CO_2 diffusion control like birds, but it lacks the O_2 supply efficiency. The important question of the specific intrapulmonary distribution of air in mammals is under intensive study by histological as well as physiological techniques. A major organ of ventilation found only in mammals is the muscular diaphragm which divides thorax and abdomen. Because it lies as a dome with its convex face toward the thorax, when it contracts and flattens it augments thoracic muscles in increasing the capacity of the thoracic cavity. The abdominal wall relaxes at the same time to accommodate viscera which are displaced by the diaphragm. As a result of these movements, air under pressure differential of a few millimeters of mercury (for example, -1.5 mmHg in the nasal cavity of the horse) passes into the pulmonary system. The abdominal muscles cannot contribute directly to inspiration in mammals or any other vertebrates except turtles, in which a special arrangement is associated with the shell as previously mentioned.

Breathing mechanism. It is in the nature of muscular membranes that they must be oriented as the mammalian diaphragm is; that is, they must insert along their periphery into a resistant structure and bulge into the cavity if contraction is to produce negative pressure in the cavity. A positive expiratory force is a different matter, however, and abdominal muscles contribute such force in all vertebrates. The extent varies with the degree of ventilation,

a much greater contribution in exercise than at rest; with the species; and even with the sex. In the human female during eupneic breathing, for example, costal movement predominates and expiration results largely from the passive recoil of lung and chest, whereas in males there is somewhat more abdominal involvement. In mammals generally, abdominal muscles are more involved in expiration than is true for humans, and in the larger quadrupeds the work of displacing heavy pendant viscera requires their continuous activity. Forced breathing (hyperpnea) with increased amplitude, as during exercise; labored breathing (dyspnea), as during strenuous exercise or at high altitude; and compressatory acts all require the abdominal muscles. These supplement the internal intercostal muscles which pull the ribs to resting position. On the other hand, they are little involved in panting (polypnea) or sniffing.

Myoelastic fibers. Myoelastic fibers of the lung itself not only provide a passive component of expiration, but also account for collapse of the lung if the chest cavity is opened to the atmosphere. Because from the first filling at birth these fibers are stretched during the entire life of the animal, the surface of the lung always tends to recoil from adjacent structures. This recoil is limited in birds because the lung is structurally attached, but in mammals and all other animals the lung surface is free from attachment. The visceral pleura of mammals, a covering membrane, adheres to the lungs, but this is in turn separated from all adjacent structures, which are covered by a parietal pleura, only by a film of mucoid fluid. Measurement of the elastic pulling force exerted by the lung against the surface tension of this fluid reveals the equivalent of a pressure which is negative with respect to atmosphere, when the animal is at rest or during eupnea. It is negative with respect to intrapulmonic pressure at all times. This is called intrapleural pressure (Table 5).

Pressure change and ventilation. The relationships between various pressure changes and ventilation activity in mammalian breathing are illustrated in **Fig. 29**. Here are shown tracings from instruments which recorded abdominal movements in a horse during one breathing cycle, along with concomitant pressure changes in the designated areas. The shapes of such graphs vary in detail among different species of mammals, as can be expected from a comparison of some ventilation parameters for a few animals (Table 5). The forces of respiratory exchange are apparent from such measurements as those in Table 3, made on air and blood entering, within, and leaving the lungs of humans.

Vertebrate respiratory regulation. Ventilation and circulation through the lungs are controlled and adjusted to maintain an efficient exchange of gases in relationship to the needs of the animal and to the suitability of the environment. Sensory units (tension receptors) for pressure or distension occur in lung and breathing elements and in related circulatory organs. These signal the medulla oblongata of the brain, from which the breathing rhythm itself originates, as well as other parts of the brain, and breathing is
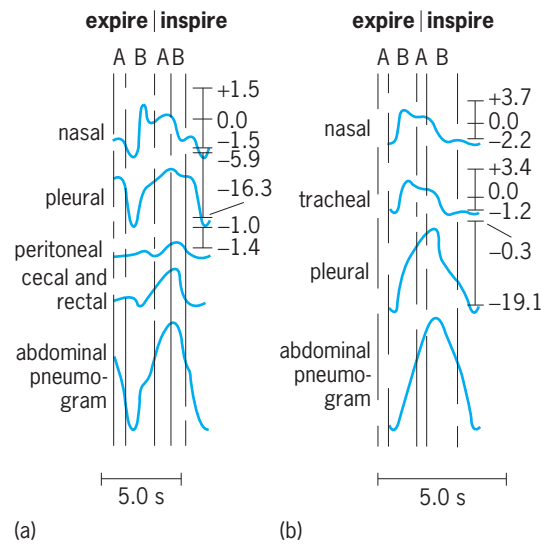

Fig. 29. Composite diagram of the factors in the breathing cycle of the horse, analyzed to show the correlation of intracoelomic and intrapassage pressures with the external musculature. (*a*) Eupnea. (*b*) Hyperpnea. Numbers indicate mmHg.

adjusted to physiological changes and physical conditions. Other receptors, especially in certain major blood vessels near the heart and in the medulla itself, signal the respiratory areas of the brain about blood chemical conditions, in particular, about oxygen, carbon dioxide, and acidity. Signals from other parts of the body, such as the olfactory and tracheal membranes and the joints, and from other parts of the brain, such as heat-regulating, olfactory, and phonation areas, also reach respiratory areas. Here the correlating and integrating role of the brain determines how breathing will proceed: whether to speed up, stop, cough, or to do whatever is appropriate. Oxygen tension in blood is a basic regulatory element for ventilation rate in lower vertebrates, but it becomes secondary to carbon dioxide in birds and mammals. Because carbon dioxide is a major factor in blood acidity, ventilation also comes to play a major role in critical regulation of hydrogen ion concentration (blood acidity) in higher animals. *See* CAROTID BODY.

Respiration in aquatic animals. The critical need for oxygen in amounts available only by continuous air breathing becomes more pronounced the higher an animal is in the evolutionary scale. This reflects the increasing energy requirements of such biological advantages as rapid sustained locomotion and critical continuous neuroregulatory activity. Through such attributes animals gain increasing independence from environmental limitations. Because there is about 30 times less oxygen in an aquatic environment, volume for volume, than in the atmosphere above it, no air-breathing animal except a few fish and most amphibians can maintain vital processes with oxygen gained by aquatic respiration. In amphibians the skin is free from such protective structures as scales, corneum, and hair; thus diffusion is much less restricted than in other animals, and the skin is an important respiratory

organ in water as well as in air. Permeability to respiratory gases includes permeability to water, however, and this restricts even terrestrial amphibians, such as toads, to a very humid atmosphere. Some aquatic amphibians, such as the mud puppy (*Necturus*), possess lungs as well as gills, but no such animals are capable of the high energy output which characterizes each of the strictly terrestrial and air-breathing vertebrate classes: the reptiles, birds, and mammals. Even these classes do include some species which are partly and some strictly aquatic in habitat. All the aquatic species remain strictly air-breathing, but they have modified respiratory and circulatory processes which accommodate diving and underwater activity during long suspension of breathing.

Suspension of breathing. In aquatic reptiles, such as snapping turtles and alligators, breathing suspension for dives is simply a special instance of the apnea which characterizes normal periodic breathing of reptiles on land. In diving birds and mammals, however, the suspension is a departure from the eupneic pattern, for eupnea is normally interrupted only at rest by an occasional deep breath and a pause of about 5 s duration, such as the sigh in humans. A dive of more than 2–3 min duration in birds and mammals requires special adaptations, because no animal has means of storing oxygen to last more than a few minutes; however, seals may submerge for 15–25 min and whales for 1 h or more. Important means of submergence in such animals include (1) restriction of blood flow from the bulk of the muscles, thus sequestering lactic acid and building up a debt for oxygen to be paid on access to air; (2) higher tolerance for carbon dioxide, which is always toxic in excess, and which must accumulate until breathing is resumed; (3) conservation of movement while submerged; (4) reflexes around nostrils which stop breathing on contact with water; and (5) great reduction in lung volume when the high pressure of great depths is involved, which reduces diffusion area of the lung. The last of these does not concern respiration itself; rather it serves to limit nitrogen buildup in tissues, with its attendant hazard of bubble formation, or bends. Such bubbles may form on too rapid ascent from depths, and they can cause immobilization and death. Also, lungs and air sacs may serve as buoyancy organs for swimming and diving, because their contents can be altered in volume to change the displacement of the animal.

Adaptations in terrestrial animals. The adaptations for diving in birds and mammals have been found in counterpart among some nonaquatic animals such as rabbits, which may become immobile and suppress breathing under certain conditions, for example, in evading predators. For active locomotion in air and for highest brain activity, however, oxygen is the most urgently needed of all substances required from the environment; therefore, essentially continuous ventilation is a necessity for strictly terrestrial animals such as the chicken, cat, and human. Such animals survive without actual ventilation only before the time of birth, or hatching, although nonrespira-

tory periodic breathing movements do occur in egg or uterus as appropriate structures develop. Early respiration utilizes special diffusion respiratory structures, such as the allantois and placenta. In birds an actual pulmonary ventilation, utilizing the air chamber of the egg, begins shortly before hatching. In mammals it can begin only when the fetus gains direct access to air at birth. *See* PLACENTATION; RESPIRATORY SYSTEM DISORDERS.   F. Harold McCutcheon

Bibliography. W. Bloom and D. W. Fawcett, *A Textbook of Histology*, 11th ed., 1986; J. N. Cameron, *The Respiratory Physiology of Animals*, 1989; R. M. Cherniack (ed.), *Lung Disease: State of the Art, 1988–1989*, 1990; P. Dejours et al. (eds.), *Comparative Physiology: Life in Water and on Land*, 1987; J. Graham, *Air-Breathing Fishes*, 1997; W. S. Hoar and D. J. Randall (eds.), *Fish Physiology*, vol. 10A and B: *Gills*, 1984; D. J. Randall, The control of circulation in fish during exercise and hypoxia, *J. Exp. Biol.*, 100:275–288, 1982; I. V. Seller (ed.), *Bird Respiration*, vol. 1, 1987; N. B. Slonim, *Respiratory Physiology*, 5th ed., 1987; P. D. Sturkie (ed.), *Avian Physiology*, 4th ed., 1986.

Respiratory system disorders

Dysfunction of the respiratory system, which supplies the body with the oxygen needed for metabolic activities in the cells and removes carbon dioxide, a product of cellular metabolism. For this purpose the right ventricle of the heart pumps venous blood, containing carbon dioxide, through the pulmonary arteries into the lung capillaries, where gaseous exchanges between the blood and air occur. The arterial or oxygenated blood is then drained through the pulmonary veins into the left atrium. *See* RESPIRATION.

The respiratory system includes the nose, mouth, throat, larynx, trachea, bronchi, lungs, and the muscles of respiration such as the intercostal muscles and the diaphragm. The inspiratory movements of the chest wall and diaphragm suck air through the nose, throat, larynx, trachea, and bronchi into the lung alveoli, where it comes into close contact with the blood circulating in the capillaries and gas exchange occurs. During expiration this air is expelled through the same pathway.

Symptoms and signs. The lung has a great reserve capacity, and therefore a significant amount of disease usually must be present to produce clinical signs and symptoms. Shortness of breath (dyspnea) on exertion is the most common symptom of a respiratory disorder. Shortness of breath while at rest is indicative of severe respiratory disease and usually implies a severe abnormality of the lung tissue. If the respiratory system is so diseased that normal oxygenation of the blood cannot occur, blood remains dark, and a bluish color can be seen in the lips or under the fingernails; this condition is referred to as cyanosis. If the process continues and insufficient oxygen is supplied to vital organs such as the brain and heart, a person will become unconscious and die. Other

signs and symptoms of respiratory disorder can include fever, chest pain, coughing, excess sputum production, and hemoptysis (coughing up blood). Most of these signs and symptoms are nonspecific. For example, hemoptysis can result from various conditions, though in smokers it is commonly a sign of lung cancer. *See* HYPOXIA.

Diseases of air passages. Most diseases of the airways increase the resistance against which air is sucked in and pushed out of the lungs. The effect of a disease of the airways depends largely on its localization. Diseases of the nose usually have little influence since collateral respiration through the mouth compensates easily. Diseases of the throat, larynx, and trachea can significantly inhibit the flow of air into the lungs. Infections in the back of the throat, such as in diphtheria, can cause marked swelling of mucous membranes, resulting in air obstruction. Edema (swelling) of the mucosal lining of the larynx can also cause a reduction in air flow. Likewise, air flow can be inhibited in asthma, in which the smooth muscle in the trachea and bronchi episodically constricts. Chronic bronchitis, which is caused predominantly by cigarette smoking, results in inflammation of and excess mucus production by the bronchi and this also can lead to a reduction in air flow. Bronchiolitis, a condition that usually occurs in children and is often caused by a respiratory virus, results in narrowing and inflammation of small airways and a decrease in air flow.

Diseases of the lung. Pneumonia, cancer, and emphysema are the most common lung diseases and are a major cause of morbidity and mortality in the United States. Pneumonia is caused by bacteria, viruses, and other infectious agents, that result in inflammation and consolidation (solidification) of the gas exchange units of the lung (alveoli). Persons with pneumonia frequently develop fever, chest pain, and rapid breathing. When the disease involves a significant portion of the lung, shortness of breath may develop. *See* PNEUMONIA.

There are four main types of pulmonary emphysema. The most common is centrilobular emphysema, and it is caused by cigarette smoke. This disease usually involves the upper lobes of the lungs and is characterized by the breakdown and permanent loss of lung tissue, resulting in large cystic spaces in the lung referred to as blebs. Emphysema is usually associated with chronic bronchitis, and individuals with the combination of diseases will often have shortness of breath on exertion and usually have excess sputum production. *See* EMPHYSEMA.

Of the four major types of lung cancer, approximately 90% can be attributed to the carcinogens present in cigarette smoke. Lung cancer may be detected in asymptomatic persons with a routine chest x-ray, or it may be discovered because of pain, excess coughing, or hemoptysis. *See* CANCER (MEDICINE).

Pulmonary circulatory diseases. Among the diseases of pulmonary circulation, congenital malformations of the heart and pulmonary artery account for many cases of respiratory insufficiency in newborn and younger children; those affected are called blue babies because they appear cyanotic. In adults, acquired heart diseases, such as narrowing of the valve between left atrium and left ventricle (mitral stenosis), can cause backup of blood into the lung and an increased pressure in the pulmonary circulatory system. Also, blood clots, which usually develop first in the deep veins of the legs, can break free and flow to the heart. There they enter the pulmonary arteries and wedge in their small branches, where the clots are referred to as pulmonary thromboemboli, and can cause areas of death in the lung tissue (pulmonary infarcts). Persons who develop thromboemboli usually have chest pain and shortness of breath, and some have hemoptysis. *See* CARDIOVASCULAR SYSTEM; CIRCULATION DISORDERS.

Neuromuscular diseases. Some neuromuscular diseases, such as poliomyelitis and amyotrophic lateral sclerosis (Lou Gehrig's disease), can cause dysfunction of the muscles of respiration. The resulting inability of these muscles to move air into the lungs can cause severe shortness of breath and predispose the patient to pneumonia. Artificially assisted ventilation may be necessary for breathing. *See* MUSCULAR SYSTEM DISORDERS; POLIOMYELITIS; RESPIRATORY SYSTEM.   Samuel P. Hammar

Bibliography. D. H. Dail and S. P. Hammar (eds.), *Pulmonary Pathology*, 2d ed., 1994; A. P. Fishman, *Pulmonary Diseases and Disorders*, 3d ed., 1998.

Response

The reaction or result produced by applying an input signal to a system. Given a system, or interconnection of devices to perform a desired function, the signals processed by that system can be divided into several categories, some of which are referred to as causes, stimuli, or input signals; and others, as responses, effects, or output signals. Although there can be multiple input and output signals for a given system, attention will be focused on single-input single-output systems. Often, the input and output signals of interest are functions of time only, although the discussions given can be generalized to more than one independent variable, such as space and time. Also, attention will be limited to continuous-time signals and systems, although the discussion can be easily put in terms of discrete-time signals and systems.

Linear, time-invariant systems. It is convenient to represent the relationship between stimulus and response of a system as Eq. (1), which is read "$y(t)$ is

$$y(t) = \mathcal{H}[x(t)] \qquad (1)$$

the response of $\mathcal{H}$ to the input $x(t)$," where initial conditions are included if pertinent. The symbol $\mathcal{H}$ serves the dual role of identifying the system and specifying the operation to be performed on $x(t)$ to produce $y(t)$.

It is difficult to say anything explicit about the response of a general system, as described by Eq. (1), and usually attention is limited to linear, time-invariant systems. A system is linear if an arbitrary

linear combination of two inputs produces the same linear combination of the outputs to each input applied separately. A system is time-invariant if the response to an input delayed by a constant amount is the response to the undelayed output delayed by the same amount; that is, if $x(t)$ in Eq. (1) is replaced by $x(t - \tau)$, where τ is a constant, the output is $y(t - \tau)$. *See* LINEAR SYSTEM ANALYSIS; LINEARITY.

Complete, free, and forced responses. Many linear, time-invariant systems can be described by constant-coefficient, linear differential equations. When a system is described in this way, the derivative of highest degree on $y(t)$ in the differential equation is called the order of the system. It is well known from the theory of ordinary differential equations that the solution to a constant-coefficient, linear differential equation consists of two parts: the complementary and particular solutions. The complementary solution is found by setting $x(t) = 0$ and solving for the roots of the characteristic equation (that is, the equation which results by assuming that $y(t) = e^{pt}$, and canceling the common factor e^{pt} after differentiation and substitution). The complementary solution, also known as the homogeneous solution, is then a sum of n exponential functions, if the n roots to the characteristic equation are distinct. Since this solution results by setting the input to zero, it represents the natural behavior of the system in response to an initial stored energy when the input is removed. Thus the complementary solution is also called the free response of the system. The other part of the solution, the particular integral, is of the same form as the input and can be determined by the method of undetermined coefficients. Since it depends on the form of the forcing function, or $x(t)$, the particular solution is also called the forced response of the system. The complete response of the system is the sum of the homogeneous solution and particular integral of the describing differential equation of the system, or the sum of the free and forced responses. *See* DIFFERENTIAL EQUATION.

Gain and phase response. It is useful to consider the forced response of a system described by a constant-coefficient, linear differential equation to an input of the form $e^{j\omega t} = \cos \omega t + j \sin \omega t$, where j is the square root of -1, and ω is a parameter, called the radian frequency of the input, which is to be varied later. The forced response must be of the same form as the input, so it is assumed that $y(t) = Ae^{j(\omega t+\theta)}$. After differentiation of both the input and assumed response, substitution, and cancellation of the factor $e^{j\omega t}$ on both sides of the equation, Eq. (2) is obtained,

$$\left. \frac{y(t)}{x(t)} \right|_{x(t)=e^{j\omega t}} = H(j\omega) \qquad (2)$$

where $H(j\omega)$ is a rational function of $j\omega$. The function $H(j\omega)$, called the transfer function or frequency response, of the system, is a complex function of the variable ω, which is the radian frequency of the sinusoidal input. (Although a complex exponential was assumed, the response to $\cos \omega t$ of the system can be obtained by multiplying $H(j\omega)$ by $e^{j\omega t}$ and

taking the real part.) Writing the complex function $H(j\omega)$ in terms of its magnitude and complex argument yields Eq. (3), where $A(j\omega)$ is called the ampli-

$$H(j\omega) = A(j\omega)e^{j\theta \, (j\omega)} \qquad (3)$$

tude response, or gain, and $\theta(j\omega)$ is called the phase response. *See* GAIN; PHASE (PERIODIC PHENOMENA).

These are important functions for a number of reasons. The amplitude response shows how the sinusoidal frequencies that are used to resolve an input (that is, its spectral components) are changed in amplitude, or filtered, in going through a system. Filter theory deals with how to synthesize transfer functions that give certain desired filter characteristics such as low-pass, band-pass, or high-pass. The phase response can be used to obtain the group delay of the system (group delay is the negative derivative of the phase response), which can be used to determine how the various spectral components of the input are delayed with respect to one another. For no distortion of the input, the system should impose the same gain and delay on all spectral components that compose the input. In the field of control theory, the frequency response of the open loop of a feedback control system plays an important role in determining system stability. It can be shown that oscillations within the closed loop are possible if the open-loop phase response is an odd multiple of $180°$ at a frequency or frequencies and the closed-loop gain is unity at any of these frequencies. The amounts by which phase and gain of the open-loop system do not meet these criteria are called the phase and gain margins, respectively. *See* CONTROL SYSTEM STABILITY; CONTROL SYSTEMS; ELECTRIC FILTER.

Superposition integrals. The impulse response is of special interest in the description of a linear time-invariant system. For such a system, it is the response to a unit impulse function applied as the input at time zero. The impulse response provides a means to obtain an explicit expression for the response in terms of the input via the superposition integrals, as opposed to the implicit relationship given in a differential equation.   Rodger E. Ziemer

Bibliography. C. M. Close and D. K. Frederick, *Modeling and Analysis of Dynamic Systems*, 3d ed., 2001; E. Kreyszig, *Advanced Engineering Mathematics*, 8th ed., 1998; R. E. Ziemer, W. H. Tranter, and D. R. Fannin, *Signals and Systems: Continuous and Discrete*, 4th ed., 1998.

Rest mass

A constant intrinsic to a body which determines its inertial and energy-momentum properties. It is a fundamental concept of special relativity, and in particular it determines the internal energy content of a body. It is the same as the inertial mass of classical mechanics. According to the principle of equivalence, the basic physical principle of general relativity, the inertial mass of a body is also equal to its gravitational

mass. *See* CLASSICAL MECHANICS; GRAVITATION; RELATIVISTIC MECHANICS; RELATIVITY.

Newton's law. The rest mass or inertial mass of a body, m, is a measure of its resistance to being accelerated at **a** by a force **F**; in classical mechanics the relation between inertial mass, acceleration, and force is given by Newton's law, Eq. (1). In special rel-

$$\mathbf{F} = m\mathbf{a} \tag{1}$$

ativity Newton's law holds exactly only in the body's rest frame, that is, the frame in which the body is instantaneously at rest. *See* NEWTON'S LAWS OF MOTION.

Energy-momentum. Rest mass serves to define the energy-momentum of a body. In special relativity the trajectory of a body is specified by the position 4-vector, given in Eq. (2), where t denotes time, c

$$(x^0, x^1, x^2, x^3) = (ct, x, y, z) \tag{2}$$

is the speed of lights, and (x,y,z) is the body's position in cartesian coordinates. The 4-velocity is the derivative of the position 4-vector with respect to the proper time, that is, the time measured in the body's rest frame. A body's energy-momentum 4-vector is defined as the product of its 4-velocity and its rest mass. *See* VECTOR METHODS (PHYSICS).

This 4-vector is the relativistic generalization of both the energy and vector momentum of classical mechanics, with the zeroth component, p^0, being the energy divided by c, and the components 1 to 3 being the momentum vector. Thereby, the separate conservation laws for energy and momentum in classical mechanics are combined into one conservation law for the energy-momentum 4-vector in special relativity. *See* CONSERVATION LAWS (PHYSICS); CONSERVATION OF ENERGY; CONSERVATION OF MOMENTUM; ENERGY; MOMENTUM.

Rest energy. Associated with the rest mass of a body, there is an internal or rest energy, which follows from the above definition. In the system where the body is at rest, the zeroth component of the velocity 4-vector is equal to c since the proper time and the laboratory time, t, are then the same. The energy of the body is given by Eq. (3).

$$p^0 c = E = mc^2 \qquad \text{(body at rest)} \tag{3}$$

A more physical explanation of this mass-energy equivalence based on the Doppler-shift relation of special relativity, follows that of A. Einstein. If a slowly moving body with velocity v emits one light pulse in the forward direction and one in the backward direction, each of energy $\Delta E/2$ as seen in its rest frame, then by symmetry the body does not change velocity, but it does lose energy, ΔE, in its rest frame. However, as seen in the laboratory frame, the light pulses will be Doppler-shifted and will carry a larger total energy, $\Delta E(1 - v^2/c^2)^{-1/2}$, so that the body's energy is there seen to decrease by this larger amount. A difference between the energies of the body in the rest and laboratory frames must correspond to a difference in kinetic energy, $mv^2/2$. Since the body's

velocity does not change as the light pulses are emitted, this decrease in kinetic energy must reflect a decrease in the body's rest mass, Δm. Comparison of kinetic energy loss, $\Delta mv^2/2$, with the difference between the energy loss in the two frames, $\Delta E[1 - (1 - v^2/c^2)^{-1/2}]$, yields Eq. (4), in the limit of low velocity, in accord with Eq. (3).

$$\Delta m = \frac{\Delta E}{c^2} \tag{4}$$

Reactions. The experimental realization of the interconversion of mass and energy is accomplished in the reactions of nuclei and elementary particles. In particular, the energy source of nuclear bombs and nuclear fission reactors is a small decrease in the total mass of the interacting nuclei, which gives rise to a large energy release because of the large numerical value of c^2. *See* ELEMENTARY PARTICLE; NUCLEAR FISSION.   Ronald J. Adler

Bibliography. H. A. Lorentz et al., *The Principle of Relativity*, 1923; H. M. Schwartz, *Introduction to Special Relativity,* 1968, reprint 1977.

Restionales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Commelinidae of the class Liliopsida (monocotyledons). The order consists of 4 families and about 450 species, some 400 of them belonging to the Restionaceae. The vast majority of the species grow in temperate regions of the Southern Hemisphere. The Restionales are wind- or self-pollinated, with reduced flowers and a single, pendulous, orthotropous ovule in each of the 1–3 locules of the ovary. The Restionaceae are much like the Gramineae in the related order Cyperales and are sometimes referred to as Southern Hemisphere grasses. *See* COMMELINIDAE; CYPERALES; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM.
                          Arthur Cronquist; T. M. Barkley

Restoration ecology

A field in the science of conservation that is concerned with the application of ecological principles to restoring degraded, derelict, or fragmented ecosystems. The primary goal of restoration ecology (also known as ecological restoration) is to return a community or ecosystem to a condition similar in ecological structure, function, or both, to that existing prior to site disturbance or degradation. *See* ECOLOGY.

Most restorationists trace their discipline to Aldo Leopold, a professor at the University of Wisconsin who began the first restoration project in the United States in 1935. Working with the Civilian Conservation Corps, a federal relief work program during the Dust Bowl era, Leopold replanted 24 hectares (60 acres) of North American tallgrass prairie, a formerly widespread ecosystem that to this day remains highly fragmented and endangered. By working to

restore this species-rich grassland, Leopold established a national conservation ethic that still serves as the basis of this ecological discipline. Restoration ecology today, however, involves more than replanting native species. It is a highly synthetic discipline that draws upon older ecological fields such as applied ecology and wildlife management, and newer ones such as invasion biology and landscape ecology. *See* CONSERVATION OF RESOURCES; ECOLOGY, APPLIED; ECOSYSTEM.

Ecologists engage in restoration activities for different reasons. For example, wetlands are restored as a consequence of legal mandates, such as the federal Clean Water Act. This legislation requires the restoration or creation of wetlands to mitigate the unavoidable impacts of human activities such as commercial construction or habitat conversion. Other ecologists are involved in restoration-related activities because they recognize that conservationists cannot afford to be concerned only with the preservation of pristine areas. Many ecosystems have nearly disappeared altogether, and what remains is highly fragmented or degraded. The only alternative in such cases is to restore the fragmented patches that remain. Some ecologists restore specific segments of an ecosystem (for example, a rare prairie orchid or a top carnivore) to reestablish the processes that control the structure and function of the food web. *See* FOOD WEB.

Restoration projects. A reference framework is needed to guide any restoration attempt—that is, to form the basis of the design (for example, desired species composition and density) and monitoring plan (for example, setting restoration milestones and success criteria for restoration projects). Such a reference system is derived from ecological data collected from a suite of similar ecosystems in similar geomorphic settings within an appropriate biogeographic region. Typically, many sites representing a range of conditions (for example, pristine to highly degraded) are sampled, and statistical analyses of these data reveal what is possible given the initial conditions at the restoration site. Two examples of restoration deserve mention, because they represent the range of activities in which restoration ecologists are engaged: Guanacaste National Park and San Diego Bay.

Guanacaste National Park. Perhaps the most well known (and most ambitious) restoration project of the 1990s was the restoration of thousands of hectares of tropical dry forest in Guanacaste National Park in Costa Rica. The original forest cover of Guanacaste was composed of canopy trees 20–30 m in height, with many understory shrubs and vines. However, much of the dry forest ecosystem had been converted in the 1800s to pasturelands of exotic African grasses. These grasses appear to be maintained by fire, so the cessation of annual burning was key to the restoration process. Costa Rican biologists and United States scientists worked together using ecological principles and standard land management techniques to keep fire out and increase seed dispersal. Horses were used to disperse one of the dominant tree species, the Guanacaste tree

(*Enterolobium cyclocarpum*, Fabaceae), by allowing them to wander through the restoration site and deposit in manure piles the seeds they had eaten. As of 2001, a significant portion of Guanacaste National Park consisted of a secondary forest that formerly was only pastureland. Scientists believe that this restoration project continues to be successful because some dry forest patches still exist that serve as sources of restoration propagules (units of dispersal), and because the long time frame (decades) is appropriate for this kind of long-term restoration endeavor. *See* FOREST AND FORESTRY; FOREST ECOSYSTEM; REFORESTATION; LAND RECLAMATION.

San Diego Bay. Since 1984, the U.S. Fish and Wildlife Service has put considerable effort into the restoration of the wetlands in San Diego Bay, California, to provide a habitat for the light-footed clapper rail (*Rallus longirostris levipes*), a bird on the Endangered Species List. Two wetland ecosystems totaling 29 acres were constructed for the rails, but as of early 2001 rails had not nested at either site. The scientists involved in this restoration project believe that the constructed wetlands were perceptively different from natural systems in such ecological characteristics as plant height, abundance of epibenthic (living on the soil surface at the bottom of a body of water) invertebrates, and concentrations of organic matter and nitrogen in the soil. Lessons learned from this project include the sobering realization that restorationists cannot guarantee long-term success. Ecologists may not know of, much less understand, all of the ecological linkages in the ecosystem being restored. This lack of understanding can lead to the creation of a different, but not necessarily restored, ecosystem. *See* ENDANGERED SPECIES; WETLANDS.

Challenges. Restoration ecology faces both practical and scientific challenges. Practical issues involve the costs associated with restoration projects, as well as the scale. Typically restoration is expensive, and land acquisition, nursery stock, soil preparation, grading, irrigation, monitoring, maintenance, and labor can contribute to a total construction cost that is prohibitive. This is especially relevant in developing countries, where national budgets for conservation-related activities are often minimal. In terms of scale, most restoration projects involve the restoration of less than 10 acres. A 50-acre project is considered quite large, and there are fewer than 50 restoration projects in the United States that involve more than 100 acres. The ability of restoration ecologists to effectively restore ecosystems on a long-term basis is unknown; therefore concerns about the effectiveness of restoration efforts are justified. Restoration ecologists are turning to creative ways to cut costs and expand their efforts in a land-based metric. Utilization of the large volunteer base of nonprofit conservation organizations such as The Nature Conservancy is one way that labor costs have been diminished.

Scientifically, restoration ecologists face equally challenging issues. Perhaps the most important guiding ecological principle applied to restoration

ecology is that of plant succession. Many restoration projects must accelerate the species replacement process or simply "jump-start" communities without allowing time for natural succession to occur. Ecologists are now turning to experimental approaches to determine which species should be planted when, and how best to assemble a community in a restored environment. *See* ECOLOGICAL SUCCESSION.

Peggy L. Fiedler

Bibliography. A. D. Bradshaw, Ecological principles and land reclamation practice, *Landsc. Plan.*, 11:35–48, 1984; J. Cairns, Jr., *Rehabilitating Damaged Ecosystems*, 2d ed., CRC Press, Boca Raton, 1995; R. J. Hobbs and D. A. Norton, Towards a conceptual framework for restoration ecology, *Restor. Ecol.*, 4:93–110, 1996; W. R. Jordan III, M. E. Gilpin, and J. D. Aber (eds.), *Restoration Ecology: A Synthetic Approach to Ecological Restoration*, Cambridge University Press, 1987.

Restriction enzyme

An enzyme, specifically an endodeoxyribonuclease, that recognizes a short specific sequence within a deoxyribonucleic acid (DNA) molecule and then catalyzes double-strand cleavage of that molecule. Restriction enzymes have been found in bacteria, where they serve to protect against the deleterious effects of foreign DNA, and in viruses that infect *Chlorella*-like organisms, the only eukaryotes from which those enzymes have yet been isolated. *See* DEOXYRIBONUCLEIC ACID (DNA).

Types. There are three known types of restriction enzymes.

Type I enzymes recognize a specific sequence on DNA but cleave the DNA chain at random locations with respect to this sequence. They have an absolute requirement for the cofactors adenosine triphosphate (ATP) and *S*-adenosylmethionine, and during cleavage they hydrolyze massive amounts of ATP. Because of the random nature of the cleavage, the products are a heterogeneous array of DNA fragments. *See* ADENOSINE TRIPHOSPHATE (ATP).

Type II enzymes also recognize a specific nucleotide sequence but differ from the type I enzymes in that they do not require cofactors and they cleave specifically within or close to the recognition sequence, thus generating a specific set of fragments. It is this exquisite specificity which has made these enzymes of great importance in DNA research, especially in the production of recombinant DNAs. More than 1200 such enzymes have been characterized, and almost 150 different specific sequences are recognized. In those cases where enzymes from different sources recognize the same sequence, the enzymes are called isoschizomers. The sequences that are recognized range from four to eight nucleotides in length and often contain one or two positions at which several sequence combinations can be present. Many type II enzymes recognize sequences that contain a dyad axis of symmetry (commonly called palindromes) and cleave to leave fragments with cohesive ends.

Type III enzymes have properties intermediate between those of the type I and type II enzymes. They recognize a specific sequence and cleave specifically a short distance away from the recognition sequence. However, it is often difficult to obtain complete digestion. They have an absolute requirement for the cofactor ATP, but they do not hydrolyze it. Their activity is stimulated by *S*-adenosylmethionine, although it is not an absolute requirement.

Biological importance. Given the capacity of a restriction enzyme to cleave DNA, it is interesting that the bacterium containing such an enzyme does not kill itself as a result of the restriction enzyme cutting the bacterial DNA. The reason is found by considering the circumstances under which the restriction enzyme is present in the bacterium, and relates to the biological phenomenon of restriction modification.

A restriction modification system contains two components, the restriction enzyme and the modification enzyme. The modification enzyme recognizes exactly the same sequence as is recognized by the restriction enzyme and then specifically modifies that sequence so that it can no longer be cleaved by the restriction enzyme. In all known cases, this modification is achieved by methylation. In the bacterium containing the restriction enzyme, all of the recognition sequences on the bacterial DNA are modified by the companion modification enzyme, and hence the bacterium is immune to the effects of its own restriction enzyme. However, if DNA from some foreign source enters the bacterium, as might occur during bacteriophage infection, then it will not usually be premodified in the specific manner required to protect its recognition sequences from the restriction enzyme, and so it will be cleaved and effectively destroyed. In this way the restriction modification system serves to protect the bacterium from the deleterious effects of incoming foreign DNA.

Use in research. Restriction enzymes are important to research in molecular biology. Their discovery was honored by the award of the 1978 Nobel Prize in Physiology or Medicine to Werner Arber, Hamilton O. Smith, and Daniel Nathans. Arber was the first to postulate the existence of these enzymes, and subsequently isolated the first I enzyme. Smith was responsible for the discovery and characterization of the first II enzyme, *Hin*dII from *Haemophilus influenzae*. Nathans was the first to use that enzyme to cleave SV40 DNA into discrete fragments and then to use the fragments generated to study the molecular biology of the monkey virus SV40.

Naturally occurring DNA molecules are usually rather large and therefore inaccessible to direct biochemical experimentation. For instance, one of the smallest known DNA molecules is the chromosome of SV40, which contains 5243 nucleotide pairs; a typical human chromosome may contain 10^8 nucleotide pairs. The II restriction enzymes allow molecular biologists to cut these large DNA molecules into a series of smaller specific fragments, each of which

contains a unique section of the larger molecule. For example, the enzyme *Hae*III, which recognizes the sequence GGCC, cuts the SV40 chromosome into 19 specific fragments. These fragments can be separated by size by using gel electrophoresis and then can be purified individually. The relative order of these fragments on the genome can be determined, so that a map can be produced that shows the location of each restriction enzyme site. The construction of such a physical map for one or more enzymes is often the first step in a more detailed study of a DNA molecule. Restriction enzyme cleavage sites provide a convenient set of markers along the chromosome by which genes and their regulatory signals may be ordered and their positions assigned. The aim of this kind of approach is a direct correlation of physical and genetic maps, ultimately taken to the nucleotide sequence level. In more complex cases where very large numbers of fragments are present, they can still be separated effectively by gel electrophoresis, and individual fragments can be detected by hybridization using radioactivity labeled probes. *See* GENETIC MAPPING.

Genetic engineering. A key feature of the fragments produced by restriction enzymes is that when mixed in the presence of the enzyme DNA ligase, the fragments can be rejoined. For instance, when the bacterial plasmid pBR322 is cleaved with the restriction enzyme *Eco*RI, for which there is a single site, the original circular DNA molecule is converted to a linear form. Upon addition of DNA ligase, the ends of the linear molecule are joined to reform a circular pBR322 molecule that is indistinguishable, chemically and biologically, from the starting material. This rejoining is very efficient when the restriction enzyme used produces fragments with cohesive ends. If two different linear fragments produced by cleavage with *Eco*RI are mixed in the presence of DNA ligase, then in addition to producing circular versions of the two original DNA molecules there will result recombinant molecules in which the two linear molecules have been joined together. If one of the molecules is the plasmid pBR322, which contains the genetic information necessary for replication, then the new recombinant molecule, containing the additional fragment, will be able to replicate when placed in a bacterial cell. Thus, by simple laboratory manipulation, a new molecule has been created that comprises the genetic information of the plasmid pBR322 together with new genetic information present on the additional fragment. If, instead of a single *Eco*RI fragment, a mixture of 10 fragments was present plus the linear pBR322, then 10 different recombinant plasmids would result from the experiment, each deriving its replicative ability from the pBR322 sequences. By using standard techniques of bacterial genetics, colonies of bacterial cells containing individual recombinant plasmids can be easily isolated. Indeed, extremely complex mixtures of restriction fragments can be resolved into individual components, each of which is in a readily accessible form by this same technique.

Should the new fragment carry genetic informa-tion that can be interpreted by the bacterial cell containing the recombinant molecule, then the information will be expressed as a protein, and the bacterial cell will serve as an ideal source from which to obtain that protein. For instance, if the DNA fragment carries the genetic information encoding the hormone insulin, the bacterial cell carrying that fragment will produce insulin. Because many plasmids are present in high copy number in bacterial cells, the amount of the protein produced can be extremely high. By using this method, the human gene for insulin has been cloned into bacterial cells and used for the commercial production of human insulin. The potential impact of this technology forms the basis of the genetic engineering industry. *See* ENZYME; GENETIC ENGINEERING.   Richard Roberts

Bibliography. F. C. Neidhardt et al. (eds.), *Escherichia coli and Salmonella typhimurium*: *Cellular and Molecular Biology*, 1987; R. J. Roberts, S. M. Linn, and R. S. Lloyd (eds.), *Nucleases, Monograph 25*, 2d ed., 1994; G. G. Wilson, Type II restriction-modification systems, *Trends Genet.*, 4:314–318, 1988.

Resultant of forces

A system of at most a single force and a single couple whose external effects on a rigid body are identical with the effects of the several actual forces that act on the body. For analytic purposes, forces are grouped and replaced by their resultant. Forces can be added graphically (**Fig. 1**) or analytically. The sum of more than two vector forces can be found by extending the method of Fig. 1c to a three-dimensional vector polygon in which one force is drawn from the tip of the previous one until all are laid out. The resultant force is the force vector required to close the polygon directed from the tail of the first force vector to the tip of the last. A force system has a zero force resultant if its vector polygon closes. *See* CALCULUS OF VECTORS.

The resultant of force along an axis may be desired. In that case, all forces are resolved into their
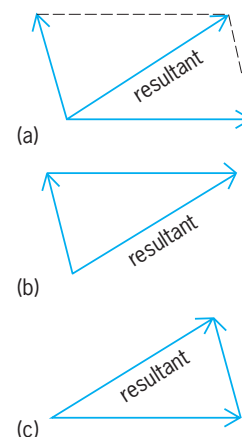

Fig. 1. Resultant of two forces acting through common center. (*a*) Diagonal of parallelogram. (*b*, *c*) Hypotenuse of triangle.
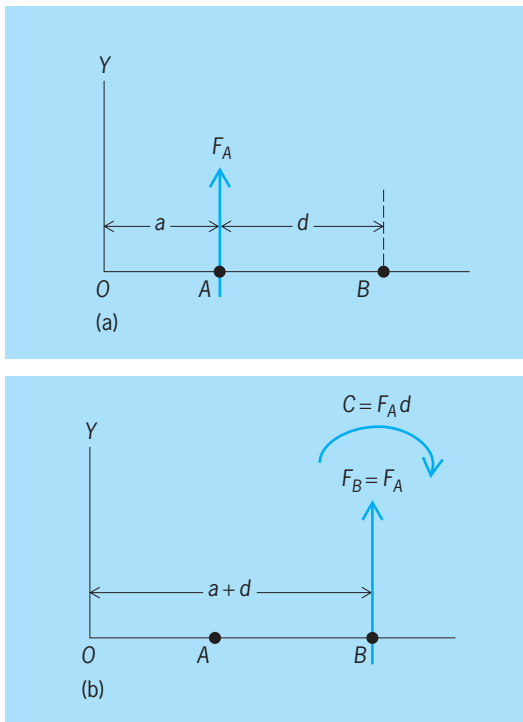
Fig. 2. Resolution of a force. (a) Force F_A. (b) Resolved force F_B and couple C.

respective components along the axis, and the components are added algebraically to obtain the desired resultant.

The resolution of a force into components is an inverse operation to the composition of multiple forces into a resultant. However, whereas the resultant force is unique to a given force system, a force can be resolved into any variety of components.

Two force systems are equivalent if their resultant forces, as described above, are equal and if their total vector moments about the same point are also equal. Vector moments are combined in the same manner as forces, that is, by parallelograms, triangles, or polygons.

In **Fig. 2** the force F_A acting at point A may be resolved into an equivalent system consisting of equal parallel F_B acting at point B and couple C; the magnitude of C is the moment of F_A about B. Also, the total moment of F_B and C about A is the moment of F_A about A, namely zero.

A resultant is the equivalent force system having the fewest possible forces and couples. The resultant of concurrent forces is a force equal to their vector sum and acting through the point of concurrence. A special case is a collinear system in which the resultant is collinear with the forces of the system as well as equal to their vector sum.

When all forces of a system are coplanar, the resultant may be a force or a couple. If it is a force, the resultant is positioned to produce the same moment about a reference point as the system. Should the vector sum of forces be zero, the resultant is a couple that develops the same moment as the system. *See* COUPLE.

In a three-dimensional force system, the resultant consists of a force element passing through an arbitrary point and equal to the total force of the system and a couple element that produces a moment equal to the total moment of the system about any point on the line of action of the force element. *See* FORCE; STATICS.   Nelson S. Fisk

Retaining wall

A generic structure that is employed to restrain a vertical-faced or near-vertical-faced mass of earth. The earth behind the wall may be either the natural embankment or the backfill material placed adjacent to the retaining wall. Retaining walls must resist the lateral pressure of the earth, which tends to cause the structure to slide or overturn.

Types. There are several types of retaining walls. A gravity wall is typically made of concrete and relies on its weight for stability (**illus.** *a*). The mass of the structure must be sufficient to develop enough frictional resistance to sliding, and the base or footing of the structure must be wide enough to develop sufficient moment to resist overturning earth forces. The base width of the gravity wall must also be sufficient that the stresses on the soil below the base are not greater than the bearing capacity of this soil. Resistance to sliding may be increased by including a key that projects down into the soil.

A cantilever retaining wall (illus. *b*) gains a larger effective mass by virtue of the soil placed on the horizontal cantilevered section of the wall. Reinforced counterforts are spaced along the wall to increase its strength. A key may be included to resist sliding. *See* CANTILEVER.

A variation of the gravity retaining wall is the crib wall (illus. *c*) is usually constructed of prefabricated interlocking concrete units. The crib is then filled with soil before the backfill adjacent to the crib is placed. Crib retaining walls require less concrete than gravity structures, and the individual units can be prefabricated for quicker construction.

Bulkhead retaining walls (illus. *d*) consist of vertical sheet piling that extends down into the soil and is stabilized by one or more tiebacks and anchors periodically spaced along the structure. The sheet piling may be made of reinforced concrete, steel, or aluminum.

A novel method for stabilizing excavations and freshly cut slopes in place is soil nailing. As excavation proceeds, the segmented face of the retaining wall is secured by many long steel bars (nails) that are driven into the embankment or placed in predrilled holes and grouted. This type of retaining wall is popular because of the economy of time and materials used in construction.

Marine installations. Retaining walls are often used in the marine environment, where they separate the retained soil from the water. Gravity walls (known as seawalls) can be constructed where strong wave and current forces are exerted on the wall. Bulkheads
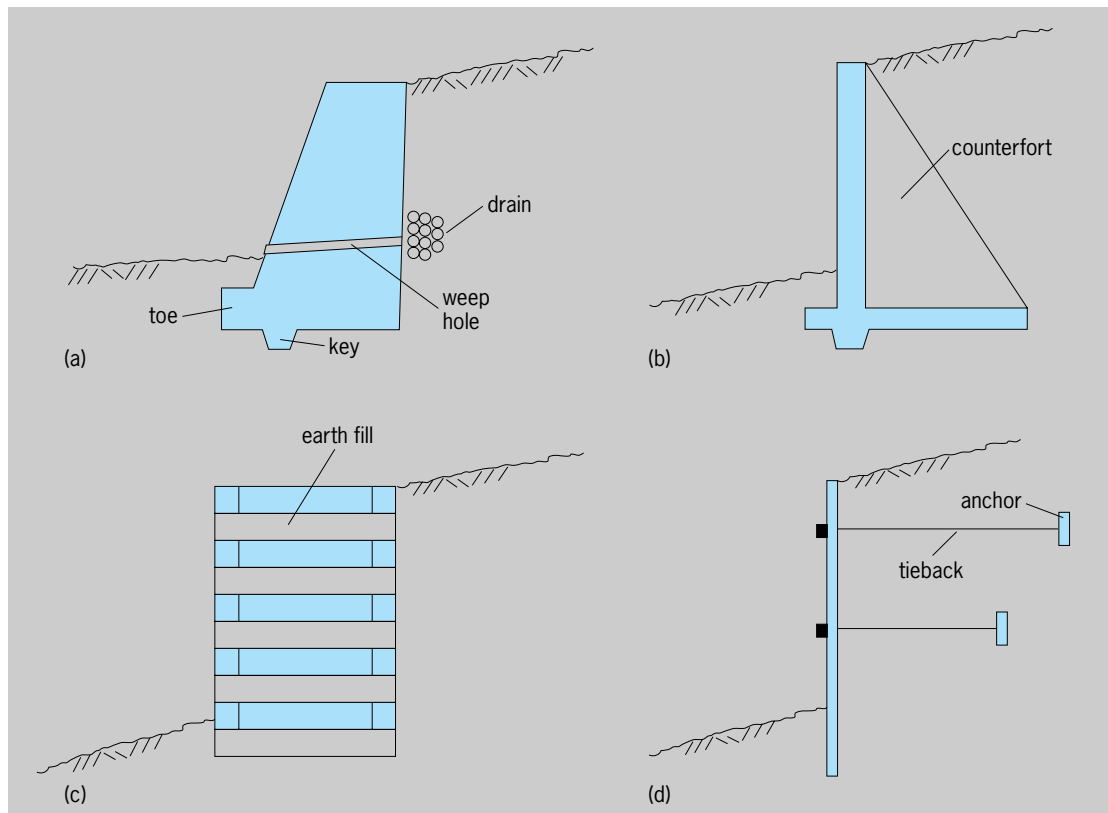
Common types of retaining walls. (*a*) Gravity wall. (*b*) Cantilever wall. (*c*) Crib wall. (*d*) Bulkhead.

are more commonly found in sheltered areas such as harbors and navigation channels. Where there is strong wave or current action, there is a potential for scour at the toe of the structure and eventual undermining of the retaining wall. To counter scour, a protective blanket of stone is placed at the structure toe. Also, the structure may be extended more deeply into the earth to allow for toe scour without causing the structure to be undermined.

Stability. The retaining wall must not fail by sliding or rotating under the combined action of all of the forces acting on the wall. These forces primarily include the vertical gravitational force acting through the center of mass of the structure, the lateral earth pressure on the wall including any dead loads (for example, buildings) or live loads (such as roadway traffic) that act on the embankment and transmit to the wall, and the frictional and bearing forces of the foundation soil.

The embankment earth pressure may be increased because of hydrostatic pressures caused by water in the soil if the embankment is not adequately drained. Water in the embankment may also cause swelling if the soil is cohesive, and ice thrust when the water freezes and thaws repeatedly. If the foundation soil is saturated, there will be a hydrostatic uplift force on the structure base.

Soil pressures at the toe of the wall will generally increase structure stability by resisting sliding of the structure. These forces should be included in the stability analysis only if the designer can be certain that this soil will remain at the toe. Lateral earth pres-

sures may be increased by earthquake loadings or other vibratory loading from machines or traffic.

Drainage. The embankment behind and under a retaining wall should be well drained. This can be accomplished by paving the top surface of the embankment and installing a gutter at the toe of the structure to carry water runoff away from the structure. A drain along the lower section of the embankment and weep holes through the structure (illus. *a*) will prevent water from accumulating in the embankment. If fill is placed in the embankment next to the structure, it should be of highly permeable material. *See* SOIL MECHANICS.   R. M. Sorensen

Bibliography. R. Brown, *Practical Foundation Engineering Handbook*, 2d ed., 2000; D. D. Driscoll, *Retaining Wall Design Guide*, Forest Service, U.S. Department of Agriculture, 1979; P. C. Lambe and L. A. Hansen, *Design and Performance of Earth Retaining Structures*, Geotech. Spec. Publ. 25, American Society of Civil Engineers, 1990; G. P. Tschebotarioff, *Foundations, Retaining and Earth Structures*, McGraw-Hill, 1978.

Reticular formation

Characteristic clusters of nerve cell bodies (gray matter) and their meshwork, or reticulum, of fibers which are found in the brainstem and the diencephalon. The reticular formation is thought to be a complex, highly integrated mechanism which exerts both inhibition and facilitation on almost

every type of activity of the central nervous system.

Anatomy. The brainstem, composed of the medulla oblongata, the pons, and the midbrain or mesencephalon, is the basic integrating and connecting unit of the central nervous system. The long ascending sensory fibers pass upward from all parts of the body and are distributed largely in the brainstem to the cerebellum, cerebral cortex, and other related higher brain centers. Similarly, the descending, or motor, fibers pass downward from higher centers through the stem to be distributed to appropriate lower levels, such as the spinal cord. In addition, the brainstem is itself the site of many important structures, notably the cell body clusters, or nuclei, of cranial nerves. *See* BRAIN; NERVOUS SYSTEM (VERTEBRATE).

The reticular formation lies in and around these better-defined structures, which break the reticular formation into many small islands of gray matter connected by large numbers of relatively short nerve fibers. These fibers pass in an apparently haphazard manner in all directions; but investigation has revealed that this reticulum is in reality a highly complex, intricately organized, master system of communication which alters many body activities.

Many authorities disagree on the exact anatomic or physiologic extent of the reticular formation because of differences in study methods in the several disciplines. Although certain reticular formation cell clusters are well circumscribed, others are not. This fact, in addition to the multiplicity of connections that are difficult to trace, has prevented a more detailed analysis.

Influence on behavior. For many years the reticular formation was largely ignored, except for a vague descriptive acknowledgment of its presence. However, increased attention has focused on this portion of the brain, mainly because of increasing evidence of its vital, though often subtle, role in many body activities. Sleep, wakefulness, attention, and other aspects of consciousness, as well as effects on muscular coordination, vascular tone, blood pressure, and many other aspects of everyday adjustments to environment are all affected by the condition and responses of the reticular system. *See* SLEEP AND DREAMING.

The discovery of drugs, such as chlorpromazine, which seem to affect selectively the reticular structures created an upsurge of interest and investigation, in which many pharmacological methods were added to those of anatomy and physiology to gain further insight in this field.

The use of electrical stimulation and electroencephalography were added to the microscopic study of neurons, fibers, and degeneration phenomena following experimental injury to selected sites. *See* ELECTROENCEPHALOGRAPHY.

These refinements in techniques led to observations which support the general view that the reticular formation may exert a facilitatory or an inhibitory action on at least three major areas of nervous activity.

Sensory impulses. One area of activity so affected is the modification of ascending impulses received from almost any sensory receptor in the body. The passage of such impulses apparently stimulates the reticular system through a collateral system. The level of activation depends upon previous correlation between cortical areas and the reticular structures involved, so that a kind of presetting mechanism is achieved. An illustration may be helpful. A person goes to bed with the conscious or unconscious knowledge that waking will be signaled by the noise of an alarm clock. Despite many other stimuli, some of a high degree of intensity, the person sleeps until the ring of the alarm. Often, in fact, the alarm need not actually ring because the warning to wakefulness may be elicited by the preliminary click of the alarm control. In a similar manner, a mother may sleep through the crying of other babies but wake at the first cry of her child.

This illustrates at least two principles involved in the modification of ascending impulses: first, the inhibitory or damping effect of the reticular formation on most sensory impulses; and second, the primary importance of some level of cortical association and control over the reticular alarm system.

Similar reactions have been repeatedly demonstrated in animal experiments, wherein electroencephalographic records can be correlated with cortical patterns in many states relating to consciousness.

The various tranquilizers are thought to work through a similar mechanism to produce an inhibition of ascending stimuli and also a decrease in reverberation within the brain itself. *See* TRANQUILIZER.

Motor activity. The second area in which nervous activity is altered by the reticular formation is in its relations with the cerebellum. Under particular conditions, the reticular formation may facilitate or inhibit certain actions involving motor movement and coordination. Somewhat conflicting observations have been reported, but it is apparent that many cerebellar functions are dependent upon intact nuclei in the reticular area, notably the paramedian reticular nuclei, the lateral reticular nuclei, and others.

Much of this relationship is a complex feedback system, wherein continuing motor activity requires continuing adjustment or coordination of both sensory and motor impulses by the modifying reticular centers. Inhibition of extensor tone in decerebrate animals is an example of one aspect of this relationship. Many more subtle influences, particularly those involving localization of motor functions between specific sites in the cerebellum and reticular nuclei, have been reported. The implications of this are of prime significance in terms of the so-called primitive reflex acts and other motor activities.

Integrative impulses. The third area influenced by the reticular formation is that of impulses mediated by the spinal cord. Facilitation or inhibition of cortical or reflex movements by reticular elements has been demonstrated repeatedly. In addition, general and specific effects on vasomotor tone, muscular tonus, and the inspiratory and expiratory phases of

respiration have been traced to alterations induced by reticular activity, or lack of it.

Summary. It may be useful to visualize the reticular formation as a complex, highly integrated mechanism which exerts some degree of inhibition or facilitation on almost every type of nerve-body activity. The dualistic response cannot be overemphasized, particularly because the cortically arranged selectivity of the inhibition-facilitation mechanism is a primary feature. The mediation of sensory, motor, and integrative impulses touches upon specific somatic states of wakefulness, sleep, attention, and related conditions of whole body activity. Finally, coordination and reflex activity require reticular formation participation. Thus, the reticular formation emerges from obscurity to become a fascinating regulatory mechanism, at present only dimly perceived. *See* CONDITIONED REFLEX; PSYCHOLOGY.   Douglas B. Webster

Reticulosa

An order of the subclass Hexasterophora in the class Hexactinellida. This is a group of Paleozoic hexactinellids with a branching form. Each branch is provided with dermal, parenchymal, and gastral spicule reticulations. *Titusvillia* from the Mississippian Period is an example. *See* HEXACTINELLIDA; HEXASTEROPHORA.   Willard D. Hartman

Retinoid receptor

A protein in the cell nucleus that mediates the actions of retinoids (the natural and synthetic analogs of vitamin A) by regulating the rates of transcription of retinoid-responsive genes. Retinoid receptors facilitate the normal biological actions of vitamin A within the body. Thus, they play an important role in the maintenance of normal growth and development; in the immune response; in male and female reproduction; in blood cell development; and in the maintenance of healthy skin and bones; and ultimately in the general good health of the organism. *See* VITAMIN A.

The biochemical details of how retinoid receptors regulate gene transcription have been worked out over the last decade and are well understood. The retinoid receptors recognize specific deoxyribonucleic acid (DNA) sequences, called retinoid response elements, within the regulatory (promoter) regions of genes. The transcription of a retinoid-responsive gene is either increased or diminished through interactions with these response elements, with coactivator or corepressor proteins present in the nucleus, and with the proteins that compose the basal transcription machinery. In general, retinoid binding to a retinoid receptor facilitates gene transcription by making the gene more accessible to enzymes and other factors responsible for the synthesis of ribonucleic acid (RNA) from the gene, whereas the absence of such binding lessens RNA synthesis.

As many as 500 genes may be regulated by retinoids and retinoid receptors. While some of these genes are likely direct targets of retinoid-dependent transcription, many others are undoubtedly indirect targets of retinoid receptor action, reflecting retinoid involvement in modulating communications within the nucleus that are needed to coordinate gene expression. *See* CELL NUCLEUS; DEOXYRIBONUCLEIC ACID (DNA); GENE; PROTEIN.

Receptor structure and distribution. The retinoid receptors are members of the steroid/thyroid/retinoid superfamily of ligand-dependent transcription factors. Each of the proteins that compose this superfamily shares certain common structural features, including a ligand-binding domain and a DNA-binding domain (**Fig. 1**). Often the members of this family of proteins are referred to as hormone nuclear receptors or, more simply, nuclear receptors. Each member of the steroid/thyroid/retinoid superfamily recognizes a specific natural ligand (or ligands), and in response to the availability of this ligand acts to influence either positively or negatively the expression of genes that are responsive to the ligand. *See* NUCLEAR HORMONE RECEPTOR.

Six different retinoid nuclear receptors have been identified, and each is the product of its own individual gene. Based on similarities in their protein structure, three of these are classified as retinoic acid receptors (RARs) and three are classified as retinoid
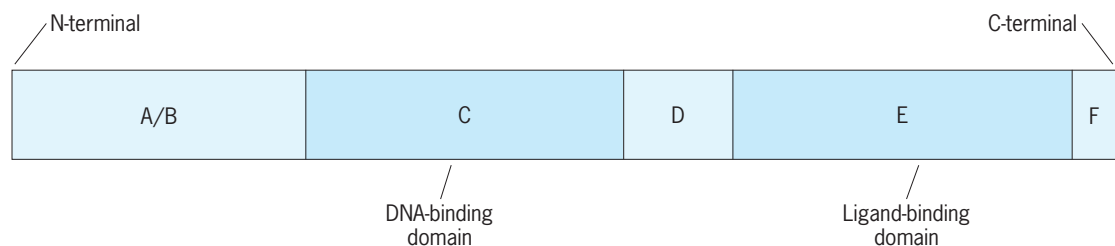

Fig. 1. Schematic of the structural and functional organization of retinoid receptors. The C region of the receptor interacts with response element DNA [DNA-binding domain (DBD)], and the E region with the retinoid ligand [ligand-binding domain (LBD)]; both are highly conserved evolutionarily. All members of the steroid/thyroid/retinoid superfamily of nuclear receptors have the same structure-function organization as illustrated here. This conserved structure-function relationship is central for identifying and classifying nuclear receptors to this superfamily. The A/B, D, and F regions are divergent and not conserved evolutionarily across the members of the steroid/thyroid/retinoid superfamily. However, these regions do play important roles in mediating the functions of the retinoid receptors and other superfamily members. Modifications to these nonconserved regions can have a significant impact on the transcription-modulating activity of the retinoid receptor.

X receptors (RXRs). The three distinct RAR subtypes are termed RAR-α, RAR-β, and RAR-γ; and the RXR subtypes are termed RXR-α, RXR-β, and RXR-γ. Furthermore, for each subtype there are at least two isoforms that are generated by differential promoter usage and alternative splicing. All-*trans*- and 9-*cis*-retinoic acid stereoisomers of retinoic acid are naturally occurring retinoid forms that have physiologically important roles in activating the retinoid receptors. All-*trans*-retinoic acid activates members of the RAR family of nuclear receptors and 9-*cis*-retinoic acid activates members of both the RAR and RXR family. Thus, different retinoid species can activate differentially the RAR and RXR retinoid receptor families.

Different combinations of RAR and RXR species are found in tissues and cells. Thus, not all tissues and cells have the same complement of RARs or RXRs. It is not fully understood what significance these differences in RAR and RXR distribution have for mediating retinoid actions in the body. It is clear, however, that almost all tissues and cells within the human body have one or more of the RARs or RXRs. This implies that retinoids and retinoid receptors play an important and diverse role in maintaining the health of all tissues within the body.

Mechanism of transcriptional regulation. Upon binding to a cognate response element, the retinoid receptors must interact with other nuclear proteins that function as either coactivators or corepressors of transcription. These interactions are essential for regulating transcriptional rates of retinoid responsive genes because they influence the degree of acetylation or deacetylation of histone proteins. The binding of retinoids to their receptors enables the receptors to interact with coactivators that have histone acetyltransferase activity (**Fig. 2**). This results in the local acetylation of histone proteins and their dissociation from the gene and an opening up of chromatin structure. This allows the transcriptional machinery access to the DNA and results in enhanced rates of gene expression. In the absence of retinoid, unliganded retinoid receptors can interact with corepressors that have histone deacetylase activity (**Fig. 3**). This results in the local deacetylation of histones and a closure of chromatin structure, thus lessening the activity of the transcriptional machinery for that gene. Consequently, corepressor binding by retinoid receptors serves to diminish transcription rates.

The response elements within genes to which the retinoid receptors bind are well characterized for many retinoid-responsive genes. Most, but not all, of these retinoid response elements share certain common characteristics, including a direct repeat (DR) of a six-nucleotide sequence, in which the two repeated sequences are separated by either two or five nucleotides (base pairs). These response elements, which consist of two 6-nucleotide half-sites separated by two or five nucleotides, are referred to as DR-2 or DR-5 motifs. The strongest, or most retinoid responsive, of these elements has a consensus six nucleotide half-site sequence of AGGTCA
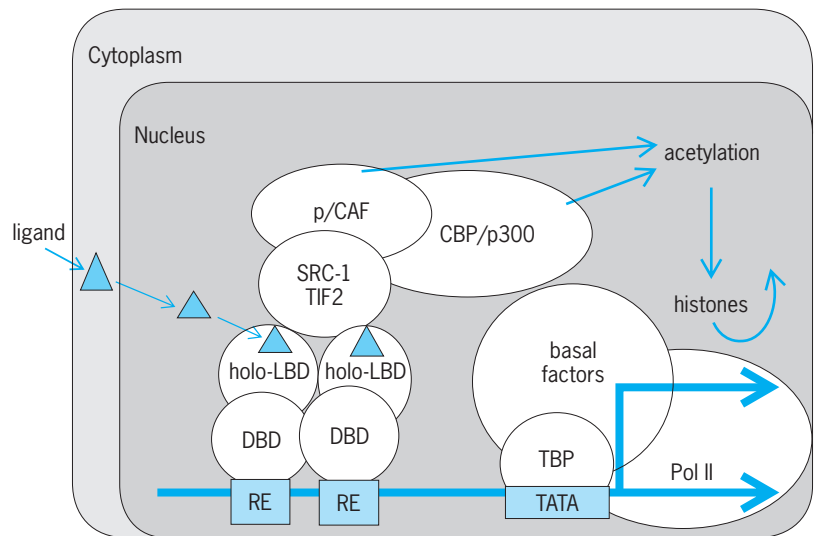

Fig. 2. Model for retinoid receptor-mediated gene activation. The DBD is able to bind the retinoid response element (RE), and an activating retinoid is available to bind the LBD, yielding a holo-LBD form. This displaces corepressor that may have bound the receptor in the absence of ligand, and facilitates binding of coactivators (p/CAF, SRC-1, TIF2, and CBP/p300 in this example). These bound coactivators catalyze the local acetylation of histones associated with the DNA composing the gene. This has the effect of releasing the histones from the DNA and making the gene more accessible to the basal transcriptional machinery (the basal factors, TBP and Pol II). This has the overall effect of enhancing the actions of this gene product within the cell.

(A = adenylate, C = cytidylate, G = guanylate, and T = thymidylate). Although minor variations in this half-site nucleotide sequence are found in different genes across different species, the presence of two repeated stretches of a sequence consisting of AGGTCA separated by two or five nucleotides in the promoter region of a gene suggests that the gene
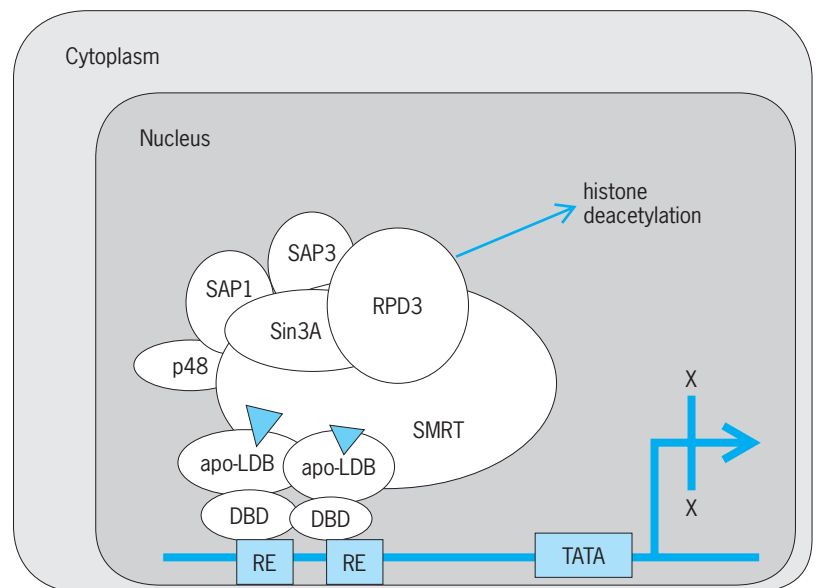

Fig. 3. Model for retinoid receptor-mediated gene silencing. The DBD is able to bind the RE. In the absence of retinoid, the LBD is unoccupied or in the apo-LBD form. This allows corepressors (p48, Sin3A, SAP1, SAP3, RPD3, and SMRT) to bind to the retinoid receptor and to catalyze local deacetylation of histones. As a result, the histones bind the DNA more tightly, thus preventing the basal transcription machinery from efficiently transcribing the gene. This has the overall effect of lessening the actions of this gene product within the cell.

is responsive to retinoid receptor action. *See* GENE ACTION; NUCLEIC ACID; NUCLEOPROTEIN.

Physiologic actions. The actions that retinoids and their receptors have in regulating gene activity affect many essential physiologic processes within the body. Living organisms require retinoids to maintain normal embryonic growth and differentiation, epithelial linings throughout the body, reproduction, an immune response, and normal brain function, in addition to assuring normal blood cell development and proper bone growth and remodeling. At the cellular level, retinoids are needed to maintain a balance between cellular differentiation and proliferation and cell death. A review of even a partial listing of the more than 500 genes reported to be responsive to retinoids provides insight into the diversity of responses regulated by retinoid receptors and their importance in maintaining the health of living organisms (see **table**).

This importance is further underscored by targeted gene disruptions (gene knockouts) in mice, in which one or more of the genes encoding RARs or RXRs have been deleted. Single RAR or RXR gene knockouts have resulted in mice that show severe developmental or functional impairments in organ systems such as the cardiovascular system, the central nervous system, the reproductive system, and the gastrointestinal system. Disruption of combinations of two RAR and/or RXR genes results in embryonic lethality. Presumably, the absence of marked embryonic lethality for single-gene knockouts arises from a functional redundancy in the actions of the RARs and/or RXRs.

Clinical applications. Because retinoids are such potent regulators of cellular processes, there has been considerable interest in the use, or potential use, of retinoids to treat chronic diseases. This interest centers especially on disease states involving aberrant cell proliferation and differentiation. Over 10,000 synthetic retinoids have been produced by medicinal chemists in the hope of developing effective pharmacological agents for use in the treatment of disease. Some of these synthetic retinoids can act as retinoid receptor agonists and bring about the same biologic actions as all-*trans*- or 9-*cis*-retinoic acid. Others act as retinoid receptor antagonists and block the actions of all-*trans*- or 9-*cis*-retinoic acid.

Skin disease. At present, both naturally occurring and synthetic retinoids are used clinically to treat skin disease. Among the retinoids commonly used in clinical dermatology are 13-*cis*-retinoic acid (also known as isotretinoin or Accutane®) and all-*trans*-retinoic acid (also known as tretinoin or RetinA®). Natural and synthetic retinoids have been used effectively in the treatment of skin disorders such as severe cystic acne, psoriasis, many cutaneous disorders of keratinization, several dermatoses, acne vulgaris, and the damaging effects of sun exposure. Although retinoids are highly effective for use in clinical dermatology, their systemic use is associated with a high incidence of birth defects (teratogenic effects). They must never be taken by pregnant women or women considering pregnancy.

Cancer. Since retinoids have an important role in regulating cellular proliferation and differentiation, there has long been interest in their use as cancer chemopreventive and chemotherapeutic agents. The most successful use of retinoids as cancer therapeutic agents has been for treatment of some forms of acute promyelocytic leukemia (APL). Retinoids, administered either as all-*trans*- or 13-*cis*-retinoic acid, have been reported to have clinical activity in cancer therapy and prevention in several other settings. These include treatment of the premalignant lesions of oral leukoplakia, actinic keratosis, cervical dysplasia, and xeroderma pigmentosum. *See* CANCER (MEDICINE); LEUKEMIA.

Cardiovascular disease and diabetes. Much recent work in animal models suggests that the pharmacological use of retinoids to target retinoid receptor action will be of benefit to patients with cardiovascular disease and with type II diabetes. Data from these studies indicate that rexinoids have efficacy for lowering plasma lipid levels, for preventing restenosis injury, and for increasing insulin responsiveness. (A rexinoid is an RXR-specific agonist or antagonist.) Clinical trials to demonstrate rexinoid efficacy in human patients are being initiated. *See* DIABETES; HEART DISORDERS.

Toxicity. Because retinoids are such potent modulators of cellular actions and because they influence many diverse physiologic actions, retinoids can have many undesirable toxic effects when administered in pharmacologically effective doses. The toxicity of retinoids is the major factor that impedes their widespread use for treating disease. Nevertheless, there is a large body of clinical research focused on using retinoids pharmacologically to treat or to prevent chronic disease.

Solveig Halldorsdottir; William S. Blaner

Bibliography. P. Chambon, A decade of molecular biology of retinoic acid receptors, *FASEB J.*, 10:940–954, 1996; A. Chawla et al., Nuclear receptors and lipid physiology: Opening the X-files, *Science,* 294:1866–1870, 2001; IARC Working

Examples of genes regulated by retinoid receptors	
Substance encoded in gene	Action
Oxytocin	Needed for reproduction
Growth hormone	Needed to maintain growth
Phosphoenol pyruvate carboxykinase	Involved in glucose metabolism
Class I alcohol dehydrogenase	Needed for alcohol oxidation
Tissue transglutaminase	Needed to regulate cell growth and death
Laminin B1	Needed to facilitate cell-cell interactions
Matrix gla-protein	Needed for bone formation
Several keratin genes	Needed for healthy skin formation
Cellular retinol-binding protein, type I	Active in retinoid metabolism
RAR-β	Affects retinoid action
Cytochrome 26	Catalyzes catabolism of all-*trans*-retinoic acid
Several hox genes	Needed to control embryologic development
Dopamine D2 receptor	Needed in central nervous system function

Group on the Evaluation of Cancer Prevention Agents, *IARC Handbooks of Cancer Prevention*, vol 3: *Vitamin A*, World Health Organization, International Agency for Research on Cancer, Lyon, 1998; IARC Working Group on the Evaluation of Cancer Prevention Agents, *IARC Handbooks of Cancer Prevention*, vol 4: *Retinoic Acid*, World Health Organization, International Agency for Research on Cancer, Lyon, 2000; H. Nau and W. S. Blaner (eds.), *Retinoids: The Biochemical and Molecular Basis of Vitamin A and Retinoid Action*, vol. 139 of *Handbook of Experimental Phyarmacology*, Springer Verlag, Heidelberg, 1999; L. Packer et al. (eds.), *Carotenoids and Retinoids: Molecular Aspects and Health Issues*, AOCS Press, Champaign, 2005; M. B. Sporn, A. B. Roberts, and D. S. Goodman (eds.), *The Retinoids: Biology, Chemistry, and Medicine*, 2d ed., Raven Press, New York, 1994; L.-N. Wei, Retinoid receptors and their coregulators, *Annu. Rev. Pharmacol. Toxicol.*, 43:47–72, 2003.

Retortamonadida

An order of parasitic flagellate protozoa belonging to the class Zoomastigophorea. P. P. Grassé suggested that there should be eight orders of these parasitic (and symbiotic) zooflagellates, all in a superorder Metamonadina, but this may be an oversimplification. All retortamonads are medium to large in size and have a complicated blepharoplast-centrosome-axo-style apparatus. Retortamonadida have two or four flagella, one turned ventrally into a cytostomal depression. The nucleus, containing a distinct endosome, is located at the anterior tip. The body is twisted. These organisms are actually symbionts,
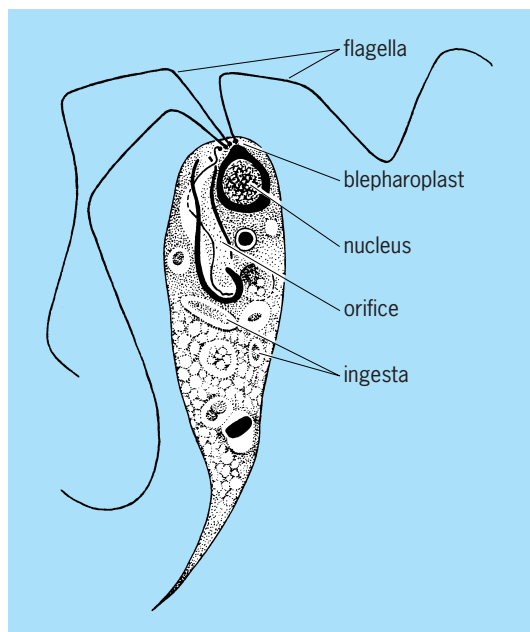
A retortamonad, *Chilomastix aulastomi*.

ingesting bacteria in the digestive tracts of their hosts.

Retortamonas has several species, some of which infest insects and vertebrates. *Chilomastix* also has a number of species, found in vertebrates and invertebrates (see **illus.**). *Chilomastix mesnili* is probably the best-known species; it is found in humans, but is believed to be harmless. It shows the sinistral spiraling very well, and the cytostome with the fourth flagellum lying therein. The cell divides by a well-defined mitosis; the nucleus migrates to midcell, where it forms a nuclear spindle. The poles are blepharoplasts, which give rise to flagella. *See* PROTOZOA; SARCOMASTIGOPHORA; ZOOMASTIGOPHOREA.   James B. Lackey

Bibliography. J. A. Pechenick, *Biology of the Invertebrates*, 5th ed., McGraw-Hill, 2005.

Retrograde motion (astronomy)

In astronomy, either an apparent east-to-west motion of a planet or comet with respect to the background stars or a real east-to-west orbital motion of a comet about the Sun or of a satellite about its primary. The majority of the objects in the solar system revolve from west to east about their primaries. However, near the time of closest approach of Earth and a superior planet, such as Jupiter, because of their relative motion, the superior planet appears to move from east to west with respect to the background stars. The same apparent motion occurs for an inferior planet, such as Venus, near the time of closest approach to Earth.

Actual, rather than apparent, retrograde motion occurs among the satellites and comets; the eighth and ninth satellites of Jupiter and the ninth satellite of Saturn are examples. *See* ORBITAL MOTION.
                                    Raynor L. Duncombe

Retrovirus

A family of viruses distinguished by four characteristics: (1) genetic information is in the form of ribonucleic acid (RNA); (2) each single virus particle or virion (**illus.** *c*, arrow) possesses the enzyme reverse transcriptase; (3) virion morphology consists of two proteinaceous structures, a dense core (illus. *a*) and an envelope (illus. *a*) that surrounds the core; and (4) the viral reproductive cycle requires the synthesis of a deoxyribonucleic acid (DNA) copy of the viral genome and insertion of that DNA into the genome of the host cell. *See* ANIMAL VIRUS; RIBONUCLEIC ACID (RNA).

The genome is composed of two identical molecules of single-stranded RNA. This genomic RNA is similar in function to messenger RNA (positive polarity). The genome contains two structural genes, termed *gag* and *env*, that code for the structural coat of the virus. *Env* codes for the viral proteins of the outer envelope surface of the virus and *gag* codes for the core proteins. The core contains the RNA genome of the virus.
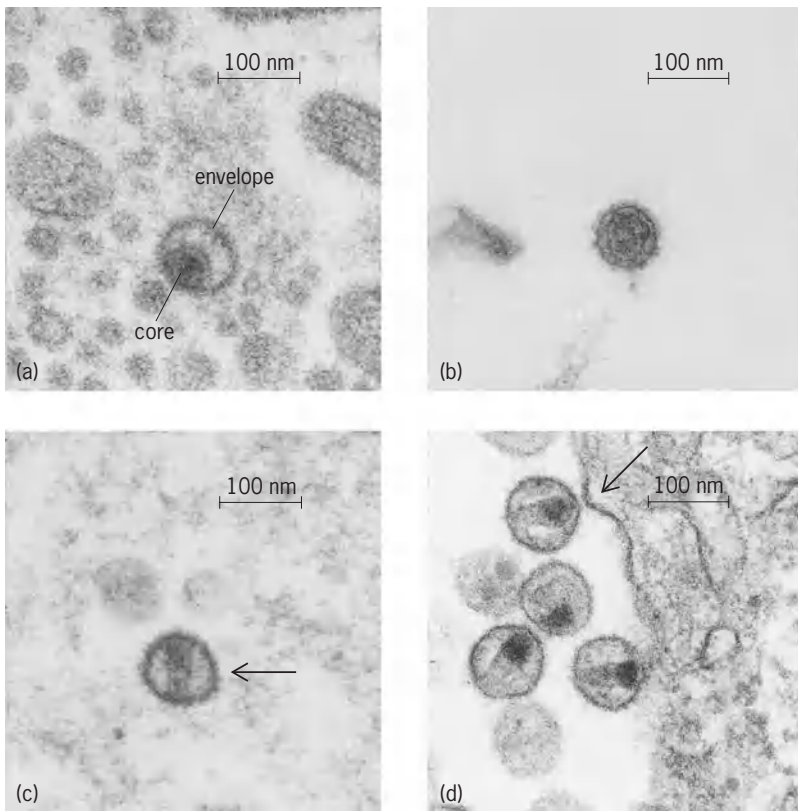
Different morphological types of retroviruses. (*a*) Murine mammary tumor virus, a betaretrovirus, shown with its off-center core surrounded by an envelope. (*b*) Feline leukemia virus, a gammaretrovirus, shown with its centrally located core. (*c*) SRV, a betaretrovirus of Asian macaque monkeys, shown with its rodlike core. (*d*) Human immunodeficiency virus (HIV), a lentivirus that causes acquired immune deficiency syndrome (AIDS), shown with its conical core.

Other genes, such as *pol*, encode enzymes such as the reverse transcriptase that is needed for viral multiplication. Some retroviruses have additional genes as well. DNA is not present in the virions of retroviruses. The reverse transcriptase in each virion copies the RNA genome into DNA shortly after entry of the virus into the animal or human host cell. The discovery of this enzyme changed thinking in biology. Previously, the only known direction for the flow of genetic information was from DNA to RNA, yet retroviruses make DNA copies of their genome by using an RNA molecule. This reversal of genetic information was "backwards," and hence the family name retrovirus, meaning backward virus. *See* REVERSE TRANSCRIPTASE.

The DNA copy of the RNA genome is inserted directly into one of the chromosomes of the host cell. This results in new genetic information being acquired by the host cell. Acquiring new genetic information can induce changes in that cell that may lead to tumors. The study of reverse transcriptase led to other discoveries of how retroviruses may add a variety of new genetic information into the host. One such class of new genes carried by retroviruses is oncogenes (tumor genes). Oncogenes are carried either in place of or in addition to one of the normal genes of retroviruses. Retroviral oncogenes appear to be involved in the induction of tumors in animals. *See* ONCOGENES.

Multiplication. Viral multiplication occurs in six steps: adsorption, penetration, uncoating, synthesis of viral proteins and nucleic acids, assembly and maturation, and release. In the first step, the external envelope of the virus attaches itself to the surface cellular membrane of an animal or human cell (illus. *d*, arrow). The cell takes up the bound virus (penetration) and the virus releases its genome (uncoating) and begins to multiply in the host cell. Viral synthesis begins when the reverse transcriptase copies the RNA genome into DNA. A viral integrase enzyme then inserts the newly made DNA into the genome of one of the host chromosomes in the nucleus of the infected cell. The viral DNA, called a provirus in this state, is transcribed by cell enzymes to form viral messenger RNAs as well as new viral RNA genomic molecules. The messenger RNAs are then translated into viral proteins such as the reverse transcriptase, gag, and envelope proteins. These newly made components assemble into new progeny virus, signaling the end of viral synthesis and the beginning of maturation. The newly made virions are released from the cell surface by budding through the outer cell membrane. The newly made virions will continue the cycle by infecting new cells in that host or by transmission to a new individual of that species.

Specific genera of retroviruses and diseases. The family Retroviridae contains all of the Retroviruses and is divided into seven genera; Alpharetrovirus, Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Epsilonretrovirus, Lentivirus, and Spumavirus. Retroviruses are either exogenous or endogenous. Exogenous retroviruses are transmitted horizontally by infection of new cells or hosts, just as other viruses are transmitted. Endogenous viruses, unlike other viruses, are transmitted vertically, as their DNA proviruses are inherited through the genome of a species.

1. Alpharetroviruses contain the avian leucosis virus and the Rous sarcoma virus. All members infect avian species and some are associated with tumors.

2. Betaretroviruses contain viruses that infect mice, certain species of monkeys, and sheep. Mouse mammary tumor virus is associated with mammary tumors in mice. Mason-Pfizer monkey virus, SRV-1 and SRV-2, are natural infections of Asian macaque monkeys. Infected monkeys develop an AIDS-like disease and also may have fibrosarcoma, a tumor similar to Kaposi sarcoma, a skin cancer seen in humans with AIDS. SRV-1, SRV-2, and Mason-Pfizer monkey virus are transmitted by a bite that pierces the skin. The betaretroviruses of langur and squirrel monkeys are endogenous retroviruses, meaning that they are not acquired by infection but are inherited. Ovine pulmonary adenosarcoma, a tumor of sheep, is associated with the Jaagsiekte sheep retrovirus.

3. Gammaretrovirus contains retroviruses of diverse species, including snakes and the domestic cat. Feline leukemia virus (FeLV) (illus. *b*) infects the domestic cat and is of significant economic importance as a cause of an AIDS-like disease and leukemia in that species. There is no evidence of infection of humans

by FeLV. Vaccines against FeLV are available and are widely used. Retroviruses that are associated with leukemia in gibbon apes and in mice are also known. Endogenous retroviruses of mice are also included in this genus.

4. Deltaretroviruses contain leukemia viruses of cattle, humans, and simians. The human T cell leukemia virus (HTLV) is one of two retroviruses known to cause disease in humans. The other is HIV, the cause of AIDS. HTLV is associated with human leukemias of T cell origin and neurologic diseases. The closest known relative of HTLV is simian T cell leukemia virus (STLV). STLV naturally infects a number of simian species in Africa and Asia. Genetic comparisons between HTLV and STLV indicate that HTLV originated from cross-species transmission of STLV to humans. Leukemia in humans develops in about 2% of infected persons.

5. Epsilonretrovirus contains exogenous retroviruses of fish. The Walleye dermal sarcoma virus, perch hyperplasic virus, and snakehead retrovirus have been described.

6. The genus Lentivirus is named for the slow progression to disease in the infected host. Only exogenous lentiviruses are known. Members include Visna-Maedi virus, associated with pulmonary and central nervous system diseases of sheep, and the human immune deficiency virus (HIV), which is the cause of AIDS in humans.

Two lentiviruses found in chimpanzees and sooty mangabey monkeys, SIVcpz and SIVsm, are the ancestral viruses of HIV type 1 and HIV type 2, respectively. Although HIV has its origins in African simian species, it is not correct to think of AIDS as a zoonosis (a disease acquired directly from animals, like rabies); there is no evidence that SIV directly causes AIDS in human beings. Over 30 SIVs have been identified in simian species on the African continent. It is noteworthy that SIV rarely causes AIDS in its natural simian hosts in Africa.

Other animal lentiviruses are known, including feline immunodeficiency virus, bovine immunodeficiency virus, and equine infectious anemia virus.

7. Spumavirus contains the foamy viruses, so called because they appear to cause foam when infecting tissue culture cells. These viruses are found in humans, chimpanzees, monkeys, and domestic cats. The virus is present in saliva. There is evidence that the simian foamy viruses gave rise to human infections. There is no known disease associated with the spumavirus group.

Epidemiology. Endogenous retroviruses are inherited as part of the genome of the host species. Exogenous retroviruses are acquired by infection. The exogenous retroviruses gain entry into the host through the respiratory system, circulatory system, genital tract, or by a bite or other action that pierces the skin. Retroviruses may be sexually transmitted, as is HIV. HIV can also be passed from mother to infant during gestation, at birth, or through infected breast milk. Treatment of the infected mother may effectively prevent transmission to her child. Infection through the respiratory tract occurs when a virus enters the mouth or nose. Entry into the circulatory system can occur from a blood transfusion, use of contaminated needles, or from insect bites, as is the case for equine infectious anemia. However, HIV is not transmitted through insect bites.

Origin of the AIDS virus. AIDS was identified in Los Angeles in 1981. The causative agent, HIV-1, was discovered by scientists at the Pasteur Institute in Paris in 1983. Thus far, all available evidence indicates that the AIDS virus is spreading worldwide in humans for the first time. There are no known past epidemics. Computer models that attempt to calculate the age of HIV based on its genetic makeup place its origin from 50 to 75 years ago. The SIV viruses that are ancestral to HIV are found in naturally infected simian hosts (the chimpanzee and sooty mangabey) that live along the west coast of Africa. Yet, despite 300 years of slave trading from that same African coast, HIV did not spread to North or South America, indicating HIV's origin after the end of the slave trade. This fact makes AIDS a new disease, relatively speaking.

Although the African origin of HIV from infected monkeys and apes is certain, it is not known what launched HIV and how it became a potent human pathogen. HIV probably adapted to humans in the twentieth century by mechanisms still unknown.   Preston A. Marx

Bibliography. C. Apetrei, D. L. Robertson, and P. A. Marx, The history of SIVs and AIDS: Epidemiology, phylogeny and biology of isolates from naturally SIV infected non-human primates (NHP) in Africa, *Frontiers Biosci.*, 9:225–254, 2004; E. Drucker, P. G. Alcabes, and P. A. Marx, The injection century: Consequences of massive unsterile injecting for the emergence of human pathogens, *Lancet*, 358:1989, 2001; D. M. Knipe et al. (eds.), *Fields Virology*, 2 vols., 4th ed., pp. 1871–2123, 2001.

Reverberation

After sound has been produced in, or enters, an enclosed space, it is reflected repeatedly by the boundaries of the enclosure, even after the source ceases to emit sound. This prolongation of sound after the original source has stopped is called reverberation. A certain amount of reverberation adds a pleasing characteristic to the acoustical qualities of a room. However, excessive reverberation can ruin the acoustical properties of an otherwise well-designed room. A typical record representing the sound-pressure level at a given point in a room plotted against time, after a sound source has been turned off, is given in the decay curve shown in **Fig. 1**. The rate of sound decay is not uniform but fluctuates about an average slope.

Reverberation time. Because of the importance of the proper control of reverberation in rooms, a standard of measure called reverberation time (abbreviated t_{60}) has been established. Reverberation time is the time required for sound to die away to one-thousandth of its initial pressure, that is, to drop 60 dB in sound-pressure level.
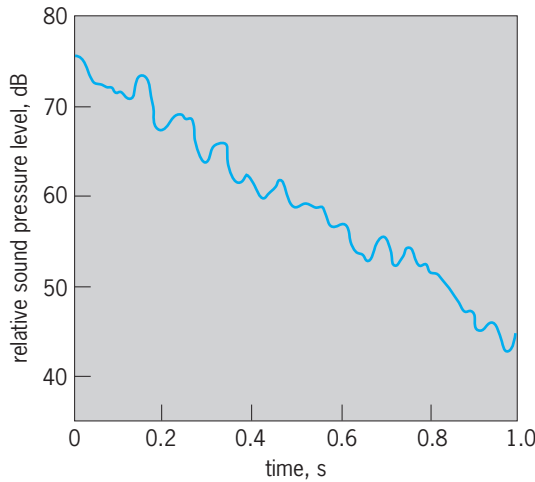
Fig. 1. Typical decay curve illustrating reverberation.

Optimum reverberation time is a matter of individual preference. A critical study of empirical data based upon preference evaluations in the United States and abroad has been made by V. O. Knudsen and C. M. Harris (**Fig. 2**). Since the optimum reverberation time for music depends on the type of music, it is represented in the figure by a broad band. The optimum reverberation time for a room used primarily for speech is considerably shorter; reverberation times longer than those shown for speech result in a decrease in speech intelligibility. The optimum reverberation time at frequencies other than 500 Hz is obtained by multiplying the 500-Hz value by the ratio R, which is given in **Fig. 3**. Note that R is unity for frequencies above 500 Hz and is given by a band for frequencies below 500 Hz. For large rooms R may have any value within the indicated band; for small rooms preferred ratios are in the lower part of the band.

Mean free path. According to the principles of geometrical acoustics, sound radiated from a source in an enclosure is successively reflected by its boundaries. The average distance between reflections is defined as the mean free path. The mean free path of a sound ray in a room depends on the shape and size of the room, and to some extent on the distribution and nature of the absorptive material. However, in most cases, it is approximately $4V/S$, where V is the volume of the room and S the total surface area.

Decay rate. The number of reflections per second of a decaying sound wave is numerically equal to the distance sound will travel in 1 s, that is, the velocity of sound c, which is about 343 m/s (1130 ft/s) in air at 20°C (68°F), divided by the average distance between reflections, or the mean free path. Hence the number of reflections per second is $cS/4V$. Each time a wave strikes one of the boundaries, on the average, a fraction (α) of the energy is absorbed, and a fraction $(1 - \overline{\alpha})$ is reflected. Here $\overline{\alpha}$ is the average absorption coefficient given by Eq. (1), where α_1 is

$$\overline{\alpha} = \frac{\alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \cdots}{S_1 + S_2 + S_3 + \cdots} \qquad (1)$$

the coefficient of absorption of surface S_1, and so forth. Because sound pressure is proportional to the square root of sound intensity, the ratio of the average reflected pressure to incident pressure is given by $(1 - \overline{\alpha})^{1/2}$, and the average decrease in the sound pressure level is therefore given by notation (2).

$$10 \, \log_{10} \left(\frac{1}{1 - \overline{\alpha}} \right) \text{dB/reflection} \qquad (2)$$

Since there are $cS/4V$ reflections per second, the average decay rate is given by Eq. (3a), where S is in square meters and V is in cubic meters, or (3b), where S' is in square feet and V' is in cubic feet.

Decay $= 373(S/V)$
$$\cdot \, [-2.30 \log_{10} (1 - \overline{\alpha})] \text{ dB/second} \quad (3a)$$

Decay $= 1230(S'/V')$
$$\cdot \, [-2.30 \log_{10} (1 - \overline{\alpha})] \text{ dB/second} \quad (3b)$$

Reverberation-time formulas. From Eqs. (3) for decay rate, it follows that the time it takes for the second-pressure level to decay 60 dB, that is, the reverberation time, is given by Eqs. (4).

$$t_{60} = \frac{0.161V}{S[-2.30 \log_{10} (1 - \overline{\alpha})]} \text{ seconds} \qquad (4a)$$
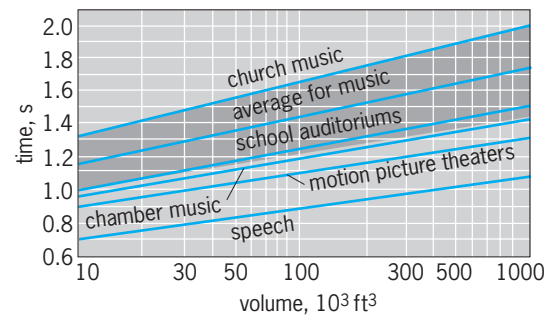

Fig. 2. Optimum reverberation time at 500 Hz for different types of rooms as a function of room volume. This figure should be used in conjunction with Fig. 3 to obtain optimum reverberation time as a function of frequency. 10^3 ft $=$ 28.3 m³. (*After V. O. Knudsen and C. M. Harris, Acoustical Designing in Architecture, John Wiley and Sons, 1950*)
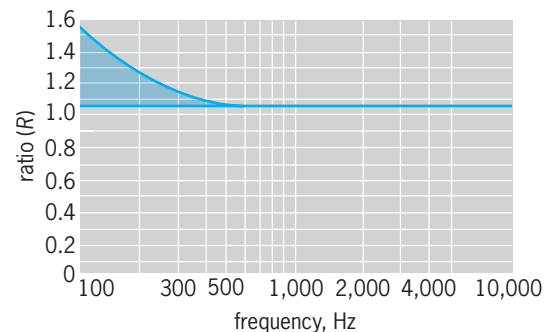

Fig. 3. Chart for computing optimum reverberation time. The time at any frequency is given in terms of the ratio R, which should be multiplied by the optimum time at 500 Hz (from Fig. 2) to obtain the optimum time at that frequency. (*After V. O. Knudsen and C. M. Harris, Acoustical Designing in Architecture, John Wiley and Sons, 1950*)
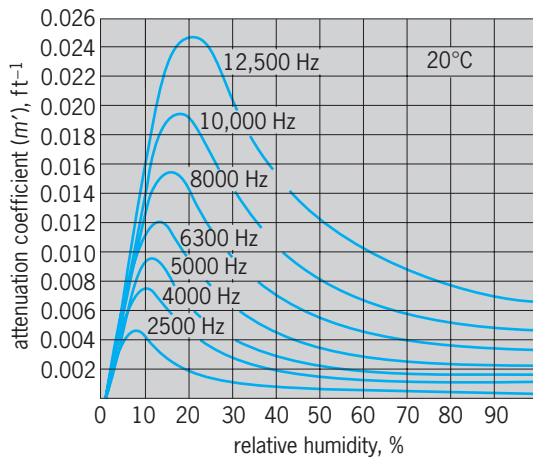
Fig. 4. Values of the attenuation coefficient *m* as a function of relative humidity for different frequencies. 1 ft^{-1} = 3.28 m^{-1}. 20°C = 68°F. (*After C. M. Harris, Absorption of sound in air versus humidity and temperature, J. Acoust. Soc. Amer., 40:148–159, 1966*)

$$t_{60} = \frac{0.049V'}{S'[-2.30 \log_{10}(1-\overline{\alpha})]} \text{ seconds} \quad (4b)$$

When $\overline{\alpha} \ll 1$, Eqs. (4) become Eqs. (5). For fre-

$$t_{60} = \frac{0.161V}{S\overline{\alpha}} \text{ seconds} \quad (5a)$$

$$t_{60} = \frac{0.049V'}{S'\overline{\alpha}} \text{ seconds} \quad (5b)$$

quencies above 2000 Hz, especially in large auditoriums, the effects of air absorption must be included in the reverberation time formulas. The corresponding equations are then Eqs. (6) and (7), where *m* is

$$t_{60} = \frac{0.161V}{S[-2.30 \log_{10}(1-\overline{\alpha}) + 4mV]} \text{ seconds} \quad (6a)$$

$$t_{60} = \frac{0.049V'}{S'[-2.30 \log_{10}(1-\overline{\alpha}) + 4m'V']} \text{ seconds} \quad (6b)$$

$$t_{60} = \frac{0.161V}{S\overline{\alpha} + 4mV} \text{ seconds} \quad (7b)$$

$$t_{60} = \frac{0.049V'}{S'\overline{\alpha} + 4m'V'} \text{ seconds} \quad (7b)$$

the attenuation coefficient in inverse meters and m' is the attenuation coefficient in inverse feet, given in **Fig. 4** . It can be shown that air absorption is molecular in origin.

The above reverberation time formulas apply only in rooms in which sound is diffuse, a condition that was assumed in their derivation. Thus, large discrepancies may be noted between the calculated and observed values of reverberation time in a room whose shape or absorptive treatment (or both) do not promote diffuse conditions of sound. *See* ARCHITECTURAL ACOUSTICS; SOUND.   Cyril M. Harris

Bibliography. T. Cox and P. D'Antonio, *Acoustic Absorbers and Diffusers*, Spon Press, London, 2004; M. D. Egan, *Architectural Acoustics*, McGraw-Hill, 1988; C. M. Harris (ed.), *Handbook of Acoustical Measurements and Noise Control*, 3d ed., McGraw-Hill, 1991; L. E. Kinsler et al., *Fundamentals of Acoustics*, 4th ed., Wiley, 2000; V. O. Knudsen and C. M. Harris, *Acoustical Design in Architecture*, Wiley, 1978; Z. MacKawa and P. Lord, *Environmental and Architectural Acoustics*, E & FN Spon, London, 1994.

Reverse transcriptase

Any of the deoxyribonucleic acid (DNA) polymerases present in particles of retroviruses, that is, in virions of ribonucleic acid (RNA) tumor viruses and related viruses. These DNA polymerases are coded by the retroviruses and are able to carry out DNA synthesis using an RNA template. This reaction is called reverse transcription since it is the opposite of the usual transcription reaction, which involves RNA synthesis using a DNA template. Thus, the viral DNA polymerases that carry out this reverse reaction have been called reverse transcriptases, and the viruses with this enzyme are called retroviruses because they reverse the usual flow of genetic information. Such viruses are much discussed now, since the etiologic agent of acquired immune deficiency syndrome (AIDS) is a retrovirus. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); RETROVIRUS.

Genetic information flow. DNA carries the genetic information of all living systems except RNA viruses. As shown in **Fig. 1**, the genetic information in DNA for proteins is first transferred to an RNA molecule, messenger RNA, by a process called transcription and then to the final product, protein, by a process called translation. The other process of information transfer in living systems except for RNA viruses is the multiplication of DNA by a process known as replication. In all living systems except some viruses, genetic information flows only from DNA to DNA and from DNA to RNA to protein. *See* GENETIC CODE; PROTEIN.

Virus classification. Viruses are simple genetic systems that consist of genes coding for viral structural components and for proteins necessary for virus replication. Viruses can replicate only in living cells.
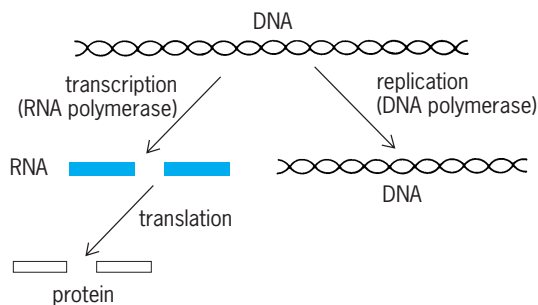

Fig. 1. Modes of genetic information transfer found in all living systems except some viruses.
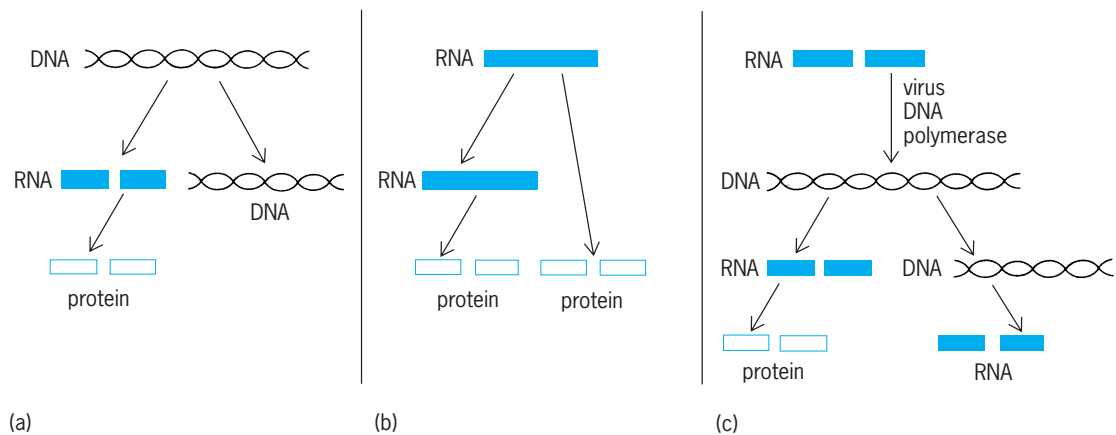
Fig. 2. Modes of information transfer that characterize the replication of viruses in three major classes. (*a*) DNA viruses or deoxyviruses (example: smallpox virus); (*b*) RNA viruses or riboviruses (example: poliovirus); and (*c*) retroviruses (example: Rous sarcoma virus and human immunodeficiency virus).

Viruses are grouped into three major classes according to the kind of nucleic acid contained in the extracellular and intracellular viral genomes (**Fig. 2**). There are viruses whose genome is only DNA (DNA viruses or deoxyviruses), viruses whose genome is only RNA (RNA viruses or riboviruses), and viruses whose genome is RNA in extracellular virus particles and DNA inside cells (retroviruses). There are also viruses whose genome is DNA in extracellular virus particles and RNA inside cells (pararetroviruses, that is, hepadnaviruses such as hepatitis B virus and caulimoviruses). Deoxyviruses use the same modes of information transfer as do cells. Riboviruses use information transfer from RNA to RNA in addition to information transfer from RNA to protein. Retroviruses use an additional mode of information transfer from RNA to DNA. *See* VIRUS.

DNA provirus. The transfer of genetic information from RNA to DNA in retrovirus replication was proposed in 1964 by H. M. Temin in the DNA provirus hypothesis for the replication of Rous sarcoma virus, an avian retrovirus which causes tumors in chickens and transformation of cells in culture. Rous sarcoma virus is the prototype retrovirus.

The DNA provirus hypothesis was based on the results of studies of the effects of inhibitors of nucleic acid synthesis on the replication of Rous sarcoma virus. The replication of Rous sarcoma virus is inhibited by actinomycin D, an inhibitor of DNA-dependent RNA synthesis. In addition, when DNA synthesis is inhibited by compounds such as cytosine arabinoside early after infection of nondividing cells by Rous sarcoma virus, no Rous sarcoma virus replication or cell transformation is found. These studies indicated that a DNA intermediate is involved in the replication of Rous sarcoma virus. The discovery in 1970 of an enzyme, reverse transcriptase, that carries out RNA-directed DNA synthesis in the particles of retroviruses provided convincing evidence in support of the DNA provirus hypothesis. (Reverse transcriptase also carries out DNA-directed DNA synthesis.) Direct evidence for the DNA provirus hypothesis was then obtained by nucleic acid hybridization experiments that detected new Rous sarcoma virus–specific DNA sequences in the DNA of Rous sarcoma virus–infected cells, and by the isolation of infectious DNA from Rous sarcoma virus–infected cells which, after addition to cells in culture (by transfection), resulted in the production of complete infectious progeny Rous sarcoma virus.

The role of reverse transcriptase during retrovirus replication was established by the study of temperature-sensitive mutants of Rous sarcoma virus with a temperature-sensitive reverse transcriptase. These virus mutants do not replicate at temperatures that inactivate the DNA polymerase activity in their virions. In addition, retrovirus particles that lack reverse transcriptase are not infectious. Therefore, the reverse transcription of information from the viral RNA genome to the viral DNA intermediate by the reverse transcriptase in retrovirus virions is an essential early step in retrovirus replication.

Now, RNA-directed DNA polymerase activity or genes coding for proteins homologous to reverse transcriptase have been reported in several cellular genetic elements. These are listed in the **table** and are discussed below.

Properties of reverse transcriptase. Reverse transcriptase has been purified from virions of many retroviruses. The avian, murine, and human retrovirus DNA polymerases have been extensively studied.

All reverse transcriptases have ribonuclease H activity on the same molecule as the DNA polymerase activity. Ribonuclease H activity is an enzymatic activity that degrades only the RNA strand of RNA-DNA hybrid molecules. It is involved in retrovirus DNA synthesis. The ribonuclease H activity of the retrovirus DNA polymerase is endonucleolytic. A DNA endonuclease activity, in the form of an enzyme that can cut DNA and is involved in retrovirus integration, is also coded by the reverse transcriptase gene.

Most of the other biochemical properties of the purified reverse transcriptases are common to them and to other DNA polymerases. For example, like cellular DNA polymerases, reverse transcriptases

Names and properties of DNA sequences encoding reverse transcriptase (retroelements) or resulting from it				
Reverse transcriptase	Transposition or integration	Long terminal repeats	Virions	Name
Present	Absent	Absent	Absent	Retron
Present	Present	Absent	Absent	Retroposon or LINE element
Present	Present	Present	Absent	Retrotransposon
Present	Present	Present	Present	Retrovirus
Present	Absent	Present	Present	Pararetrovirus (hepadnavirus; caulimovirus)
Absent	Absent	Absent	Absent	Retrosequence (cDNA genes; *Alu* sequences)

require for DNA synthesis a primer with a free $3'$-hydroxyl group, a template, a divalent cation (Mg^{2+} or Mn^{2+}), and all four deoxyribonucleoside triphosphates (**Fig. 3**). The direction of DNA synthesis for reverse transcriptases and all other DNA polymerases is from the $5'$ to the $3'$ end of the newly synthesized DNA. However, the retrovirus reverse transcriptases, in contrast to most cellular DNA polymerases, are able to utilize natural RNA templates for DNA synthesis as efficiently as DNA templates. Moreover, retrovirus reverse transcriptases do not seem to have strict requirements for specific RNA templates, unlike other RNA virus replicases or RNA-directed RNA polymerases, which usually recognize only their own genome as a template.

One consequence of this lack of strict template specificity of reverse transcriptases is that they can be used to make DNA copies of any RNA. This property of reverse transcriptase has become very useful in molecular biology, especially in genetic engineering. DNA copies of purified RNAs are made by reverse transcriptase, and then the DNA is cloned in bacteria. Several potentially important molecules have been produced in bacteria as a result of this cloning. *See* GENETIC ENGINEERING.

It has also been determined that a reverse transcriptase activity is involved in the formation of telomeres (chromosome ends). This is an essential cellular enzyme activity.

Mechanisms of synthesis of retrovirus DNA. Much has been learned about the detailed mechanism of synthesis of retroviral DNA. This mechanism depends upon the properties of the reverse transcriptase and ribonuclease H described above and properties of the viral RNA and DNA described in **Figs. 4** and **5**. The viral RNA contains a 15- to 250-base direct repeat (R) at its $5'$ end next to the cap and at its $3'$ end next to the poly (A) tail. At the $5'$ end of the viral RNA, about 150–350 bases from the end is a sequence (PBS or PB) complementary to the eighteen $3'$ nucleotides of a specific transfer RNA molecule which acts as a binding site for that tRNA. The tRNA serves as a primer for minus-strand DNA synthesis. The cellular tRNA is tRNA^{trp} for avian retroviruses and tRNA^{pro} for some mammalian retroviruses. In the viral DNA there are additional sequences outside the sequences that are directly repeated in the viral RNA such that the viral DNA has a direct repeat of 300–1200 base pairs, depending upon the species of virus. This direct repeat is called the long terminal repeat (LTR)

and consists of sequences called U_3-R-U_5. The U_3 sequences are from the $3'$ end of the RNA, and the U_5 sequences are from the $5'$ end of viral RNA.

Figure 4 shows a possible mechanism for the synthesis of this DNA structure from viral RNA, making use of the RNA direct repeats, the tRNA primer, the ribonuclease H activity of the reverse transcriptase, movement of the DNA product from one end of the viral RNA to the other end of another viral RNA (there are two identical molecules of viral RNA in a retrovirus virion), and specific priming of the plus strand of the viral DNA, by a polypurine tract next to U_3. The steps in this model for the synthesis of retrovirus DNA are as follows: (*a*) Viral RNA. RU_5PB and U_3R represent two contiguous sequences at the $5'$ and $3'$ ends of the RNA, respectively. R is a terminal repeat in the RNA; PB (primer binding) is a sequence complementary to the end of
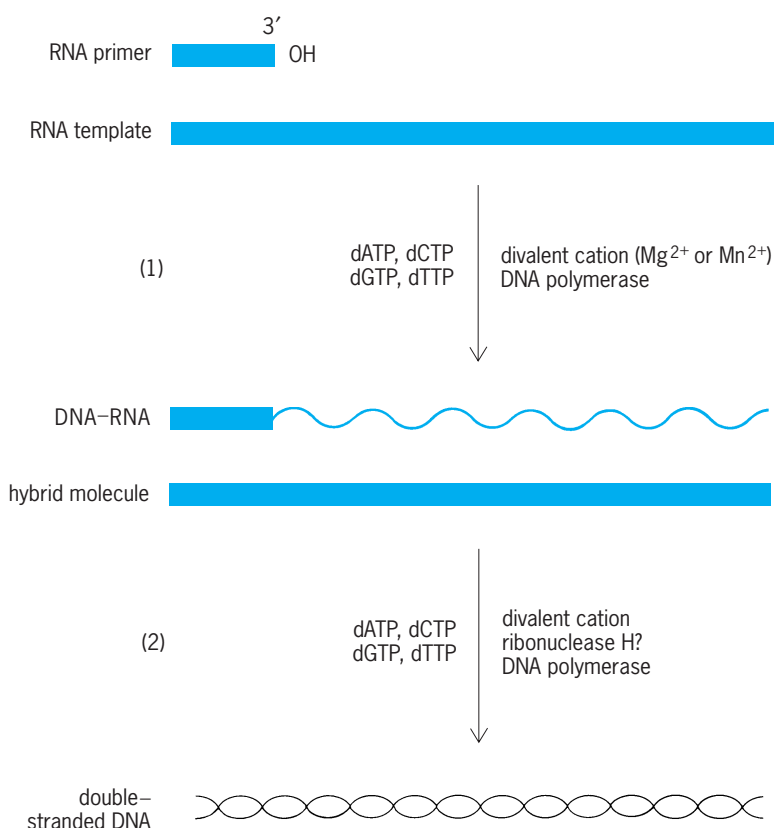

Fig. 3. Two reactions in retrovirus RNA-directed DNA synthesis. The viral RNA primer is actually attached close to the $5'$ end of the viral RNA template.
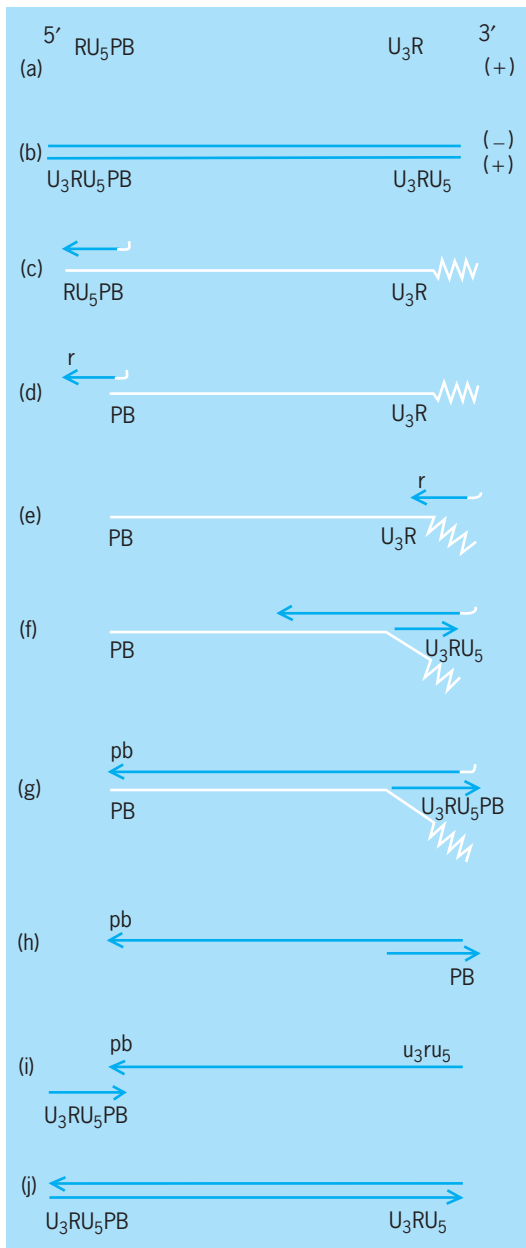
Fig. 4. Model for the synthesis of retrovirus DNA.
Uppercase letters represent the plus strand; lowercase letters represent the minus strand, RNA is shown in white, DNA in color.

a cellular tRNA, the primer. (*b*) Unintegrated viral DNA. The steps in its synthesis from *a* are as follows: (*c*) Viral DNA synthesis begins with minus-strand synthesis using the tRNA as a primer and proceeding through U_5 and R to the 5′ end of the RNA. (*d*) Viral ribonuclease H digests the part of the RNA template (viral RNA) that is hydrogen-bonded to minus-strand DNA. (*e*) The newly synthesized minus-strand DNA hydrogen- bonds to the repeat (R) at the 3′ end of the other molecule of viral RNA. (*f*) Minus-strand DNA synthesis proceeds using viral RNA as a template. Plus-strand DNA synthesis is initiated by using ribonuclease H-cut viral RNA as a primer and minus-strand DNA as a template. (*g*) Minus- and plus-strand viral DNA synthesis proceed to the end of their

respective templates. (*h*) Viral ribonuclease H further degrades the template RNAs, exposing homologous sequences, pb and PB on the minus- and plus-strand DNAs, respectively. (*i*) The plus-strand and minus-strand DNAs hydrogen-bond at PB. (*j*) Synthesis of the unintegrated viral DNA is completed.

Evolution. Studies indicate that reverse transcriptase is widely distributed in living organisms and that all reverse transcriptases are evolutionarily related. For example, the organization of the nucleotide sequence of integrated retroviral DNA described in Fig. 5 has a remarkable resemblance to the structure of bacterial transposable elements, in particular, transposons. Transposons have a five-base-pair repeat of bacterial DNA next to an inverted repeat of transposon DNA, which often is in a direct repeat of transposon DNA. The similarity of this structure to that of retroviruses indicates that these genetic elements are very similar and that it is not possible to determine from the structure of an integrated retrovirus or transposon whether this structure was formed as a result of reverse transcription or of transposition involving DNA only. *See* TRANSPOSONS.

These organizational similarities supported the hypothesis that retroviruses evolved from cellular movable genetic elements such as transposons. More direct support for this hypothesis has come from cloning and sequencing of reverse transcriptase genes and movable genetic elements from *Drosophila*, yeast, and many other organisms, including even myxobacteria (see table). Reverse transcriptase genes are present in the eukaryotic organisms in retrotransposons, which are DNA sequences with long terminal repeats (LTRs) and reverse transcriptase coding sequences, and in retroposons or long interspersed (LINE) elements, which are DNA sequences without LTRs but with reverse transcriptase coding sequences. Both of these types of elements can transpose in cells. In bacteria, coding sequences for reverse transcriptase are found as part of DNA sequences coding for unusual small molecules composed of both RNA and DNA and
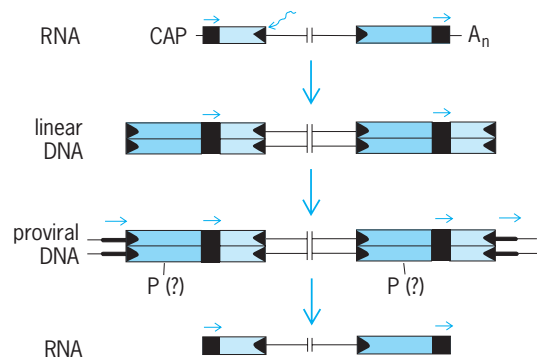

Fig. 5. Schematic representation of the structures at the termini of retroviral nucleic acid molecules. Solid areas represent terminal sequences at either terminus which are directly repeated in the RNA (as indicated by the arrows above these areas in the bottom RNA diagram). Shaded areas represent sequences unique to either end of the RNA. Heavy lines outside the long terminal repeats in the provirus represent direct repeats of cell DNA formed in the integration process.

called msDNA. The DNA in these molecules is synthesized by reverse transcriptase. The function of msDNA molecules is unknown. *See* DEOXYRI-BONUCLEIC ACID (DNA); RIBONUCLEIC ACID (RNA).

Howard M. Temin

Bibliography. A. J. Davison and R. M. Elliott (eds.), *Molecular Virology*, 1993; R. F. Doolittle et al., Origins and evolutionary relationships of retroviruses, *Quart. Rev. Biol.*, 64:1–30, 1989; S. P. Goff, Retroviral reverse transcriptase: Synthesis, structure, and function, *JAIDS*, 3:817–831, 1990; D. Shippen-Lentz and E. H. Blackburn, Functional evidence for an RNA template in telomerase, *Science*, 247:546–552, 1990; H. M. Temin, Retrons in bacteria, *Nature*, 339:254–255, 1989; H. Varmus, Retroviruses, *Science*, 240:1427–1435, 1988; R. Weiss et al. (eds.), *RNA Tumor Viruses: Molecular Biology of Tumor Viruses*, 2d ed., Cold Spring Harbor (New York) Laboratory, 1982.

Revetment

A facing or veneer of stone, concrete, or other materials constructed on a sloping embankment, dike, or beach face to protect it against erosion caused by waves or currents. The revetment may be a rigid cast-in-place concrete structure; but more commonly it is a flexible structure constructed of stone riprap or interlocking concrete blocks. It is sometimes an articulated block structure where the armor blocks are set in a form known as a flexible carpet; that is, the blocks interlock for stability, but the interlocking makes them flexible enough to respond to settlement of the underlying soil. A flexible revetment provides protection from exterior hydraulic forces, and it also can tolerate some settlement or consolidation of the underlying soil.

Components. A typical revetment might employ stone riprap as the armor material (see **illus.**). A revetment typically has three major components: (1) the armor layer, which resists the wave or current-induced hydraulic forces; (2) a filter layer under the armor layer to allow water seepage out of the underlying soil without the removal of fine soil particles; and (3) a mechanism to stabilize the structure toe. Toe stabilization is particularly important where
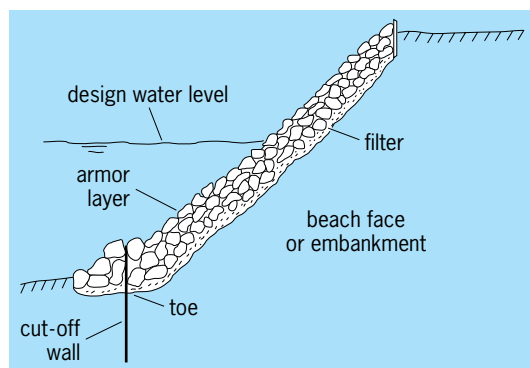
Cross-sectional profile of a typical stone revetment.

waves break on the structure, but may not be necessary if the revetment extends to sufficient depths where hydraulic forces will not erode the toe of the slope. The design water level (see illus.) for the structure may be higher than the normal water level during nonstorm conditions. If the revetment is exposed to waves that will break and run up the face of the revetment, the upper extent of the revetment must be sufficiently high to counter the force exerted by the waves.

Materials. Although stone riprap is the most commonly used material for revetment armor layers, a wide variety of other materials have been used, including cast-in-place concrete and poured asphalt, wire bags filled with stone (gabions), interlocking concrete blocks, soil cement, cement-filled bags, interlocked tires, woven wooden mattresses, and vegetation. Vegetation would only be used for surfaces exposed to very low waves or slow-moving currents.

A rough porous armor layer with a surface that is as uniform as possible has advantages when the revetment is exposed to wave action. This surface condition will decrease the elevation of wave runup on the revetment face and thus allow for a lower structure for the top elevation. Armor units that protrude above the surface of the revetment are exposed to stronger hydraulic forces and thus are more prone to removal from the structure, leading to possible structure failure.

An advantage that a revetment with a stone riprap armor layer has over an armor layer consisting of interlocking concrete units is that it is self-healing to some extent. If a few armor stones are dislodged, others will shift into the vacated position. Stone revetments also are generally more esthetically pleasing as they are more natural in appearance.

Design geometry. Typical reveted slopes vary in vertical to horizontal proportion between 1:2 and 1:4. A steeper slope requires less material for the revetment. However, the armor layer will be more stable on a flatter slope, as will the underlying soil embankment. Other considerations such as revetment constructability, slope accessibility for maintenance and repair, and movement over the revetment to access a beach may also affect the design geometry.

Filter layer. The filter layer typically consists of one or more layers of finer stone or a woven plastic fabric or a combination of both. While the filter must be sufficiently fine to prevent embankment material from being pulled through the armor layer, it must not be so fine that there is a buildup of excessive pore water pressure in the embankment. Stone filter layers also distribute the load from larger armor stone in the armor layer. *See* GEOSYNTHETIC.

Stabilization. The revetment toe may need to be stabilized. This can be accomplished by maintaining sufficient soil or beach sand at the toe. This would involve placing sand fill with or without groins at a wave-exposed shore or constructing spur dikes perpendicular to the bank of a river to trap river sediment. Or the toe may be stabilized by placement of a cut-off wall (see illus.) or larger armor stone that prevents the toe from being undermined. If the

revetment toe fails, the entire structure is prone to failure.

At some point along the river or shoreline, the revetment must terminate or merge with other structures. This point is often critical to the stability of the entire structure. If the revetment just ends at a natural unprotected section of embankment, the revetment must be stabilized to allow for the possible erosion of the adjacent natural embankment without undermining the revetment. This stabilization may involve the placement of a more massive segment of revetment or the construction of a cut-off wall perpendicular to the embankment surface that extends well back into the embankment. *See* RETAINING WALL; RIVER ENGINEERING.

R. M. Sorensen

Bibliography. M. S. Peterson, *River Engineering*, 1986; L. L. Whiteneck and L. A. Hockney, *Structural Materials for Harbor and Coastal Construction*, 1989.

Reynolds number

The number Re, a dimensionless parameter that determines the behavior and characteristics of viscous flow patterns. It is defined by Eq. (1), where ρ is

$$\mathrm{Re} = \frac{\rho V L}{\mu} \qquad (1)$$

fluid density, V is stream velocity, L is a characteristic length scale, and μ is fluid viscosity. This parameter, the dominant factor in viscous flow analysis, was formulated in 1883 by Osborne Reynolds and named in his honor about 40 years later.

The form of Eq. (1) is such that Re has the same value regardless of the unit system used to define its constitutents. In the International System (SI), for example, if brackets denote the dimensions of a variable, Eq. (2) is obtained. Thus the Reynolds number

$$[\mathrm{Re}] = \frac{[\mathrm{kg/m^3}][\mathrm{m/s}][\mathrm{m}]}{[\mathrm{kg/m \cdot s}]} = [1] \qquad (2)$$

is not subject to any ambiguity.

Laminar and turbulent flow. It had been well known since the eighteenth century that viscous flows were sometimes smooth (laminar) and sometimes disorderly (turbulent). Reynolds revealed the reason in 1883 in a classic experiment which introduced a dye streak down the centerline of water flowing in a glass tube (see **illus.**). For values of Re less than 2000 (illus. *a*) the dye streak remained smooth and undisturbed—that is, laminar flow resulted. For Re greater than 3000 (illus. *b*), the dye streak seemed to burst into turbulence and fill the whole tube with color. However, a spark-flash photograph (illus. *c*) showed that the turbulence structure was really a series of minute whirls and eddies, oscillating too fast for the eye to see. This classic demonstration laid the foundation for what is now a rich field of turbulent flow analysis and experimentation. *See* LAMINAR FLOW; TURBULENT FLOW.

For Re between 2000 and 3000 (not shown in the illustration), the flow pattern cycled erratically be-

(a)   (b)

(c)
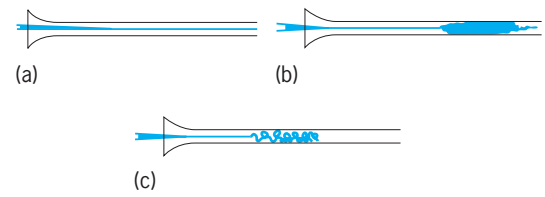
Reynolds's original sketches of dye streak behavior in pipe flow: (*a*) laminar flow, Re > 2000; (*b*) turbulent flow, Re > 3000; (*c*) a spark-flash photograph of condition *b*. (*After O. Reynolds, An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous and of the law of resistance in parallel channels, Phil. Trans. Roy. Soc., London, A174:935–982, 1883*)

tween laminar and turbulent regimes. This is now known as the transition region between the two more stable flow patterns.

The transition Reynolds number is not always equal to 2000 but rather varies somewhat with flow geometry. It is less than 2000 for unstable patterns, such as jet flow, and greater than 2000 for highly stable flows such as converging nozzles. *See* BOUNDARY-LAYER FLOW; JET FLOW.

Model testing. In experimenting with model tests of viscous flows, the Reynolds number is the primary parameter by which the model and prototype are matched. Fulfillment of Eq. (3) ensures the dynamic

$$\left(\frac{\rho V L}{\mu}\right)_{\mathrm{model}} = \left(\frac{\rho V L}{\mu}\right)_{\mathrm{prototype}} \qquad (3)$$

similarity of the model test. One can vary density, velocity, length, or viscosity in any manner as long as the overall grouping $(\rho V L/\mu)$ remains the same. Of course, the model must also be geometrically similar to the prototype, that is, must have exactly the same proportions and orientation to the fluid stream. Then model test data can be used to predict the drag, pressure distribution, friction, and even the convection heat transfer from the prototype, if the data are properly interpreted. An exception is free surface flow—such as ship resistance, ocean wave motion, and buoy dynamics—where the Froude number of the flow is also extremely important. *See* DIMENSIONAL ANALYSIS; DYNAMIC SIMILARITY; FLUID FLOW; FROUDE NUMBER; MODEL THEORY.

Frank M. White

Bibliography. H. Rouse and S. Ince, *History of Hydraulics*, 1963; H. Schlichting, *Boundary Layer Theory*, 8th ed., 2000; I. H. Shames, *Mechanics of Fluids*, 3d ed., 1992; F. M. White, *Fluid Mechanics*, 4th ed., 1998.

Rh incompatibility

A condition in which red blood cells of the fetus become coated with an immunoglobulin (IgG) antibody [anti-Rh D antibody] of maternal origin which is directed against an antigen (Rh D antigen) of paternal origin that is present on fetal cells. *See* ANTIBODY; ANTIGEN; IMMUNOGLOBULIN.

Etiology. A pregnant woman may develop an antibody to a red blood cell antigen that the fetus has but

that she does not possess. This occurs because the fetus has inherited the antigen from the father; fetal red blood cells pass into the maternal circulation and are recognized as foreign by the mother. The mother develops antibodies to the foreign protein; the antibodies pass across the placenta and bind to the fetal red blood cells. The fetal mononuclear phagocytic system recognizes the antibody-sensitized red blood cells as foreign and removes them from the circulation producing a hemolytic anemia. The maternal immune response is known as red blood cell alloimmunization. Alloimmunization of the mother results in hemolytic disease of the fetus and newborn (also known as erythroblastosis fetalis). It is important to understand that hemolytic disease of the fetus and newborn can occur from an incompatibility to a number of blood group antigens. The incidence of Rh incompatibility leading to hemolytic disease of the newborn has decreased dramatically from the development of passive immunization to prevent maternal Rh D alloimmunization. *See* ANEMIA; AUTOIMMUNITY.

Blood group nomenclature. Rh antigens are essential to maintain structural stability of the red cell membrane and are composed of protein and lipid (proteolipid). The Rh blood group system has a two gene loci (RHD and RHCE), and a combination of Rh alleles is inherited from each parent. The Rh antigen of greatest importance is the D antigen. If both D genes are inherited, the individual is said to be homozygous Rh positive. If one D gene is inherited, the blood type will be Rh positive and the individual is said to be hemizygous due to the deletion of one D gene from one chromosome. If neither D gene is present, the individual will be Rh negative. Total absence of both the RHD and RHCE genes results in the condition known as Rh null; it is very rare. *See* BLOOD GROUPS.

Risk of sensitization. The development of Rh incompatibility is determined by the Rh types of the parents. If the mother is Rh positive, there is no risk of Rh D incompatibility. If the mother is Rh negative, the risk to the fetus depends upon its blood type, which in turn is related to the blood type of the father. If the father is Rh negative, there is no risk to the fetus because all fetuses will be Rh negative. If the father is hemizygous Rh positive, half of the offspring will be Rh positive and, following sensitization of the mother, will be in danger of developing erythroblastosis. If the father is homozygous Rh positive, all fetuses will be Rh positive; if the mother has been sensitized after the first pregnancy, there will be the risk of hemolytic disease of the fetus and newborn for each subsequent pregnancy. The risk of alloimmunization to Rh is 16% with each ABO-compatible pregnancy and 1.5–2% with each Rh-positive ABO-incompatible pregnancy.

Some Rh-incompatible pregnancies do not result in sensitization. Factors influencing the occurrence of sensitization include the timing and extent of the transplacental hemorrhage, the degree and strength of antibody development in the mother, and the ABO status of mother and fetus. Most immunizations result from transplacental hemorrhage during placen-

tal separation at delivery. Smaller numbers of cells enter the mother's circulation during the last half of pregnancy. If the amount of cells is less than 0.1 milliliter throughout and after delivery, sensitization is less likely. ABO-incompatible Rh-incompatible fetal red blood cells are more easily recognized as foreign, due to preexisting ABO antibodies, so that the fetal red blood cells are rapidly destroyed in the maternal circulation and are less likely to elicit an antibody response. Because most sensitization occurs as a result of bleeding at delivery, hemolytic disease of the newborn almost never occurs during the first pregnancy, but subsequent Rh-positive pregnancies may result in an affected infant. Once immunization has occurred, successive Rh-positive pregnancies will result in hemolytic disease of the fetus and newborn of increasing severity. Because Rh negativity is less common in Blacks and is rare in Asians, Rh incompatibility is less likely to occur in these groups.

Pathogenesis. The symptoms and signs of hemolytic disease of the newborn from Rh incompatibility are due to the destruction (hemolysis) of the antibody-coated fetal red blood cells. The fetus responds to the hemolysis and anemia in a number of different ways. The products of hemolysis, specifically bilirubin, pass through the placenta and are cleared by the maternal circulation. The fetus compensates for the anemia by producing more red blood cells. The liver and spleen become the major sites of red blood cell production, and this is known as extramedullary hematopoiesis, the formation of blood cells outside the bone marrow. As the liver and spleen enlarge, their function decreases and serum protein levels fall. This results in anasarca, or body edema; hence the term hydrops fetalis—literally, water baby. Anemia results in heart failure, which contributes to the hydrops. In the severest forms, infants are stillborn or die soon after birth. Another major manifestation is jaundice, caused by increased amounts of the blood pigment bilirubin, derived from the breakdown of hemoglobin. Although much of the bilirubin passes across the placenta and is metabolized by the mother's liver, the newborn still may be perceptibly jaundiced, and this jaundice rapidly deepens after birth because the placenta is no longer present to clear the bilirubin. In less severe cases, the infant may be in good condition at birth but become jaundiced as the bilirubin increases in concentration. Indirect bilirubin that is not bound to albumin may cross the blood–brain barrier and affect certain brain cells, resulting in permanent brain damage. This condition of bilirubin encephalopathy is termed kernicterus. *See* BILIRUBIN; HEMATOPOIESIS; HEMOGLOBIN; JAUNDICE.

Diagnosis and management. The at-risk mother is monitored for the development and progression of Rh incompatibility with serial serum antibody titers to anti-D. The presence of anti-D and its titer do not necessarily correlate with the status of the fetus. To monitor the status, a procedure known as amniocentesis is performed in which amniotic fluid is collected by introducing a needle into the amniotic sac through the mother's abdominal wall. This

procedure permits measurement of bile pigment concentration and assessment of lung maturity. These tests are done serially to decide when to deliver the fetus or whether to transfuse the fetus until it is mature enough to be delivered safely.

Because invasive procedures to obtain fetal blood for prenatal blood group analysis may present a risk to the fetus, other methods have been explored. During pregnancy, cell-free fetal deoxyribonucleic acid (DNA) circulates in maternal blood plasma. Molecular techniques, especially real-time polymerase chain reaction, or RT-PCR, can be used to identify copies of the product of the RHD gene and to accurately predict fetal D status because the genome of the Rh-negative mother does not contain an RHD gene. In another technique, Doppler assessment of fetal middle cerebral artery blood flow is used as a reliable noninvasive predictor of fetal anemia. Peak systolic velocity increases with decreased red cell volume (that is, anemia), which is assessed in the proximal region of the artery identified by cross-sectional ultrasonography. *See* PRENATAL DIAGNOSIS.

Transfusions. Treatment of the fetus and infant depends upon the severity of the disease. Intrauterine transfusions can be administered as early as 18 weeks' gestation and are given by introducing a needle into the fetal abdominal cavity or by catheterizing a fetal umbilical vein, using ultrasound to guide the needle placement. This procedure is called PUBS (percutaneous umbilical blood sampling); both sampling and intrauterine transfusions can be administered through PUBS. Such transfusions may be done weekly or bimonthly. The blood used for the transfusion is washed, type-O, Rh-negative, cytomegalovirus (CMV) "safe" blood, and it may be specially processed in other ways.

Following delivery, the infant may require transfusion immediately or during the postnatal period. This may involve simple replacement or exchange transfusion, in which the infant's blood volume is replaced with compatible blood. The aim of exchange is to remove bilirubin and antibody-coated red blood cells and to correct the anemia while replacing the infant's blood with blood that is antigen negative for the offending antibody. Other medical treatment for jaundice may also be required. *See* TRANSFUSION.

Rh immune globulin. Rh immune globulin (RhIG) is a purified solution of IgG containing anti-D obtained from pooled human plasma of deliberately alloimmunized donors or previously pregnant women alloimmunized to the Rh D antigen. A 1-ml dose contains 300 micrograms of RhIG, and it will counteract the immunizing effects of 30 ml of Rh-positive fetal red blood cells. The mechanism of action is not well understood but is thought to interfere with antigen recognition by the maternal immune system. Because of its method of preparation, it will not transmit hepatitis or human immunodeficiency virus (HIV) infection. Special tests can be performed on the mother's blood after delivery to quantify the amount of transplacental hemorrhage. RhIG should be administered to any Rh-negative mother of an Rh-positive infant within 72 h after delivery. In cases where there is bleeding of more than 30 ml of fetal cells, additional RhIG is usually administered. Administration of RhIG at the twenty-eighth week of pregnancy followed by a second dose at birth has been very successful in all but eliminating Rh hemolytic disease. Anti-Rh immunoglobulin should also be administered in a lower dose to Rh-negative women following amniocentesis, abortion, ectopic pregnancy, and fetal death if the status of the father is Rh positive or unknown. If the pregnancy is terminated after 13 weeks of gestation, a full dose (300 mg) should be administered. *See* PREGNANCY.

Naomi L. C. Luban; Chrysanthe Gaitatzes; Aimee M. Barton

Bibliography. J. M. Bowman, The prevention of Rh immunization, *Trans. Med. Rev.*, 2:129, 1988; P. A. Grannum and J. A. Copel, Prevention of Rh isoimmunization and treatment of the compromised fetus, *Seminar Perinatol.*, 12:324, 1988; K. H. Nicolaides, Studies on fetal physiology and pathophysiology in Rhesus disease, *Seminar Perinatol.*, 13:328, 1989.

Rhabditida

An order of nematodes in which the number of labia varies from a full complement of six to three or two or none. The tubular stoma may be composed of five or more sections called rhabdions. The three-part esophagus always ends in a muscular bulb that is invariably valved. The excretory tube is cuticularly lined, and paired lateral collecting tubes generally run posteriorly from the excretory cell; some taxa have anterior tubules also. Females have one or two ovaries; when only one is present the vulva shifts posteriorly. The cells of the intestine may be uninucleate, binucleate, or tetranucleate, and the hypodermal cells may also be multinucleate. Caudal alae, when present, contain papillae.

There are eight superfamilies in the order: Rhabditoidea, Alloionematoidea, Bunonematoidea, Cephaloboidea, Panagrolaimoidea, Robertioidea, Chambersiellolidea, and Elaphonematoidea. Of these, the Rhabditoidea not only is one of the largest nematode superfamilies but also contains many important parasites of humans and domestic animals. The group is distinguished by the well-developed cylindrical stoma and three-part esophagus that ends in a valved terminal bulb. In parasitic species, adult stages and some larval stages lack the valved terminal bulb. Though most species are free-living feeders on terrestrial bacteria, others are important in the biological control of insects or as parasites of mammals.

There are eight superfamilies in the order: Rhabditoidea, Alloionematoidea, Bunonematoidea, Cephaloboidea, Panagrolaimoidea, Robertioidea, Chambersiellolidea, and Elaphonematoidea. Of these, the Rhabditoidea not only is one of the largest nematode superfamilies but also contains many important parasites of humans and domestic animals. The group is distinguished by the well-developed cylindrical stoma and three-part esophagus that ends in a valved terminal bulb (see **illustration**). In
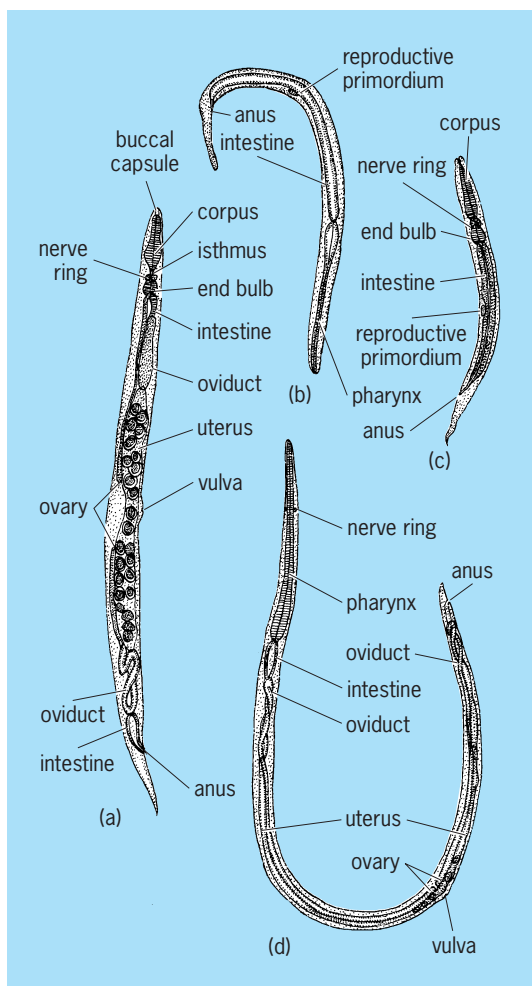
Strongyloides ransomi. (*a*) Free-living female, 1 mm long. (*b*) Filariform young. (*c*) Rhabdiform young. (*d*) Parasitic female, 4 mm long. (*After L. H. Hyman, The Invertebrates, vol. 3, McGraw-Hill, 1951*)

parasitic species, adult stages and some larval stages lack the valved terminal bulb. Though most species are free-living feeders on terrestrial bacteria, others are important in the biological control of insects or as parasites of mammals.

Neoaplectana, a genus whose known species are vectors of bacterial insect diseases, are of great interest in biological control schemes. A common bacterium transmitted is *Xenorhabdus nematophilus*. This bacterium is carried from generation to generation of nematodes in a pouch or expanded region at the anterior end of the nematode's intestine. The nematode, upon penetrating into the insect's hemocoel, ejects the bacteria from its alimentary canal. Once released, the bacteria multiply rapidly in the insect, killing it in less than 48 h. The nematode undergoes no development until this point; then it begins to feed on the bacteria, and development proceeds to adulthood. The nematode may go through several generations; eventually, when food is depleted, development is arrested in the infective stage, which once agains retains a bacterial bolus. These infective larvae remain in this state until another insect is invaded.

Strongyloides stercoralis, commonly called the threadworm, is an important parasite of humans. It can sustain itself as either a free-living bacterial feeder or as a parasite. In the free-living state it is a fecal associate feeding on bacteria. When the fecal substrate or bacteria are exhausted, the nematode arrests its development in the third (infective) stage. Morphologically it is distinguished by its lack of the valved rhabditoid esophagus. *Strongyloides stercoralis* enters the definitive host (humans) by direct skin penetration. Upon penetration it makes its way through the circulatory system to the lungs; from the lungs it is coughed up and swallowed. Upon reaching the small intestine it develops to adulthood. The eggs hatch in the intestine, and the released larvae may remain rhabditiform (free living) or metamorphose to the infective filariform that can immediately reinfect the same victim or be passed out to seek new victims. The rhabditiform larvae, when passed out with the feces, begin the free-living cycle. Human symptoms may include itching, from skin penetration by larvae; pneumonia, from lung penetration; or anemia, from blood loss due to adult penetration into the intestinal mucosa. *See* MEDICAL PARASITOLOGY; NEMATA.   Armand R. Maggenti

Rhenium

A chemical element, Re, with atomic number 75 and atomic weight 186.2. Rhenium is a transition element. It is a dense metal (21.04) with the very high melting point of $3440°$C ($6220°$F). *See* PERIODIC TABLE.

Rhenium is similar to its homolog technetium in that it may be oxidized at elevated temperatures by oxygen to form the volatile heptoxide, Re_2O_7; this in turn may be reduced to a lower oxide, ReO_2. The compounds ReO_3, Re_2O_3, and Re_2O, are well known. Perrhenic acid, $HReO_4$, is a strong monobasic acid and is only a very weak oxidizing agent. Complex perrhenates, such as cobalt hexammine perrhenate, $[Co(NH_3)_6(ReO_4)_3]$, are also known.

The halogen compounds of rhenium are very complicated, and a large series of halides and oxyhalides have been reported. Rhenium forms two well-characterized sulfides, Re_2S_7 and ReS_2, as well as two selenides, Re_2Se_7 and $ReSe_2$. The sulfides have their counterparts in the technetium compounds, Tc_2S_7

1																	18
1	2											13	14	15	16	17	**2**
H																	**He**
3	4											5	6	7	8	9	10
Li	**Be**											**B**	**C**	**N**	**O**	**F**	**Ne**
11	12											13	14	15	16	17	18
Na	**Mg**	3	4	5	6	7	8	9	10	11	12	**Al**	**Si**	**P**	**S**	**Cl**	**Ar**
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
K	**Ca**	**Sc**	**Ti**	**V**	**Cr**	**Mn**	**Fe**	**Co**	**Ni**	**Cu**	**Zn**	**Ga**	**Ge**	**As**	**Se**	**Br**	**Kr**
37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
Rb	**Sr**	**Y**	**Zr**	**Nb**	**Mo**	**Tc**	**Ru**	**Rh**	**Pd**	**Ag**	**Cd**	**In**	**Sn**	**Sb**	**Te**	**I**	**Xe**
55	56	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86
Cs	**Ba**	**Lu**	**Hf**	**Ta**	**W**	**Re**	**Os**	**Ir**	**Pt**	**Au**	**Hg**	**Tl**	**Pb**	**Bi**	**Po**	**At**	**Rn**
87	88	103	104	105	106	107	108	109	110	111	112	113					
Fr	**Ra**	**Lr**	**Rf**	**Db**	**Sg**	**Bh**	**Hs**	**Mt**	**Ds**	**Rg**							

lanthanide series	57	58	59	60	61	62	63	64	65	66	67	68	69	70
	La	**Ce**	**Pr**	**Nd**	**Pm**	**Sm**	**Eu**	**Gd**	**Tb**	**Dy**	**Ho**	**Er**	**Tm**	**Yb**

actinide series	89	90	91	92	93	94	95	96	97	98	99	100	101	102
	Ac	**Th**	**Pa**	**U**	**Np**	**Pu**	**Am**	**Cm**	**Bk**	**Cf**	**Es**	**Fm**	**Md**	**No**

and TcS_2. *See* TECHNETIUM; TRANSITION ELEMENTS.

Sherman Fried

Bibliography. R. Alberto, *Technetium Rhenium*, vol. 176 in *Topics in Current Chemistry*, Springer, 1995; R. Albrecht, *Re Rhenium-Organorhenium Compounds: Binuclear Compounds* (*Gmelin Handbooks of Inorganic and Organometallic Chemistry*, pt. 7), 8th ed., Springer, 1997; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004.

Rheology

In the broadest sense of the term, that part of mechanics which deals with the relation between force and deformation in material bodies. The nature of this relation depends on the material of which the body is constituted. It is customary to represent the deformation behavior of metals and other solids by a model called the linear or hookean elastic solid (displaying the property known as elasticity) and that of fluids by the model of the linear viscous or newtonian fluid (displaying the property known as viscosity). These classical models are, however, inadequate to depict certain nonlinear and time-dependent deformation behavior that is sometimes observed. It is these nonclassical behaviors which are the chief interest of rheologists and hence referred to as rheological behavior. *See* ELASTICITY; VISCOSITY.

Rheological behavior is particularly readily observed in materials containing polymer molecules which typically contain thousands of atoms per molecule, although such properties are also exhibited in some experiments on metals, glasses, and gases. Thus rheology is of interest not only to mathematicians and physicists, who consider it to be a part of continuum mechanics, but also to chemists and engineers who have to deal with these materials. It is of special importance in the plastics, rubber, film, and coatings industries. *See* FLUID MECHANICS; PAINT; PLASTICS PROCESSING; POLYMER; RUBBER; SURFACE COATING.

This article deals with three useful nonclassical models: linear viscoelasticity, nonlinear elasticity, and nonlinear viscoelasticity. Plasticity, of interest in metals, is rarely studied by rheologists. *See* PLASTIC DEFORMATION OF METAL; PLASTICITY.

Models and properties. To make clear the nature of the various rheological models, the discussion will focus on some very simple types of deformation. Simple shear will be discussed in most detail; similar phenomena and properties, for the most part, also occur in extension. *See* STRESS AND STRAIN.

Consider a block of material of height h deformed in the manner indicated in **Fig. 1**; the bottom surface is fixed and the top moves a distance w parallel to itself. A measure of the deformation is the shear strain γ given by Eq. (1).

$$\gamma = \frac{w}{h} \tag{1}$$

To achieve such a deformation if the block is a lin-
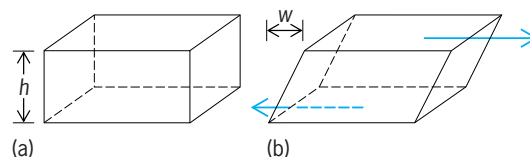

(a) (b)

Fig. 1. Simple shear. (*a*) Undeformed block of height *h*. (*b*) Deformed block after top has moved a distance *w* parallel to itself. The arrows indicate the net forces acting on the top and bottom faces. The forces which must be applied to left and right faces to maintain a steady state are not indicated.

ear elastic material, it is necessary to apply uniformly distributed tangential forces on the top and bottom of the block as shown in Fig. 1*b*. The intensity of these forces, that is, the magnitude of the net force per unit area, is called the shear stress S. For a linear elastic material, γ is much less than unity and is related to S by Eq. (2), where the proportionality

$$S = G\gamma \tag{2}$$

constant G is a property of the material known as the shear modulus.

If the material in the block is a newtonian fluid and a similar set of forces is imposed, the result is a simple shearing flow, a deformation as pictured in Fig. 1*b* with the top surface moving with a velocity dw/dt. This type of motion is characterized by a rate of shear $\dot{\gamma} = (dw/h)/dt$, which is proportional to the shear stress S as given by Eq. (3), where η is a

$$S = \eta\dot{\gamma} \tag{3}$$

property of the material called the viscosity.

Linear viscoelasticity. If the imposed forces are small enough, time-dependent deformation behavior can often be described by the model of linear viscoelasticity. The material properties in this model are most easily specified in terms of simple experiments: creep, stress relaxation, and sinusoidal deformation.

Creep and relaxation. In a creep experiment a stress is suddenly applied and then held constant; the deformation is then followed as a function of time. This stress history is indicated in the solid line of **Fig. 2***a* for the case of an applied constant shear stress S_0. If such an experiment is performed on a linear elastic solid, the resultant deformation is indicated by the full line in Fig. 2*b* and for the linear viscous fluid in Fig. 2*c*. In the case of elasticity, the result is an instantly achieved constant strain; in the case of the fluid, an instantly achieved constant rate of strain. In the case of viscoelastic materials, there are some which eventually attain a constant equilibrium strain (Fig. 2*d*) and hence are called viscoelastic solids. Others eventually achieve constant rate of strain (Fig. 2*e*) and are called viscoelastic fluids. If the material is linear viscoelastic, the deformation $\gamma(S_0,t)$ is a function of the time t since the stress was applied and also a linear function of S_0; that is, Eq. (4) is satisfied,

$$\gamma(S_0, t) = S_0 J(t) \tag{4}$$

where $J(t)$ is independent of S_0. The function $J(t)$ is
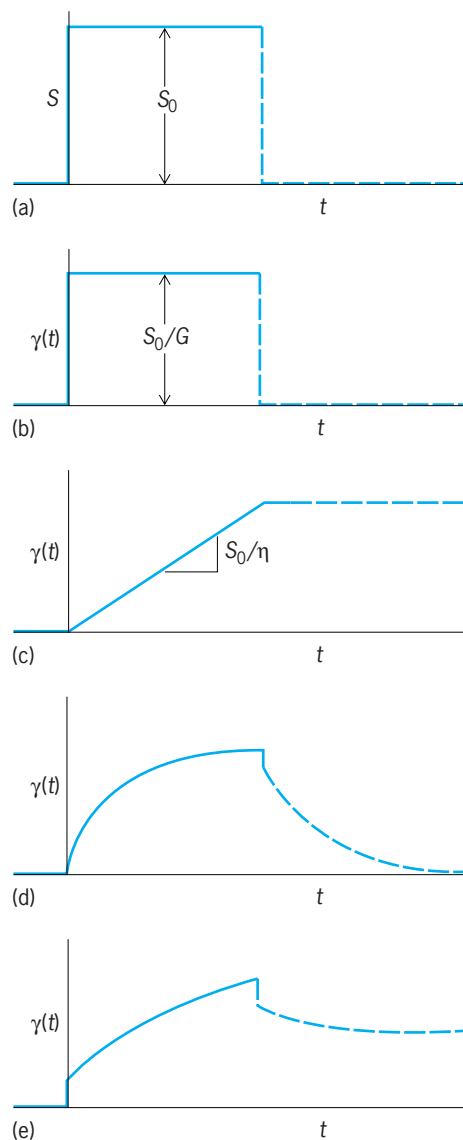
Fig. 2. Creep and recovery; solid lines indicate creep; broken lines indicate recovery. (*a*) Applied stress history. (*b*) Corresponding strain history for linear elastic solid, (*c*) linear viscous fluid, (*d*) viscoelastic solid, and (*e*) viscoelastic fluid.

a property of the material known as the shear creep compliance. Similar results are obtained for creep experiments performed in extension. *See* CREEP (MATERIALS).

The broken lines in Fig. 2 indicate the result when the stress is removed after a creep experiment. This is called a recovery experiment. The linear viscoelastic solid eventually goes back to its original undeformed shape while the fluid does not. However, the linear viscoelastic fluid recovers somewhat. The amount of recovery is taken as a measure of the elastic character of the fluid. Some of the energy which was expended on the fluid during the creep experiment is dissipated, but some is recovered. In contrast, with the newtonian fluid all the energy is dissipated and is converted to heat.

Stress relaxation. In a stress relaxation experiment, a deformation is suddenly imposed and then held con-

stant; the stress is observed as a function of time. The stress decreases with time until it reaches an equilibrium value. This equilibrium value is zero for a viscoelastic fluid, and some finite value for a viscoelastic solid. If the imposed deformation is a simple shear with a constant shear strain γ_0, the required shear stress $S(t,\gamma_0)$ depends on the time t since the strain was imposed, and is proportional to γ_0; that is, Eq. (5) is satisfied, where $G(t)$ is a property of the

$$S(t, \gamma_0) = \gamma_0 G(t) \qquad (5)$$

material, called the shear relaxation modulus, which does not depend on the strain. If the experiment performed is an extension, the corresponding property is the tensile relaxation modulus $E(t)$. Both $G(t)$ and $E(t)$ are functions which decrease as t increases.

Sinusoidal deformation. Linear viscoelasticity is useful for describing experiments where the deformation is a sinusoidal function of time. For example, if a sound wave (or other mechanical vibration) is transmitted through a thick sheet of a polymeric solid, the amplitude of the oscillation decreases as the wave progresses, and the speed of propagation and the rate of damping depend on the frequency of the oscillation and on the temperature. There usually is at least one frequency at which the rate of damping has a maximum.

Behavior of linear polymers. To illustrate some of the characteristics of linear viscoelastic properties, **Fig. 3** shows a schematic logarithmic plot of $J(t)$ for a high-molecular-weight polystyrene over a range of temperatures. This set of curves is typical of high-molecular-weight linear amorphous polymers, which include the commercial transparent thermoplastics such as poly(methyl methacrylate) (known under the trademarks Lucite and Plexiglas). These polymers are incapable of crystallization. However, they have a characteristic temperature, the glass transition temperature T_g, where the thermal expansion coefficient and the specific heat undergo a sudden change. Below T_g it takes an appreciable time to achieve an equilibrium volume if the material is subjected to a sudden decrease of temperature or a sudden change of pressure; the material is called
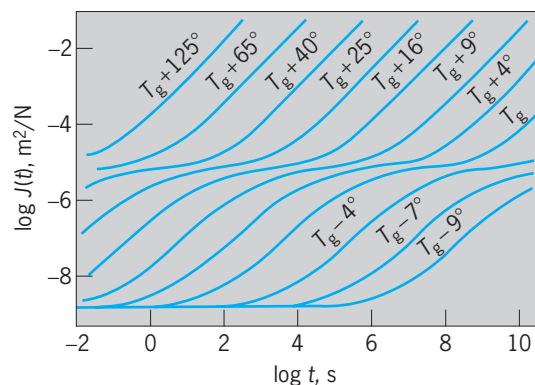

Fig. 3. Schematic shear creep compliance *J(t)* and its temperature dependence for a high-molecular-weight polystyrene. Temperatures are in degrees Celsius (1°C = 1.8°F) above or below the glass transition temperature T_g, which is 105°C (221°F).

a glass. At temperatures appreciably below T_g the polymer behaves like a hard, brittle elastic solid. *See* GLASS; GLASS TRANSITION; POLYACRYLATE RESIN; POLYSTYRENE RESIN.

In Fig. 3, it is seen that the main effect of temperature is to shift the curve along the log time axis. When a viscoelastic material has this property, the material is thermorheologically simple. Increasing the temperature does not change the processes which are occurring; it just makes them happen faster.

At long times, $J(t)$ is essentially proportional to time (the slope is unity in a logarithmic plot) and the polymer behaves much like a newtonian fluid. At temperatures considerably above T_g, this newtonian behavior begins at times less than a second. At such temperatures, the fluid polymer can be readily shaped by forcing it into a mold or extruding it through an orifice of the desired cross section. If the polymer is then cooled considerably below T_g, it becomes a stiff solid with very little creep over a time scale of hours or days. It behaves almost like an elastic solid with a $J(t)$ of about 10^{-9} m²/N, which corresponds to a Young's modulus of about 3×10^9 N/m² (4×10^5 lb/in.²). For critical and some long-term uses, even this small amount of creep may not be tolerable. Addition of solid fillers (for example, glass fibers) or replacement by a highly crystalline polymer (for example, high-density polyethylene) can reduce the amount of creep. *See* COMPOSITE MATERIAL; POLYOLEFIN RESINS.

It is in the neighborhood of T_g that viscoelastic effects are most readily evident and where they are most sensitive to temperature for experiments in ordinary laboratory time scales.

Polymers have hundreds or thousands of atoms chemically bonded to one another in long sequences. In most synthetic polymers, the presence of single C-C (or similar) bonds makes rotation about these bonds possible. As a result, such macromolecules can have many geometric shapes, most of them coiled conformations—like exceedingly long, thin coiled-up snakes. In an amorphous polymer, these snakelike molecules are highly intertwined. When forces are applied, they cannot immediately attain the steady-state response. Increasing the temperature does not change the process; it can, however, speed it up.

Nonlinear elasticity. A strip of lightly vulcanized rubber (for example, a rubber band) can be stretched to many times its undeformed length by application of a tensile force without breaking; when the force is removed, the strip returns to its original length. The equilibrium force-deformation curve (**Fig.** 4) is clearly nonlinear. Linear elasticity is obviously not applicable, and the model of nonlinear elasticity must be employed.

If a torque is applied to the ends of a rod, the rod is simply twisted if it is a linear elastic material. However, if the rod is made of rubber, it also increases in length. This normal stress effect is a natural consequence of nonlinear elasticity.

Addition of carbon black to a rubber increases the stiffness (Fig. 4). Such a rubber compound is called filled or reinforced rubber. After stretching a strip of

reinforced rubber for the first time and then removing the tensile force, the rubber is less stiff when it is stretched again. This is known as the Mullins effect and is attributed to the slippage and detachment of rubber molecules from the surface of the carbon black.

Because such high recoverable deformations are possible, a great deal of energy can be stored in rubber. This property has been used not only in toys (slingshots and motors for model airplanes) but also in supports for automobile bumpers designed so that they will absorb the impact of a low-velocity collision without damage to the vehicle.

Nonlinear viscoelasticity. If stresses become too high, linear viscoelasticity is no longer an adequate model for materials which exhibit time-dependent behavior. In a creep experiment, for example, the ratio of the strain to stress, $\gamma(t,S_0)/S_0$, is no longer independent of S_0; this ratio generally decreases with increasing S_0.

An important consequence of this nonlinearity occurs in the steady flow behavior of nonlinear viscoelastic fluids. In a steady simple shearing flow, the shear rate is not proportional to the shear stress S. In analogy with the case of the newtonian fluid [Eq. (3)], the ratio $S/\dot{\gamma}$ is also used to characterize the fluid in this nonlinear behavior, as in Eq. (6), and the
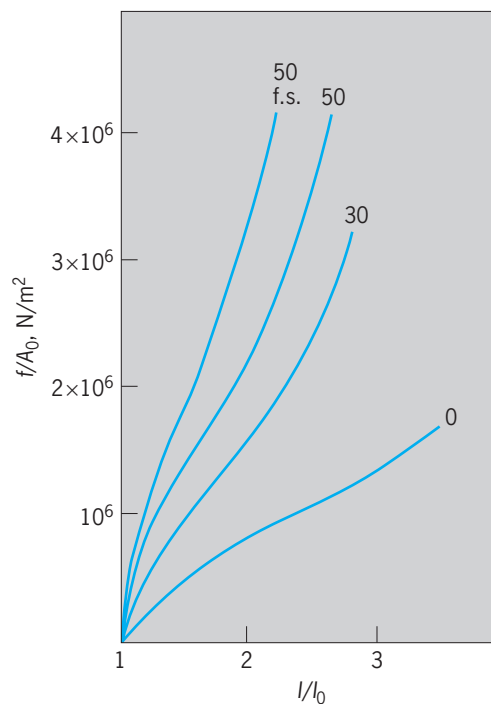
$$\frac{S}{\dot{\gamma}} = \eta(\dot{\gamma}) \qquad (6)$$

Fig. 4. Force-elongation data for a vulcanized synthetic rubber containing various amounts of carbon black (the number on the graph indicates parts of carbon black per 100 parts of rubber). The ordinate is the force *f* per initial cross-sectional area A_0. The abscissa is the ratio of the stretched length *l* to undeformed length l_0. The designation f.s. on one curve indicates that these data were obtained on the first stretching of the rubber strip; other curves show results after repeated stretching. (*After E. A. Meinecke and S. Maksin, Rubber Chemistry and Technology, 54:857, 1981*)
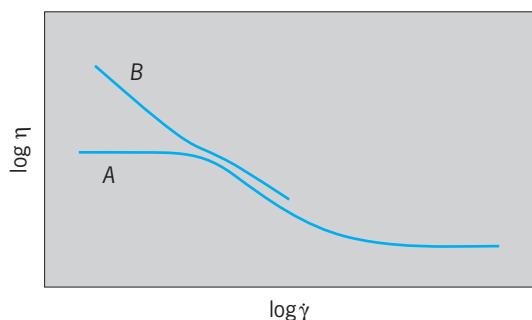
Fig. 5. Schematic logarithmic plot of the viscosity function $\eta(\dot{\gamma})$ for a polymer melt or concentrated polymer solution (curve A) and for the same polymer fluid with solid particles added (curve B).

symbol η has been adopted for this ratio, although now it is a function of $\dot{\gamma}$ as the notation indicates. It is called the viscosity function or sometimes simply the viscosity.

Shear thinning. For polymer melts, solutions, and suspensions, generally speaking, as indicated schematically in the logarithmic plot of **Fig. 5**, $\eta(\dot{\gamma})$ decreases as $\dot{\gamma}$ increases. This type of behavior, called shear thinning, is of considerable industrial significance. For example, paints are formulated to be shear-thinning. A high viscosity at low flow rates keeps the paint from dripping from the brush or roller and prevents sagging of the paint film newly applied to a vertical wall. The lower viscosity at the high deformation rates while brushing or rolling means that less energy is required, and hence the painter's arm does not become overly tired.

As $\dot{\gamma}$ goes to very small values for polymer melts and solutions, $\eta(\dot{\gamma})$ goes to a limiting value η_0, called the zero shear viscosity, as shown in curve A of Fig. 5. For high-polymer melts, η_0 is a very sensitive function of molecular weight; for a homologous series of linear polymers, η_0 is proportional to $M^{3.4}$; if the molecular weight is doubled, for example, η_0 is increased by a factor of about 10. Temperature also has a great effect on η_0, particularly in the neighborhood of T_g, where it can decrease by 75% per degree Celsius (45% per degree Fahrenheit).

On addition of solid particles (such as glass fibers or finely divided silica) to molten polymers or concentrated polymer solution, $\eta(\dot{\gamma})$ changes character as indicated in curve B of Fig. 5. A much larger effect is seen at low $\dot{\gamma}$ than at higher γ. Instead of approaching a limit at low $\dot{\gamma}$, $\eta(\dot{\gamma})$ keeps increasing as $\dot{\gamma}$ decreases. Some scientists believe that such suspensions will not flow unless the shear stress exceeds a value called the yield stress.

Normal stress effect. For newtonian fluids, steady simple shearing can be maintained by simple shear stress alone (Fig. 1). For nonlinear viscoelastic fluids, normal stresses must also be applied. If they are not, the height of the block of fluid will increase. This normal stress effect is also responsible for the observation that, when a rod is rotated in a container of some fluids (for example, some cake batters in a kitchen mixer), the fluid "climbs" around the ro-

tating rod. With newtonian fluids, the liquid level becomes depressed near the rotating rod as the fluid moves outward due to centrifugal effects.

Die swell. When some molten plastics are forced through a tube at a slow rate, the diameter of the exiting stream can be considerably larger than that of the tube. Under similar flow conditions, newtonian fluids show a diameter increase of only a few percent. This phenomenon is known as die swell or extrudate expansion. In extrusion equipment for making plastic tubing and rods, the orifice must be made smaller than the desired size of the finished product. *See* EXTRUSION.

Drag reduction. When water is pumped through a pipe at rates high enough for the flow to be turbulent, the flow rate, for a given driving pressure, can be greatly increased by adding a very small amount (a few parts per million) of a very high-molecular-weight (greater than a million) polymer. This phenomenon has been named drag reduction. It has been used to increase the range of the water stream from a firehose while using the same pumping equipment on the firetruck. Drag reduction also was used to decrease the number of pumping stations required for the Trans-Alaskan oil pipeline. *See* TURBULENT FLOW.

Thixotropy. There are suspensions (for example, bentonite clay in water) which, after remaining at rest for a long time, act as solids; for example, they cannot be poured. However, if it is stirred, such a suspension can be poured quit freely. If the suspension is then allowed to rest, the viscosity increases with time and finally sets again. This whole process is reversibly; it can be repeated again and again. This phenomenon is called thixotropy. (The word thixotropy may be used in many different senses; a standard nomenclature for rheological phenomena and properties has not been universally adopted.) *See* GEL; NON-NEWTONIAN FLUID.   Hershel Markovitz

Bibliography. J. Ferguson, *Applied Fluid Rheology*, 1991; A. I. Lurie, *Non-Linear Theory of Elasticity*, 1990; C. W. Macosko, *Rheology: Principles, Measurements, and Applications*, 1994; R. I. Tanner, *Engineering Rheology*, 2d ed., 2000; Y. G. Yanovsky, *Polymer Rheology: Theory and Practice*, 1993.

Rheostat

A variable resistor constructed so that its resistance value may be changed without interrupting the circuit to which it is connected. It is used to vary the current in a circuit. The resistive element of a rheostat may be a metal wire or ribbon, carbon disks, or a conducting liquid. *See* POTENTIOMETER; RESISTOR.
                                                Frank H. Rockett

Rheumatic fever

An illness that follows an upper respiratory infection with the group A streptococcus (*Streptococcus pyogenes*) and is characterized by inflammation of the joints (arthritis) and the heart (carditis). Arthritis typically involves multiple joints and may migrate from

one joint to another. The carditis may involve the outer lining of the heart (pericarditis), the heart muscle itself (myocarditis), or the inner lining of the heart (endocarditis). A minority of affected individuals also develop a rash (erythema marginatum), nodules under the skin, or Sydenham's chorea (a neurologic disorder characterized by involuntary, uncoordinated movements of the legs, arms, and face). These clinical manifestations may occur singly or in combination, and although most disappear completely, damage to heart valves may be permanent and progressive, leading to severe disability or death from rheumatic heart disease years after the initial attack.

Only about 3% of individuals with untreated acute streptococcal pharyngitis ("strep throat") develop rheumatic fever. The disease occurs an average of 19 days after the infection and is thought to be the result of an abnormal immunologic reaction to the group A streptococcus. Initial attacks of rheumatic fever generally occur among individuals aged 5 to 15. Those who have had one attack are highly susceptible to recurrences after future streptococcal infections.

The incidence of rheumatic fever declined steadily in the United States and other developed countries following World War II. During the 1980s, however, outbreaks of rheumatic fever were reported in a number of communities, the largest occurring in Salt Lake City and the surrounding area. In the late 1980s, two outbreaks occurred among recruits in military training camps. Streptococcal strains associated with these civilian and military outbreaks had biological characteristics similar to the strains that were responsible for rheumatic fever earlier in the century, suggesting that the resurgence of the disease may be partially explained by the reappearance of "rheumatogenic" streptococci into the population. Although outbreaks of rheumatic fever appear to have subsided for now in the United States, the disease remains rampant in developing countries and continues to be the most common cause of heart disease in children around the world.

Treatment of rheumatic fever is largely symptomatic and supportive. Aspirin is very effective in controlling the symptoms of arthritis, whereas corticosteroids may be required for management of severe carditis. Neither of these drugs prevents the development of chronic rheumatic heart disease. Initial attacks of rheumatic fever can be prevented by treatment of strep throat with penicillin for at least 10 days. Patients who have had an episode of rheumatic fever should continue taking antibiotics for many years to prevent group A streptococcal infections that may trigger a recurrence of rheumatic fever. Vaccines to prevent strep throat, and thus rheumatic fever, are in the early stages of human testing. *See* HEART DISORDERS; STREPTOCOCCUS.

A. L. Bisno; James B. Dale

Bibliography. B. F. Massell, *Rheumatic Fever and Streptococcal Infection: Unraveling the Mysteries of a Dread Disease*, Francis A. Countway Library of Medicine, Boston, 1997.

Rheumatism

Any combination of muscle or joint pain, stiffness, or discomfort arising from nonspecific disorders. It is generally used as a lay expression to indicate a chronic or recurrent condition affecting a certain area and precipitated by cold, dampness, or emotional stress.

Fibromyositis refers to acute or chronic symptoms of tenderness, stiffness, and pain in a joint and related structures which follows exposure, strain, trauma, infection, or stress. Myositis denotes an inflammation of a muscle; myalgia refers to muscle pain or tenderness without inflammation. Fibrositis is an inflammation of fibrous connective tissue, usually that of a joint region.

Rheumatism includes all of the above nonspecific disorders and is best reserved for complaints not related to a specific disease such as arthritis, rheumatic fever, muscle disorders, or others that may cause the same or similar symptoms.

Lumbago, wryneck, charley horse, and shin splints are commonly used expressions included under the catchall category of rheumatism.　　Robert Searles

Rhinoceros

An odd-toed ungulate (order Perissodactyla) belonging to the family Rhinoceratidae. The bodies and limbs of these mammals are massive and thick-skinned. Also characteristic of the family are the horns (epidermal derivatives), of which there may be one or two, composed of a solid mass of hairs attached to a bony prominence on the skull. The feet are tridactyl, with three hooves; the middle one is the most developed. Five species are found in Asia and Africa (see **table**).

Sight is poorly developed in these animals; however, hearing is quite good and the sense of smell is excellent. Most species are solitary and aggressive, and are extremely dangerous since they are unpredictable in behavior. All species are vegetarians. The great Indian rhinoceros (*Rhinoceros unicornis*) is about 14 ft (4.3 m) long and nearly 6 ft (1.8 m) high, with a single horn that may be 2 ft (0.6 m) long in both sexes. This species, found in southern Asia, is usually solitary and inoffensive. The gestation period is 16–17 months and the single offspring weighs about 100 lb (45 kg) at birth. The White rhinoceros (*Ceratotherium simum*) is found in Africa and does not stray far from water, which it requires not only for drinking but also for bathing. It is a diurnal, peaceful, and sociable species and is usually seen grazing in small groups. It is the only rhinoceros that is entirely a grazer. The Black rhinoceros (*D. bicornis*) is becoming scarce in its natural range near the East African lakes and is found mainly in game preserves (see **illustration**).

Some species show a strict territoriality, having definite tracks through the grass, definite sites for dung deposition, and personal mud wallows. Mating

Distribution and characteristics of the rhinoceros	
Name	Characteristics and distribution
Rhinoceros unicornis (*great Indian rhinoceros*)	Largest Asian species; thick skin with deep folds covered with scalelike granulations; South Asia
Rhinoceros sondaicus (Javan rhinoceros)	Rare; single short horn; granular skin; Java
Dicerorhinus sumatrensis (Sumatran rhinoceros)	Small (about $4^1/_2$ ft or 1.4 m high); two horns; Sumatra, Thailand, and Borneo
Dicerorhinus simus (white rhinoceros)	Two horns; square upper lip; flattened nose; Africa
Diceros bicornis (black rhinoceros)	Large (11 ft or 3.4 m long, 5–6 ft or 1.5–1.8 m high); two horns, the one on the nose being larger (about 2 ft or 0.6 m long) than the one behind it; Africa

may occur at any time of the year. Female White and Indian rhinoceroses breed for the first time when 6 to 8 years old; female Black rhinoceroses breed for the first time when 5 to 6 years of age. Males are usually several years older than females before they begin breeding. Gestation in most species lasts for 16 to 17 months with females normally giving birth to a single young every two to four years. Average longevity is 40 to 45 years. Major predators include humans, lions, and hyenas.

Black rhinoceros (*Diceros bicornis*). (*Dr. Lloyd Glenn Ingles,* © *1999 California Academy of Sciences*)

The gradual disappearance of the rhinoceros is attributed to habitat destruction and poaching, especially the unrestrained hunting for the horns (which were used as a medicine, although of no medicinal value), the meat, and the hide. The known Sumatran rhinoceros population now consists of only about 300 animals; only about 60 Javan rhincoeroses remain. The Javan, Black, and Sumatran species are classified as critically endangered by the IUCN (International Union for Conservation of Nature) and Endangered by the U. S. Department of the Interior; the Indian rhinoceros is classifed as endangered; and one population of White rhinoceros in the Democratic Republic of Congo is classified as critically endangered. *See* PERISSODACTYLA; TERRITORIALITY.

Donald W. Linzey

Bibliography. *Grzimek's Encyclopedia of Mammals*, McGraw-Hill, 1990; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford Ltd., 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., The Johns Hopkins University Press, 1999.

Rhinovirus

A genus of the family Picornaviridae. Members of the human rhinovirus group include at least 113 antigenically distinct types. Like the enteroviruses, the rhinoviruses are small (17–30 nanometers), contain ribonucleic acid (RNA), and are not inactivated by ether. Unlike the enteroviruses, they are isolated from the nose and throat rather than from the enteric tract, and are unstable if kept under acid conditions (pH 3–5) for 1–3 h. Most strains can be recovered by culture only in cells of human origin; they grow better when cultures are rolled (rather than held stationary) and are kept at a temperature of 91.5°F (33°C). *See* ENTEROVIRUS; TISSUE CULTURE.

Rhinoviruses have been recovered chiefly from adults with colds and only rarely from patients with more severe respiratory diseases. *See* COMMON COLD.

Rhinoviruses occur throughout the world. Natural immunity exists but is transitory. Only 30–50% of volunteers can be infected; yet the "resistant" volunteers may catch cold at other times. After being infected by one type of rhinovirus, volunteers resist the infecting type for as much as 16 weeks, and other types of rhinovirus for less than 5 weeks. *See* IMMUNITY.

In a single community, different rhinovirus types predominate during different seasons and during different outbreaks in a single season, but more than one type may be present at the same time.

Although efforts have been made to develop vaccines, none is available. Problems that hinder development of a potent and useful rhinovirus vaccine include the short duration of natural immunity even to the specific infecting type, the large number of different antigenic types of rhinovirus, and the variation of types present in a community from one year to the next. *See* ANIMAL VIRUS; PICORNAVIRIDAE; VIRUS CLASSIFICATION.   Joseph L. Melnick; M. E. Reichmann

Bibliography. R. W. Compans et al. (eds.), *Current Topics in Microbiology and Immunology*, vol. 16: *Picornaviruses*, 1990; D. Hamre, Rhinoviruses, *Monographs in Virology*, vol. 1, 1968; F. L. Horsfall and I. Tamm (eds.), *Viral and Rickettsial Infections of Man*, 4th ed., 1965.

Rhizocephala

An order of crustacean parasites related to the barnacles. Worldwide in distribution, they prey on other crustaceans, principally Decapoda, such as crabs, shrimp, and their allies. In 1949 the rhizocephalan *Loxothylacus texanus* (see **illus.**) infested 16% of the blue crab population in Arkansas Bay, Texas. Rhizocephala produce modifications affecting the abdomen of the crab, making males resemble females and causing immature females to acquire precociously the adult form. These parasites have become so modified by their mode of life that, as adults, they are no longer recognizable as barnacles, or even as crustaceans.

An adult rhizocephalan is a thin-walled sac enclosing a visceral mass, composed chiefly of ovaries and testes. It shows no trace of segmentation, appendages, or sense organs. Even an alimentary tract is missing. Instead, it possesses a threadlike root system which penetrates the interior of the host in all directions and absorbs nourishment from the body fluids of the crab. Fertilized eggs develop in a brood chamber. An opening at the summit of the sac allows seawater to enter for purposes of respiration and provides an avenue of escape for the larvae. A short stalk securely rivets the base of the sac to the abdomen of the crab.

Fertilized eggs become nauplii. These larval parasites, expelled from the mother's brood pouch, measure 0.01 in. (0.25 mm) in length. They have the appearance of barnacle nauplii, except that they lack an alimentary tract. Free-swimming nauplii metamorphose into cypris larvae which appear sexually indifferent. Those that settle on crabs become kentrogen larvae. A cell mass injected by the kentrogen develops into a tumor and root system on the intestine of the host. After 8 weeks the endoparasite emerges as a small sac. Cypris that settle on immature Rhizocephala become larval males. A cell mass, injected into the brood pouch by these males, migrates to the "testes" and differentiates into sperm. Six weeks later the sac is sexually mature.

Rhizocephala, with the exception of the family Sylonidae, were considered hermaphroditic, possessing ovaries and testes in the same individual. In 1958, A. Ichikawa and R. Yanagimachi proved
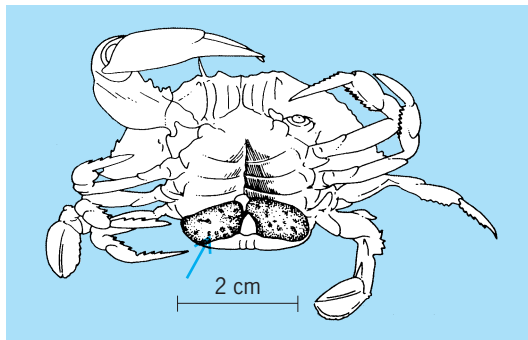
Loxothylacus texanus (arrow) attached to the abdomen of the blue crab (*Callinectes sapidus*). (*After E. G. Reinhard, Parasitic castration of Callinectes, Biol. Bull., 98(3):277–288, 1950*

that the "testes" in *Peltogasterella socialis* were really seminal receptacles. The sperm in these female repository organs is produced exclusively by cypris male cells. Adults of *Peltogasterella* are thus true females carrying hyperparasitic larval males. The hermaphroditic nature of other Rhizocephala is now questionable. *See* DECAPODA (CRUSTACEA).

Placidus G. Reischman

Bibliography. H. Boschma, The species of the genus *Sacculina* (Crustacea Rhizocephala), *Zool. Mededeel.*, 19(3–4):187–328, 1937; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; G. Smith, *Rhizocephala, Fauna und Flora des Golfes von Neapel*, vol. 29, 1906.

Rhizophorales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Rosidae of the class Magnoliopsida (dicotyledons). The order contains a single family, Rhizophoraceae, with about 100 species widely distributed in the tropics. The plants are mostly tanniferous trees and shrubs with the leaves opposite, simple, and entire. The flowers are regular, mostly perfect, and variously perigynous or epigynous. The sepals are four or five and commonly fleshy or leathery; the petals are the same number as the sepals and likewise small and fleshy. The stamens are twice as many as the petals or sometimes more. The pistil has 2–6 fused carpels with two or rarely more ovules per carpel; the fruit is berrylike or rarely a capsule. Most members of the family are inland species, but the most conspicuous group are some 17 species of shoreline shrubs, the mangroves. The fruits of mangroves are viviparous and have distinctive enlarged hypocotyls. The family has been subject to diverse interpretations, and has been associated with the Myrtales and with the Cornales. *See* MAGNOLIOPHYTA; MAGNOLIOPSIDA; PLANT KINGDOM.

T. M. Barkley

Rhizopodea

A class of Sarcodina including both parasitic and free-living species found in fresh and salt water and the soil. No species forms true axopodia; instead, pseudopodia may be filopodia, lobopodia, or reticulopodia; or there may be no pseudopodia in some cases. Rhizopodea include five subclasses: Lobosia, Fiosia, Granuloreticulosia, Mycetozoia, and Labyrinthulia. Lobosia are amebas and testate rhizopods with lobopodia. Filosia are naked or testate species with filopodia. Granuloreticulosia are noted for their reticulopodia, which often fuse into extensive networks. A bidirectional flow of cytoplasm typically moves granules into opposite directions along the two sides of a single pseudopodium. Mycetozoia are a heterogeneous collection, including both cellular and true slime molds. Labyrinthulia are unique protozoa which form no obvious pseudopodia, but move by still undetermined methods on or inside secreted tubules. Many Rhizopodea

are phagotrophic, but ingestion has not been described in Labyrinthulia. A test may be present or absent in different genera of Lobosia, Filosia, and Granuloreticulosia. *See* FILOSIA; GRANULORETIC-ULOSIA; LABYRINTHULIA; LOBOSIA; PROTOZOA; SAR-CODINA; SARCOMASTIGOPHORA.   Richard P. Hall

Bibliography. J. J. Lee et al., *An Illustrated Guide to the Protozoa*, 2d ed., Society of Protozoologists, 2000.

Rhizosphere

The volume of soil around living plant roots that is influenced by root activity, either directly or via the activities of the microbiota that are stimulated by rhizodeposits (materials released by roots). Root–soil–microbiota interactions ultimately alter the biological, biochemical, chemical, and physical properties of the surrounding soil. The rhizosphere is thus a complex, dynamic, and heterogeneous microenvironment that plays a key role in plant nutrition and health, diversity of plant and soil microbial communities, soil functioning and genesis, and biogeochemical cycling of elements. *See* ROOT (BOTANY); SOIL.

Rhizodeposition and trophic relationships. The prime rhizosphere effect is a consequence of rhizodeposition. On average, 20% of the carbon assimilated by plants via photosynthesis is released by roots in the soil as a range of rhizodeposits: sloughed-off cells and decaying root hairs, respired CO_2, and organic compounds that are released via active secretion (for example, root mucilage) or passive diffusion (for example, root exudates such as sugars, amino acids, or organic acids). Rhizodeposition is a significant component of carbon cycling in ecosystems. Although sometimes viewed as a net loss of carbon for the plant, it is a major source of carbon for soil microbiota, making the rhizosphere a hot spot for trophic relationships. Rhizodeposits consist largely of rather small molecules that can be easily metabolized by microorganisms, either beneficial or deleterious (in the case of pathogenic microorganisms) for plants. Consequently, considerable stimulation of microorganisms is known to occur around roots. This rhizosphere effect is evidenced by an increased microbial biomass in the rhizosphere relative to bulk soil, particularly for those microorganisms that are cultivable in the laboratory, which are, however, known to represent only a minor proportion of the total number of microbial communities. Rhizosphere communities are especially enriched in rather fast-growing, opportunistic microorganisms (quite many of which are easily cultivable).

Recent advances in molecular microbial ecology have shown that the composition of microbial communities is considerably altered in the rhizosphere because the stimulation of microbial growth by rhizodeposits subsequently stimulates the activity of microbial grazers such as protozoa and nematodes in the rhizosphere. Thus, rhizodeposition enhances the turnover of rhizosphere microorganisms and soil microbiota, more so than the total microbial biomass. Combined with other changes in the structure of mi-

crobial communities, this will stimulate or depress a whole range of microbial activities in the rhizosphere, which will ultimately be either beneficial or detrimental to plants.

Such microbial activities and rhizodeposits are also involved in biochemical processes that have received less attention than those involved in carbon cycling, in spite of their importance: root-borne or microbial extracellular enzymes such as phosphatases, proteases, and arylsulfatases exhibit greater activity in the rhizosphere relative to bulk soil, and may have a dramatic effect on the cycle of nutrients such as phosphorus, nitrogen, and sulfur. In addition, the release of rhizodeposits exhibiting very different carbon/nitrogen ratios than bulk soil organic matter has been sometimes shown to induce a priming effect and enhance the mineralization of soil organic matter. Rhizodeposition thus significantly impacts the fate and cycle of major nutrients and trophic relationships in the soil surrounding living plant roots. *See* BIOGEOCHEMISTRY.

Communication and biological interactions. Alteration of the structure of rhizosphere microbial communities can also result from complex communication pathways between plant roots and rhizosphere microorganisms or between these microorganisms. Minute amounts of some signaling molecules can have a dramatic effect on microbial ecology. Such processes have been much studied with the advent of new tools and approaches in molecular ecology. Communication pathways and plant-microorganism interactions have been especially studied for associations such as legume–*Rhizobia* or mycorrhizal symbioses: such interactions result in distinct morphological features at the host root–microsymbiont interface, such as the nitrogen-fixing nodules and mycorrhizae that are involved in the exchanges of matter and energy between the microsymbiont (supplying inorganic nutrients) and the host plant (supplying carbon compounds). Prior to the formation of these specific, symbiotic organs, complex interactions occur which imply a communication pathway to establish adequate recognition of the two partners. The identification of flavonoids as signaling molecules has been well documented in legume–*Rhizobia* interactions, which explains the high specificity of this symbiosis. Genes involved in the regulation of symbiosis formation on both plant and microbial sides have now been identified and, interestingly, some of these are implicated in both rhizobial and mycorrhizal symbioses.

Beside the much-studied symbioses, interactions between roots and a whole range of plant growth–promoting rhizobacteria (PGPR) have received considerable interest. The production of phytohormones by PGPR has been well documented, for instance. Antibiotics have been shown to play a key role in the interactions between rhizosphere microorganisms, with potential applications in the field of biocontrol of root-borne diseases. Quorum sensing as related to molecules such as *N*-acetylated homoserine lactones is another type of interaction that has strong implications in the regulation of bacterial activities at the population level. Helper bacteria

implied in ectomycorrhizal symbiosis are an example of interactions which involve more than two partners. Many of these intriguing processes are somehow connected to the various strategies developed by roots and microbes for sensing their biotic and abiotic environment. Deciphering the molecular basis of the crosstalk between roots and microorganisms, as well as between various rhizosphere microorganisms, is a major challenge for managing rhizosphere ecology. *See* MYCORRHIZAE; SOIL MICROBIOLOGY.

Physical interactions. Although it is much less documented, root and microbial activities are also responsible for physical processes in the rhizosphere. These affect the movement of water and solutes in and out of the root, and thus the habitat of rhizosphere microorganisms. Among root activities that directly alter physical properties of the rhizosphere are primary functions such as root growth and uptake of water. These can result in changes in bulk density and porosity, soil strength, and water potential that can have a dramatic effect on the structure of microbial communities as well as on the uptake of water and nutrients by the plant. The production of mucilage by roots as well as microbial exopolysaccharides has also been shown to alter viscosity and surface tension properties of the soil, and thereby soil–water relationships in the rhizosphere; as well as the aggregation of soil around roots and the formation of rhizosheaths, which are unique structural features of the rhizosphere. Root-borne and microbial exopolysaccharides are also implicated in chemical processes such as the protection of root apex from aluminum toxicity and the acquisition of nutrients such as phosphorus. *See* PLANT MINERAL NUTRITION.

Chemical interactions. Chemical processes occurring in the rhizosphere as a direct consequence of a range of root functions such as absorption, respiration, and exudation are well documented. Combined with similar functions of rhizosphere microorganisms, these can result in steep gradients of concentrations of nutrients, potentially toxic elements, complexing or chelating compounds, pH and redox potential, and partial pressures of gases such as O_2 and CO_2. These chemical processes have a dramatic influence on the composition of the soil solution, and on the whole environment of the root, including the soil solid phase which is also the habitat of rhizosphere microorganisms. Some of these reactions play a role in the weathering of soil minerals and soil formation processes, e.g., in the dissolution of minerals such as calcium carbonate or phosphates as a result of rhizosphere acidification. These reactions also play a major role in the fate of nutrients. In the case of phosphorus, protons and exudates such as carboxylic anions (citrate, malate, or oxalate, especially) produced by roots and rhizosphere microorganisms have been shown to play a key role in some plant species such as legumes and in phosphorus-solubilizing bacteria or ectomycorrhizal fungi.

Other well-documented chemical interactions are those involved in the fate of iron. Plants have evolved two distinct strategies of iron acquisition, based either on localized rhizosphere acidification combined with enzymatic iron-reductase activity, or on the secretion of efficient iron chelators called phytosiderophores in grasses. Microbes also produce iron chelators, which represent a broader range of molecules called siderophores. These processes not only play a key role in the iron nutrition of the plant and microbiota: Given the poor availability of iron in many soils, the rhizosphere is the site for a strong competition, favoring the most iron-efficient organisms—via the production of siderophores, fluorescent *Pseudomonas* bacteria play an antagonistic role in controlling the population of the less iron-efficient *Fusarium* fungi. As the fungi is pathogenic, such chemical interactions occurring in the rhizosphere affect not only iron nutrition but also plant health.

Changes of pH, redox potential, and the production of complexing or chelating compounds in the rhizosphere also have a dramatic effect on the speciation and bioavailability of potentially hazardous trace elements such as arsenic, cadmium, and copper. *See* IRON; IRON METABOLISM; SIDEROPHORES; SOIL CHEMISTRY.

Knowledge gaps and prospects for management. The outer boundary of the rhizosphere is loosely defined, the spatial extension of the rhizosphere gradients extending for 1 mm or less when considering the depletion of phosphorus, for instance, while extending beyond several centimeters from the root surface when considering the outward diffusion of respired CO_2. Depending on which measure of root activity is considered, the rhizosphere volume can thus represent from much less than 1% up to 100% of the soil volume colonized by roots. In the case of phosphorus, mycorrhizal hyphae in symbiotic plants can considerably expand the depletion zone, sometimes described as the mycorrhizosphere. Thus, one cannot draw a single borderline between the rhizosphere and the bulk soil, and most rhizosphere processes do not show a sharp boundary but a more or less steep gradient. The rhizosphere is not only a site of radial heterogeneity around roots but also shows considerable spatial variation along the root axes and substantial temporal variations. These arise from the dynamic nature of root growth and the short- (diurnal rhythm) and long-term (seasonal rhythm) patterns of plant activities. These spatial and temporal heterogeneities combined with the multiplicity of biological, biochemical, chemical, and physical interactions make the rhizosphere a unique and complex microenvironment. Considerable variation also occurs between plant species or even between genotypes of a given species. Inoculation of PGPR, nitrogen-fixing bacteria, either symbiotic (*Rhizobia*) or nonsymbiotic (*Azospirilum*), or ectomycorrhizal fungi has already proved efficient for improving plant growth and nutrition. Applications of our knowledge of microbial interactions in the rhizosphere have also been successful in the field of biological control. In the future, better use of adequately selected or constructed plant genotypes may prove another efficient tool for managing those intimate rhizosphere interactions that ultimately determine

plant health, nutrition, and growth. *See* MICROBIAL ECOLOGY.   Philippe Hinsinger

Bibliography. E. A. Curl and B. Truelove, *The Rhizosphere,* Springer-Verlag, Berlin, 1986; P. Hinsinger, How do plant roots acquire mineral nutrients? Chemical processes involved in the rhizosphere, *Adv. Agron.*, 64:225–265, 1998; D. L. Jones, A. Hodge, and Y. Kuzyakov, Plant and mycorrhizal regulation of rhizodeposition, *New Phytologist*, 163:459–480, 2004; J. M. Lynch, *The Rhizosphere*, Wiley, Chichester, 1990; B. T. Tinker and P. H. Nye, *Solute Movement in the Rhizosphere*, Oxford University Press, 2000.

Rhizostomeae

An order of the class Scyphozoa with the most highly organized features of this class. The umbrella is generally higher than it is wide, except in *Cassiopea* and *Cephea*. The margin of the umbrella is divided into many lappets but is not provided with tentacles. Many radial canals, which are connected with each other to form a complicated network, issue from the cruciform stomach. The oral part is very complicated and is eight-sided, with many suctorial mouths surrounded by numerous small tentacles (see **illus.**). Usually there is no central mouth.
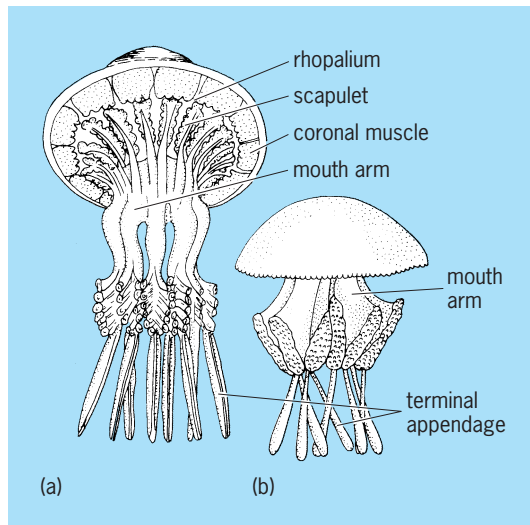
Rhizostomeae. (*a*) *Rhizostoma*; the bell may reach 2 ft (0.6 m) in diameter. (*b*) *Mastigias*. (*After L. Hyman, The Invertebrates, vol. 1, McGraw-Hill, 1940*)

The development of eggs into medusae, through the stages of planula, scyphopolyp, strobila, and ephyra, is similar to that of the Semaeostomeae. However, the metamorphosis of the ephyrae is more complicated than in the Semaeostomeae. No species of this group is injurious to the human skin. Some large forms, such as *Rhopilema*, are used as food in China and Japan.

A fair number of fossils of this order were found in the strata of the Jurassic Period. *See* SCYPHOZOA.
                                                        Tohru Uchida

Rhodium

A chemical element, Rh, atomic number 45, relative atomic weight 102.905. Rhodium is a transition metal and one of the group of platinum metals (ruthenium, osmium, rhodium, iridium, palladium, and platinum) that share similar chemical and physical properties. *See* METAL; PERIODIC TABLE; PLATINUM.

The terrestrial abundance of rhodium is exceedingly low; it is estimated to be 0.0004 part per million in the Earth's crust. It is found as a single isotope, ^{103}Rh, with a nuclear spin of 12. Since the platinum metals share common reactivities and are mined from a common source, there is an involved chemical process that is used to separate the individual elements, including rhodium. *See* ISOTOPE.

Metallic rhodium is the whitest of the platinum metals and does not tarnish under atmospheric conditions. Its surface is normally covered by a thin, firmly bound layer of rhodium(IV) oxide (RhO_2). Rhodium is insoluble in all acids, including aqua regia. It dissolves in molten potassium bisulfate ($KHSO_4$), a useful property for its extraction from platinum ores, since iridium, ruthenium, and osmium are insoluble in this melt. Important physical properties of metallic rhodium are given in the **table**. *See* ACID AND BASE; AQUA REGIA; HALOGEN ELEMENTS.

Metallic rhodium is available as powder, sponge, wire, and sheets. It is ductile when hot and retains its ductility when cold. However, it work-hardens rapidly. Molten rhodium dissolves oxygen. Upon cooling, the oxygen gas is liberated, and this can lead to ruptures in the external surface of the crust of the metal. As a result, molten rhodium is best handled under an inert atmosphere of argon, which does not dissolve in rhodium.

Complexes of Rh(III), including $RhCl_3(pyridine)_3$, $Rh(CO)Cl_3[P(C_6H_5)_3]_2$, and $RhCl_6^{3-}$, are diamagnetic six-coordinate with octahedral geometry. The most common chemical form of rhodium is $RhCl_3 \cdot 3H_2O$, a red-brown, deliquescent material that is a useful starting material for the preparation of other rhodium compounds. In contrast to the hydrated material, red anhydrous rhodium(III) chloride ($RhCl_3$) is a polymeric, paramagnetic compound that does not dissolve in water. *See* DIAMAGNETISM.

Physical properties of rhodium metal	
Property	Value
Crystal structure	Face-centered cubic
Lattice constant a, at 25°C (77°F), nm	0.38031
Thermal neutron capture cross section, barns (10^{-28} m^2)	149
Density at 25°C (77°F), g/cm^3	12.43
Melting point	1963°C (3565°F)
Boiling point	3700°C (6700°F)
Specific heat at 0°C, cal/g (J/kg)	0.0589 (246)
Thermal conductivity, 0–100°C, cal cm/cm^{2}s, °C (J · m/m^2 · s · °C)	0.36 (151)
Linear coefficient of thermal expansion, 20–100°C, μin./(in./°C) or m/(m · °C)	8.3
Electrical resistivity at 0°C, microhm-cm	4.33
Temperature coefficient of electrical resistance, 0–100°C/°C	0.00463
Tensile strength, 10^3 lb/in.^{2} (6.895 MPa)	
Soft	120–130
Hard	200–230
Young's modulus at 20°C (68°F), lb/in.^{2} (GPa)	
Static	46.2×10^6 (319)
Dynamic	54.8×10^6 (378)
Hardness, diamond pyramid number	
Soft	120–140
Hard	300
ΔH_{fusion}, kJ/mol	21.6
$\Delta H_{vaporization}$, kJ/mol	494
ΔH_f monoatomic gas, kJ/mol	556
Electronegativity	2.2

The low natural abundance and high cost of rhodium limit its uses to specialty applications. The major use is in catalysis, which accounts for over 60% of its production. Rhodium is a component of catalytic converters used in the control of exhaust emissions from automobiles. *See* CATALYTIC CONVERTER.

Rhodium is also used in the hydrogenation of olefins to alkanes. For hydrogenation, both heterogeneous catalysis and homogeneous catalysis are used. Heterogeneous conditions are achieved with rhodium metal finely dispersed on an inert support (activated carbon, charcoal, or alumina).

Rhodium complexes have been developed as catalysts for the synthesis of one optical isomer of L-dopa (used in treatment of Parkinson's disease). Greater selectivity makes rhodium catalysts more useful in hydroformylation or oxo reactions than the less expensive cobalt catalysts. A platinum-rhodium alloy is an efficient commercial catalyst for the formation of nitric acid through ammonia oxidation. *See* CATALYSIS; HETEROGENEOUS CATALYSIS; HOMOGENEOUS CATALYSIS; HYDROFORMYLATION; HYDROGENATION.

Rhodium-platinum alloys are favored for high-temperature applications. The International Temperature Scale over the range 630.5–1063°C (1134.9–1945.4°F) is defined by a thermocouple using a 10% rhodium-platinum alloy. Electroplated rhodium retains its bright surface under atmospheric conditions and finds use as electrical contacts and reflective surfaces. The reflectivity of rhodium surfaces is high (80%) and does not tarnish. About 6% of the rhodium production goes into jewelry manufacturing. *See* ELECTROPLATING OF METALS; TRANSITION ELEMENTS.
Alan A. Balch

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; R. S. Dickson, *Organometallic Chemistry of Rhodium and Iridium*, 1983; A. Earnshaw and N. Greenwood, *Chemistry of the Elements*, 2d ed., Butterworth-Heinemann, 1997; J. A. McCleverty and T. J. Meyer (eds.), *Comprehensive Coordination Chemistry II*, 2d ed., Elsevier Science, 2003.

Rho-theta system

A navigation system in which one or more signals are emitted from a facility (or colocated facilities) to produce simultaneous indication of bearing and distance. Since a bearing is a radial line of position and a distance is a circular line of position, the rho-theta system always ensures a position fix produced by the intersection of two lines of position which are at right angles to each other. This produces a minimum geometric dilution of position (GDOP), a figure of merit for all radio navigation systems, and is one of the chief advantages of a rho-theta system. Another major advantage is that it is a single-site system and can thus be installed on a ship or an island. This has also made it attractive politically, enabling small countries to have their own navigation systems. *See* ELECTRONIC NAVIGATION SYSTEMS; VOR (VHF OMNIDIRECTIONAL RANGE).

VOR. In May 1949 the Council of the International Civil Aviation Organization (ICAO), an agency of the United Nations, adopted the vhf (very high-frequency) omnidirectional radio range, commonly referred to as the VOR, to become standard on March 1, 1950. The ICAO also recommended that research and development work be actively pursued to produce a complete specification for distance-measuring equipment (DME) operating in the uhf (ultrahigh-frequency) band. The VOR was to operate in the band of 112 to 118 MHz. Equipment designed to the specification listed in the ICAO document was installed on the ground in many places

throughout the world. The location of these installations was agreed to by contracting states at their various ICAO regional meetings. This equipment operating with companion equipment on board aircraft produced static-free directional guidance to distances of over 150 mi (240 km). More than 900 VOR ground installations are in use in the United States and an equal number in the rest of the world.

It is difficult to specify the accuracy of the VOR system; accuracy of the airborne equipment varies considerably, from the highly engineered units which are utilized by the air carriers to equipment which the owner of a light pleasure craft purchases for one-tenth of the price. VOR accuracy also varies with terrain. In the early 1960s a wide baseline version of the VOR was developed and installed in the United States and Europe wherever a difficult site was encountered. This development was termed the Doppler VOR and has improved accuracy in some cases by as much as 10 times. *See* DOPPLER VOR.

According to the *2001 Federal Radionavigation Plan*, VOR/DME will continue tp provide navigation services for en route through nonprecision approach phases of flight throughout the transition to satellite-based navigation. The Federal Aviation Administration (FAA) plans to reduce VOR/DME services provided in the national airspace based on the anticipated decrease in its use for en route navigation and instrument approaches. The Phase-down is scheduled to begin in 2010. *See* SATELLITE NAVIGATION SYSTEMS.

Tacan. By the mid-1950s the military services of the United States had completed the development of a novel rho-theta system called Tacan for tactical air navigation. The system consists of ground equipment which produces a rotating radio field pattern in space very similar to that produced by the VOR; however, instead of the simple cardioid produced by the VOR, the Tacan antenna produces a cardioid plus a strong ninth harmonic which enhances its bearing accuracy. Unlike the VOR, the emission is not a continuous wave but a constant series of pulses. These pulses are emitted in time in accordance with the signals received from an airborne interrogator, so that by measuring round-trip time, distance to the ground beacon is displayed on a meter in the cockpit. The miltary need for and use of land-based TACAN will continue until aircraft are properly integrated with the Global Positioning System (GPS) and GPS is approved for all operations in national and international controlled airspace. The phase-down of TACAN usage is scheduled to begin in 2010 unless TACAN is determined to be necessary for long-term navigation services. *See* TACAN.

Vortac. The Tacan system operates in the band of 962–1213 MHz, which was the band that ICAO had recommended to be occupied by the civil DME. With the completion of the system, a facility combining the VOR and the Tacan was developed. Installations called vortacs were developed and installed where both the military and civil air operations had requirements for service. A vortac installation consists of the VOR beacon which furnishes directional guidance to civil aviation aircraft or to military aircraft that happen to be equipped with the VOR receiver.

The civil aviation aircraft carry a DME interrogator which interrogates the Tacan beacon that is colocated at the VOR site. The civil DME airborne interrogator measures round-trip elapsed time of the transmission and displays it as distance. A military aircraft carrying an airborne interrogator-responsor interrogates the ground colocated beacon and in reply receives both bearing and distance information from it. The phase-down of VORTAC sites is planned to begin in 2010.

Both the civil airborne VOR with its DME and the military Tacan can generate arbitrary courses to any point within the coverage of the ground vortac beacons by the use of simple rho-theta computers.

Other systems. The instrument landing system (ILS) and the microwave landing system (MLS) are, in essence, also rho-theta systems when they are used with DME. However, they do not normally provide full 360° coverage. The FAA has terminated the development of the MLS based on favorable GPS tests. The United States does not anticipate installing MLS equipment in the national airspace. *See* INSTRUMENT LANDING SYSTEM (ILS); MICROWAVE LANDING SYSTEM (MLS).

The advantage of the rho-theta concept has led to many other proposals, but the only one to have been implemented on a large scale is the Soviet Union's RSBN, developed before the Soviet Union joined the ICAO in 1970. It comprises a secondary radar, rotating at 100 revolutions per minute, which receives replies from airborne transponders. The aircraft derives bearing by comparing the arrival time of the radar beam with an omnidirectional pulse transmission from the radar each time the radar beam sweeps through the direction due north. The two-way link between radar and aircraft is also used by a conventional air-derived DME. RSBN, while standard throughout the Soviet Union, is not used elsewhere.   Sven H. Dodington; Richard L. Greenspan

Bibliography. B. Forssell, *Radionavigation Systems*, Prentice Hall, 1991; M. Kayton and W. Fried, *Avionics Navigation Systems*, 2d ed., Wiley, 1997; G. Pakholkov and G. Grimov, Short-range radio navigation systems, *IEEE/AESS Magazine*, January 1988; U.S. Department of Defense and U.S. Department of Transportation, *2001 Federal Radionavigation Plan*, 2002.

Rhodochrosite

The mineral form of manganese carbonate. Calcium, iron, magnesium, and zinc have all been reported to replace some of the manganese. The equilibrium replacement of manganese by calcium has been determined and found to increase with the temperature of crystallization.

Rhodochrosite is sometimes found in low-temperature veins near deposits of copper, lead, zinc, and silver, or it may occur with other manganiferous minerals of higher temperature origin.

Rhodochrosite crystals, Alicante, Colorado. (*Specimen from Department of Geology, Bryn Mawr College*)

It has also been found in sediments and in pegmatites. Well-known occurrences of rhodochrosite are in Europe, Asia, and South America. In the United States large quantities occur at Butte, Montana. As a source of manganese, rhodochrosite is also important at Chamberlain, South Dakota, and in Aroostook County, Maine.

Rhodochrosite has hexagonal (rhombohedral) symmetry and the calcite-type crystal structure (see **illus.**). It occurs more often in massive or columnar form than in distinct crystals. The color ranges from pale pink to brownish pink. Hardness is 3.5–4 on Mohs scale, and specific gravity is 3.70. *See* CARBONATE MINERALS; MANGANESE.   Robert I. Harker

Rhodonite

A mineral inosilicate with composition $MnSiO_3$. Rhodonite crystallizes in the triclinic system in crystals that are commonly tabular parallel to the base (see **illus.**). More often it is in cleavable to compact masses or in embedded grains. Crystallographically, rhodonite is closely related to the pyroxenes and thus has two cleavage directions at about 88 and 92°. Hardness is 5.5–6 on Mohs scale, and specific gravity is 3.4–3.7. The luster is vitreous and the color is rose red, pink, or brown. Rhodonite is similar in color to rhodochrosite, manganese carbonate, but it may be distinguished by its greater hardness and insol-
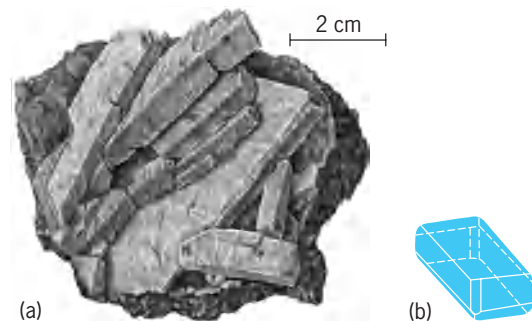
Rhodonite. (*a*) Crystals of variety fowlerite with limestone, Franklin, New Jersey (*specimen from Department of Geology, Bryn Mawr College*). (*b*) Crystal habit (*after C. Klein and C. S. Hurlbut, Jr., Dana's Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993*).

ubility in hydrochloric acid. It has been found at Langban, Sweden; near Sverdlovsk in the Ural Mountains; and at Broken Hill, Australia. Fine crystals of a zinc-bearing variety, fowlerite, are found at Franklin, New Jersey. *See* SILICATE MINERALS.
                    Cornelius S. Hurlbut, Jr.

Rhodophyceae

A large class of plants, commonly called red algae, coextensive with the division Rhodophycota. Most red algae are found in the ocean, growing on rocks, wood, other plants, or animals in the intertidal zone and to depths limited by the availability of light. A few genera and species occur in fresh water, and these are usually found in rapidly flowing, well-aerated, cold streams. Some, however, grow in quiet warm water, while a few are subaerial. Most red algae are photosynthetic, but some grow on other algae with varying degrees of parasitism. Approximately 675 genera and 4100 species are recognized. *See* ALGAE.

Characteristics. Rhodophyceae are characterized by a unique combination of biochemical, reproductive, and ultrastructural features. The primary photosynthetic pigment is chlorophyll *a*. In addition, small amounts of chlorophyll *d* are found in some species. Water-soluble tetrapyrrolic compounds called phycobilins serve as accessory photosynthetic pigments. The phycobilins, which include phycoerythrin (red) and phycocyanin and allophycocyanin (blue), are bound to proteins. The resulting pigment-protein complexes (phycobiliproteins) aggregate to form phycobilisomes, which are arranged in an orderly pattern on the outer surfaces of the chloroplast lamellae. The carotenoid pigments are similar to those of green plants. Varying ratios of chlorophylls, carotenoids, and phycobilins provide a spectrum of color from blue-green through olive-green and reddish brown to bright pink. The production of phycoerythrin is adversely affected by nitrogen deficiency and intense illumination, so that the brightest pink or red colors are found in plants growing subtidally in nutrient-rich water. The absorption by the phycobilins of energy in the green portion of the spectrum compensates for the loss of red rays, which are absorbed in the upper few meters of water. The chloroplasts, which may be stellate, discoid, or linear, are bounded by a double membrane. Each photosynthetic lamella comprises a single thylakoid. A pyrenoid is present in some genera.

The chief food reserve, floridean starch, is a branched polymer of glucose (α-1,4-glucan) similar to amylopectin of green plants. It occurs as granules in the cytoplasm. Multinucleate cells are common. In some genera, specific areas of nominally haploid or diploid plants are polyploid or polytenic, the orderly progression from one condition to another being related to ontogeny. The mitochondria have flattened cristae. The cell walls are composed of an inner fibrillar layer of cellulose and an outer mucilaginous layer of sulfated polymers of galactose that have economic importance. Throughout the order Corallinales, the cell walls are
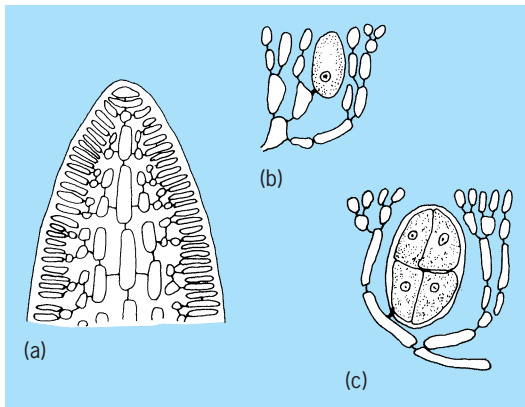
Fig. 1. *Gelidium*. (*a*) **Longitudinal section of apex of branch, showing uniaxial filamentous construction.**
(*b*) **Tetrasporangium mother cell, before meiosis.**
(*c*) **Tetrasporangium, following meiosis, cleaved into tetraspores. (*After G. M. Smith, Cryptogamic Botany, 2d ed., McGraw-Hill, 1955*)**

impregnated with the calcite form of calcium carbonate. In two common tropical noncoralline red algae (*Galaxaura* and *Liagora*), the cell walls are impregnated with the aragonite form of calcium carbonate. Rhodophyceae are distinctive among eukaryotic algae in their lack of flagella, a feature shared among major groups only by the chlorophycean order Zygnematales. Some unicellular forms (in the order Porphyridiales) and many spores and male gametes of multicellular forms are capable of gliding or feeble ameboid motion. *See* CORALLINALES.

Structure. Unicellular red algae, which may form mucilaginous colonies, are considered primitive. Most red algae have multicellular thalli of microscopic or macroscopic size, including individual filaments, blades, and complex plants of distinctive form produced by the interplay of filamentous systems (**Fig. 1***a*). The largest red algae rarely exceed a meter (3.3 ft) in length. In all but a few families, filaments grow by elongation of existing cells and cutting off of new cells by apical cells. Typically, intercalary cells initiate branches, but never divide transversely. In all members of the subclass Florideophycidae and in a few members of the subclass Bangiophycidae, the derivatives of a dividing cell remain linked. This link, the primary pit connection, results from incomplete cell plate formation. A hole left in the center subsequently becomes plugged with a proteinaceous substance. Contiguous cells of different lineages (or rarely the same lineage) may be linked secondarily by pit connections formed when a very small cell cut off from the initiating cell fuses with an adjoining cell. Secondary pit connections strengthen the thallus. In some members of the Corallinales, they are formed directly, without an intermediate conjunctive cell. They may also be formed between a host and its red algal parasite. During their formation, a nucleus is transferred from the parasite to the host. Fusions between and among cells is frequently encountered during development of the female reproductive system for the apparent purpose of increasing the supply of nutrients to the maturing cystocarp. Fusions are also found in vegetative tissues of many Corallinales, where their function is unknown.

Life histories and reproduction. Most red algae have a life history involving a succession of somatic and cytological phases. A free-living diploid phase (tetrasporophyte) bears sporangia that undergo meiosis to form tetrads of haploid spores (tetraspores; Fig. 1*b*, *c*). The tetraspores germinate, and the germlings develop into haploid thalli (gametophytes). Male and female gametes are usually produced on different plants. The tetrasporophyte and the gametophytes may be similar in size and form (isomorphic) or dissimilar to a marked degree (heteromorphic). The three free-living phases often exhibit vegetative differences that are so slight as to be appreciated only by specialists; such life histories are considered to be isomorphic. In the most common heteromorphic life history, the tetrasporophyte is a crust or a minute filamentous structure while the gametophytes are macroscopic blades. In one order (Palmariales) the male gametophyte resembles the tetrasporophyte, but the female gametophyte is greatly reduced, in at least one species to a single cell. A dwarf male gametophyte is produced by one anomalous species in a family characterized by isomorphy.

The contents of a male sex organ (spermatangium) function as a male gamete (spermatium; **Fig. 2***a*). Spermatia, which measure 2–6 micrometers in diameter, are spherical or ovoid, colorless, and produced in great abundance. Although they have no flagella, some spermatia have mucilaginous appendages that may assist in dispersal or attachment to the female organ (carpogonium), whose contents function as the female gamete. The carpogonium is usually a flask-shaped cell with an elongate receptive extension (trichogyne; Fig. 2*b*). The nucleus of an attached spermatium migrates down the trichogyne to the base of the carpogonium, where it fuses with the female nucleus to produce a zygote. In some red algae the zygote develops directly into a compact, diploid somatic phase, the carposporophyte (or gonimoblast), which remains an integral part of the female plant (Fig. 2*c*). A protective envelope (pericarp) may be produced by the female gametophyte; the carposporophyte with its pericarp is termed
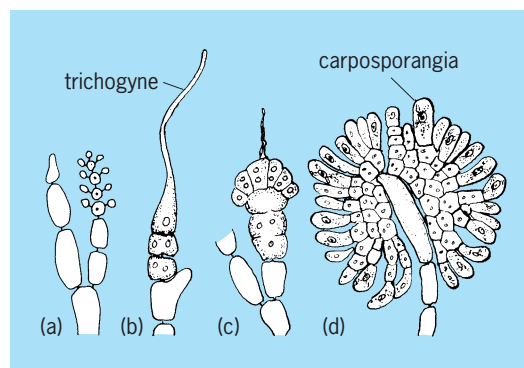

Fig. 2. *Nemalion*. (*a*) **Terminal group of spermatangia.**
(*b*) **Three-celled carpogonial branch terminating in carpogonium with trichogyne. (*c*) Early stage in direct development of carposporophyte from zygote.**
(*d*) **Longitudinal section through mature carposporophyte, showing terminal carposporangia. (*After G. M. Smith, Cryptogamic Botany, 2d ed., McGraw-Hill, 1955*)**

the cystocarp. In most red algae, by contrast, the zygote nucleus is transferred to another, often cytologically differentiated cell (auxiliary cell) which then undergoes a series of divisions to produce the carposporophyte. The carposporophyte in all red algae is partly or entirely converted into carposporangia (Fig. 2d). Carpospores are released singly from carposporangia and germinate to give rise to tetrasporophytes. In a few species, meiosis is involved in the production of carpospores, which therefore are haploid and germinate to give rise to gametophytes, thereby eliminating the tetrasporophyte phase. Some red algae bypass sexual reproduction by producing apomeiotic (unreduced) tetraspores, mitotically derived monospores, or vegetative propagules.

Classification. Two subclasses are traditionally recognized. In the Bangiophycidae the thallus is relatively simple: unicells, mucilaginous colonies, branched and unbranched discrete filaments, and thin blades. Only asexual reproduction is known for most members of this subclass, although the widespread and well-known genus *Porphyra* has a heteromorphic sexual life history. Growth is diffuse in the Bangiophycidae, and pit connections are formed in only a few genera. The Florideophycidae comprise multicellular forms, most of which are known or suspected to have sexual life histories. Growth is initiated by apical cells and all members form pit connections. Classification within the Bangiophycidae is based largely on vegetative and asexual reproductive features, while that of the Florideophycidae is based primarily on details of the

development of the female reproductive system and carposporophyte, secondarily on vegetative features.

Phylogeny. Rhodophyceae seem most closely related to Cyanophyceae in their use of chlorophyll *a* and phycobilins as photosynthetic pigments and the absence of flagella. They probably did not evolve directly from Cyanophyceae, however, but from a colorless, nonflagellate, eukaryotic ancestor that acquired pigments from an endosymbiotic blue-green alga. Fronds, presumably of red algae, are found in Cambrian shales. A more certain fossil record is provided by the reef-forming calcareous Corallinales, which extend back to the Devonian (about 400 million years ago).

Economic importance. The most salient feature of red algae is their beauty (**Fig. 3a**), which has drawn admiration from generations of seaside visitors. A popular and sometimes profitable Victorian pastime, in both Britain and the United States, was to arrange red algae artistically on cards, which were sold as souvenirs. Their greatest significance, however, is their role in the formation of coral reefs, the Corallinales being responsible for cementing together various animal and algal components. Of more apparent economic importance is their use as food, a centuries-old tradition of maritime peoples in many parts of the world. In the North Atlantic, *Palmaria palmata* (Fig. 3b, formerly, but incorrectly, placed in the genus *Rhodymenia*) was relished under the name dulse (English) or dillesk (Irish), while *Chondrus crispus* (Irish moss; Fig. 3d) was used in making a pudding by the extraction of a jelling agent. In the Far East and Indonesia, jelling agents were extracted from *Gelidium* (Fig. 1d), *Gracilaria*, and *Eucheuma*. The reddish-purple blades of *Porphyra* (asakusi-nori or simply hoshi) have long been considered a delicacy by the Japanese. *See* REEF.

Many of these folk customs have been capitalized into substantial commercial enterprises. Carrageenan, the jelling agent of Irish moss, has been found to have numerous industrial applications and is extracted in factories from *Chondrus*, the closely related genus *Gigartina*, and the less closely related genus *Eucheuma*. Natural sources of *Eucheuma* are greatly supplemented by crops grown in shallow water, especially in the Philippines, where the economy of coastal villages has been significantly strengthened. Agar, the jelling agent of *Gelidium* and *Gracilaria*, is best known for its use in preparing microbiological media. Its manufacture, although still dependent upon naturally occurring plants, has spread from Japan to nations (such as Spain) without a traditional agar industry. *See* AGAR; CARRAGEENAN.

The great economic importance of producing nori in Japan is commensurate with the social importance of eating it. Specially prepared kinds command astoundingly high prices in the luxury food market, while the ordinary product is a staple in the Japanese household. *Porphyra* has been cultivated in Japan for nearly three centuries, originally on twigs secured in the mud of shallow bays, later on rope nets suspended at tide levels optimal for
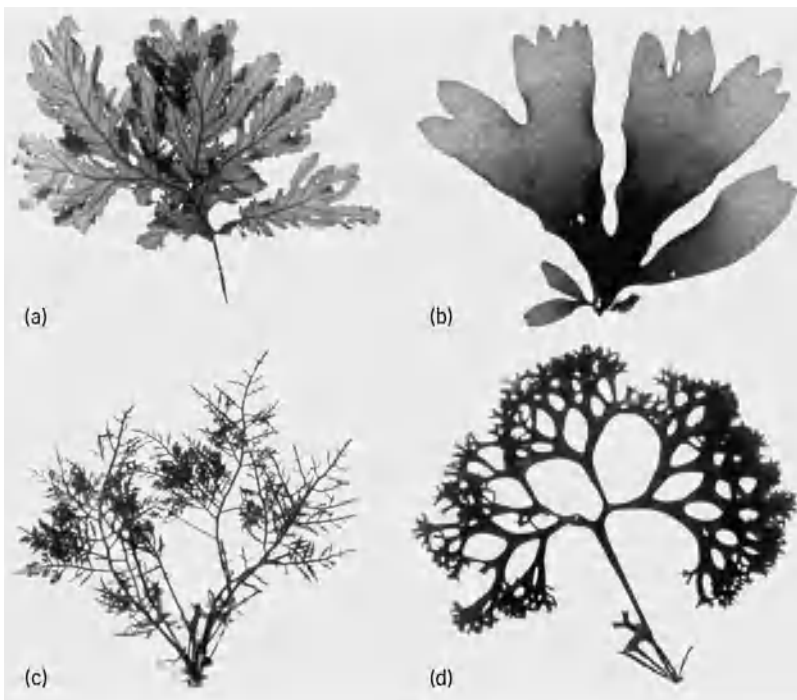

Fig. 3. Representative species of Rhodophyceae. (*a*) *Phycodrys fimbriata,* or sea oak, an alga of beauty but of no commercial importance. (*b*) *Palmaria palmata*, or dulse, eaten by maritime peoples in the North Atlantic. (*c*) *Gelidium robustum*, a commercial source of agar. (*d*) *Chondrus crispus*, or Irish moss, the chief commercial source of carrageenan.

growth. After the discovery in 1949 that the life history of *Porphyra* includes a microscopic spore-producing stage (previously described as an independent shell-boring alga, *Conchocelis*), the cultivation was modified to take advantage of natural cycles of sporulation. The "seeding" of nets, formerly done by randomly towing them in open-coast areas where *Porphyra* occurred naturally, is done in the factory by dipping them in a culture of the conchocelis stage.   Paul C. Silva; Richard L. Moe

Bibliography. H. C. Bold and M. J. Wynne, *Introduction to the Algae: Structure and Reproduction*, 2d ed., 1997; P. S. Dixon, *Biology of the Rhodophyta*, 1973; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; R. G. Sheath, The biology of freshwater red algae, *Prog. Phycol. Res.*, 3:89–157, 1984.

Rhombifera

A diverse class of extinct, marine, stalked echinoderms whose stratigraphic range extends from the Late Cambrian (505 MA) through Upper Devonian (360 MA). The class is named for rhombus-shaped structures (rhombs) that circulated seawater or body fluids through the body for the respiratory extraction of dissolved oxygen. The main body, or theca, held the visceral mass and was attached to the seafloor by a short to long stem. Atop the theca were variously developed small appendages that were used for suspension feeding in slowly moving currents. This class is composed of three distantly related groups: Glyptocystitida, Hemicosmititida, and Caryocystitida (see **illustration**). A fourth group, Polycosmititida, is too poorly known to assign to Rhombifera with certainty.

Glyptocystitids (Late Cambrian to Upper Devonian) comprise the bulk of rhombiferans. This group is characterized by a large theca, comprised of 25–27 skeletal elements, which bore numerous hairlike feeding appendages called brachioles. The short, stout stem is enlarged and flexible near the thecal attachment. The rhombs have elongate slits with little differentiation between incurrent and excurrent sides. Glyptocystitids include most of the variation with sessile and mobile species, and thecal shapes including barrel-shaped, spherical, and flattened triangular.

Hemicosmitids (Early Ordovician to Early Devonian) are characterized by a large theca with about 36 skeletal elements bearing stout erect arms upon which hairlike brachioles were born. The stem is long and stout and lacks the enlarged upper end of glyptocystitids. Rhombs have small sievelike clusters of incurrent pores and larger spoutlike excurrent pores.

Caryocystitids (Early to Late Ordovician) are characterized by an irregularly elongate to globular theca formed from numerous skeletal elements, bearing a few stout feeding brachioles. The stem is short, gracile, and irregularly plated. Rhombs of caryocystitids were fully internal, circulating body fluids instead of seawater. *See* ECHINODERMATA.
   Colin D. Sumrall

Bibliography. R. V. Kesling, Cystoids, pp. S85–S267, in R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, Part S: *Echinodermata* 1(1), Geological Society of America and University of Kansas, New York and Lawrence, 1968; J. Sprinkle, *Morphology and Evolution of Blastozoan Echinoderms*, Spec. Publ., Museum of Paleontology, Harvard University, Cambridge, 1973; J. Sprinkle, Cylindrical and globular rhombiferans, pp. 231–273, in J. Sprinkle (ed.), *Echinoderm Faunas from the Bromide Formation (Middle Ordovician) of Oklahoma*, University of Kansas, Lawrence, 1982.
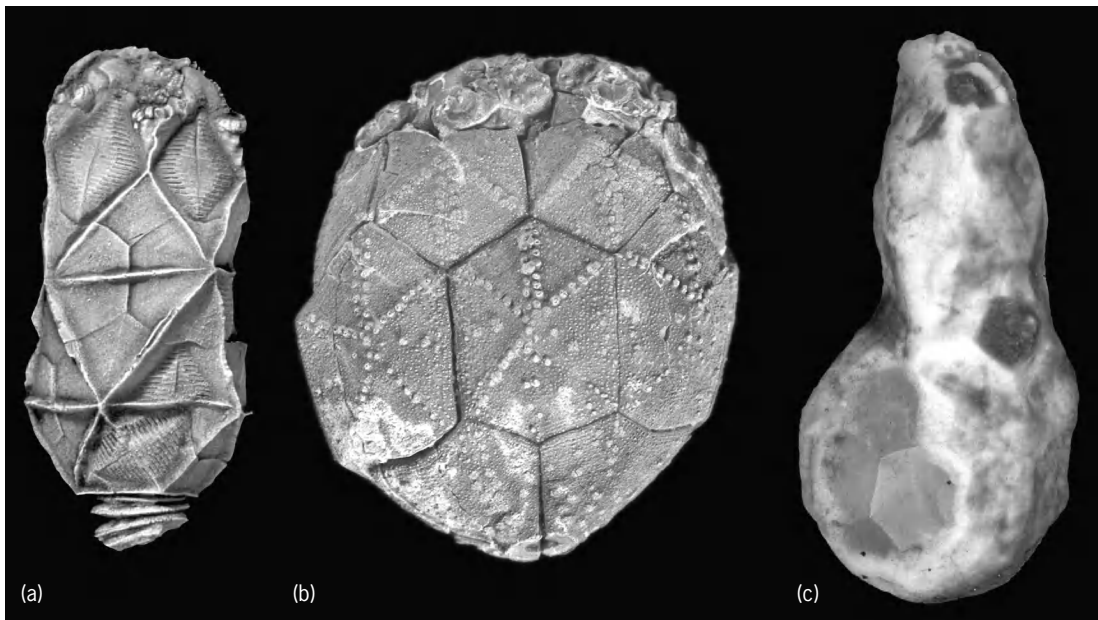
Rhombiferans of three groups. (*a*) The glyptocystitid *Cheirocystis fulotonensis*. (*b*) The hemicosmitid *Caryocrinites ornatus*. (*c*) The caryocystitid *Caryocystites angelini*.

Rhubarb

A herbaceous perennial, *Rheum rhaponticum*, of Mediterranean origin, belonging to the plant order Polygonales. Rhubarb is grown for its thick petioles which are used mainly as a cooked dessert (see **illus.**); it is frequently called the pie-plant. The
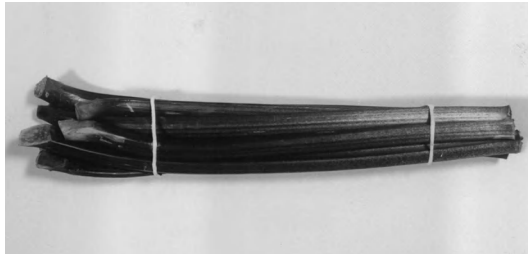
Cut stalks of rhubarb with leaves removed. (**W. Atlee Burpee Co.**)

leaves, which are high in oxalic acid content, are not commonly considered edible. Propagation is by division of root crowns. Victoria, Macdonald, and Valentine are popular varieties (cultivars). Commercial production is generally limited to areas where crowns may become dormant for 2–3 months each year. Outdoor rhubarb is a common garden vegetable in most areas of the United States except the South. Harvesting begins in the spring and continues for 6–10 weeks. Commercial plantings are renewed every 4–8 years. Michigan and Washington are important centers for forced or hothouse rhubarb. Two- or three-year-old field-grown crowns are moved into darkened forcing structures in late winter and forced at 55–60°F (12–15°C) to obtain petioles of a bright-red color. *See* POLYGONALES.   H. John Carew

Rhynchobdellae

An order of the class Hirudinea. These leeches possess an eversible proboscis and lack hemoglobin in the blood. They may be divided into two families, the Glossiphoniidae and the Ichthyobdellidea.

Glossiphoniidae are flattened, mostly small leeches occurring chiefly in fresh water. *Theromyzon* is a parasite of waterfowl, sucking blood from the nasal cavity, the mouth, or the eye. A heavy infestation on a young bird may be fatal. *Hemiclepsis* sucks the blood of fishes and amphibians and is found in stagnant waters in Europe and northern Asia. *Placobdella* is a common American parasite of aquatic turtles, frogs, and fishes, and *Glossiphonia* sucks the blood of aquatic invertebrates, especially snails (see **illus.**). The genus *Haementeria* contains reptile parasites chiefly, but also includes parasites of mammals, including humans. *Haementeria officinalis* is used in Mexico for medicinal purposes.

Ichthyobdellidae typically have cylindrical bodies with conspicuous, powerful suckers used to attach themselves to passing fish. They frequently have lateral appendages which aid in respiration. There is a tendency for a particular species to be confined to a specific host. For instance, *Callobdella lophii*
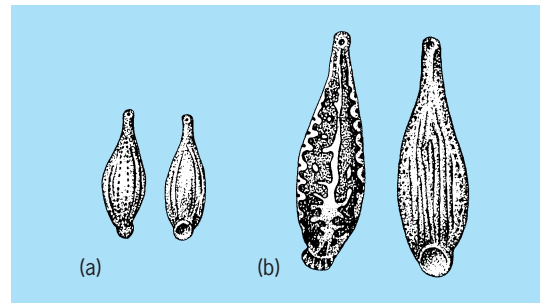
Examples of Rhynchobdellae. (*a*) *Glossiphonia complanata*. (*b*) *Placobdella parasitica*.

is confined to the angler fish *Lophius piscatorius*, and *Abranchus microstomus* to the small shore fish *Cottus scorpius*. On the other hand, *Pontobdella muricata* attacks a variety of skates and rays. *Crangonobdella* is an example of a parasite of crustaceans. *Piscicola* is one of the few fresh-water forms. *See* ARHYNCHOBDELLAE; HIRUDINEA.

Kenneth H. Mann

Rhynchocephalia

One of the two surviving clades of lepidosaurian reptiles, represented today by only two species, *Sphenodon punctatus* and *Sphenodon guntheri*, commonly known as tuatara. *Sphenodon* (a name meaning wedge-tooth) has a fully diapsid skull (**Fig. 1***a*) that was secondarily acquired from an ancestral state represented by *Gephyrosaurus*, the most primitive known member of the clade Rhynchocephalia, in which the lower temporal bar was incomplete, due to a loss of contact between jugal and quadratojugal (Fig. 1), as in lizards and basal lepidosauromorphs. The extinct *Gephryosaurus* shows some but not all of the advanced characters of more typical rhynchocephalians. Most notably, it has a primitive tooth implantation, that is, pleurodont (attached to the inside of the jaw). Typically, the teeth of rhynchocephalians are fused to the edges of the jaws (a condition known as acrodont). There are two upper rows of teeth, one of which is enlarged and palatine (on the palate), and one row of dentary (lower jaw) teeth. The dentary teeth bite between the two upper rows in a vertical manner (as in some extinct taxa, for example, *Clevosaurus*; Fig. 1*c*) or in a horizontal sliding movement known as propaliny (as in the living *Sphenodon*). *See* DIAPSIDA; LEPIDOSAURIA.

Living Sphenodon. The living New Zealand tuatara, or *Sphenodon*, is a medium-sized (2–2^1/$_2$ ft, or 60–80 mm) reptile, rather like a heavy-bodied iguana in appearance but very different in its lifestyle (**Fig. 2**). Tuatara is a Maori word meaning "peaks on the back," a reference to the jagged dorsal crest. The reptiles have an important place in Maori tradition, but, unfortunately, were also part of the cuisine. Once widespread on the main islands of New Zealand, they are now restricted to about 30 small offshore islets, threatened mainly by introduced animals such as pack-rats and cats.

Sphenodon is unique among lepidosaurs in having only very rudimentary hemipenes (the paired
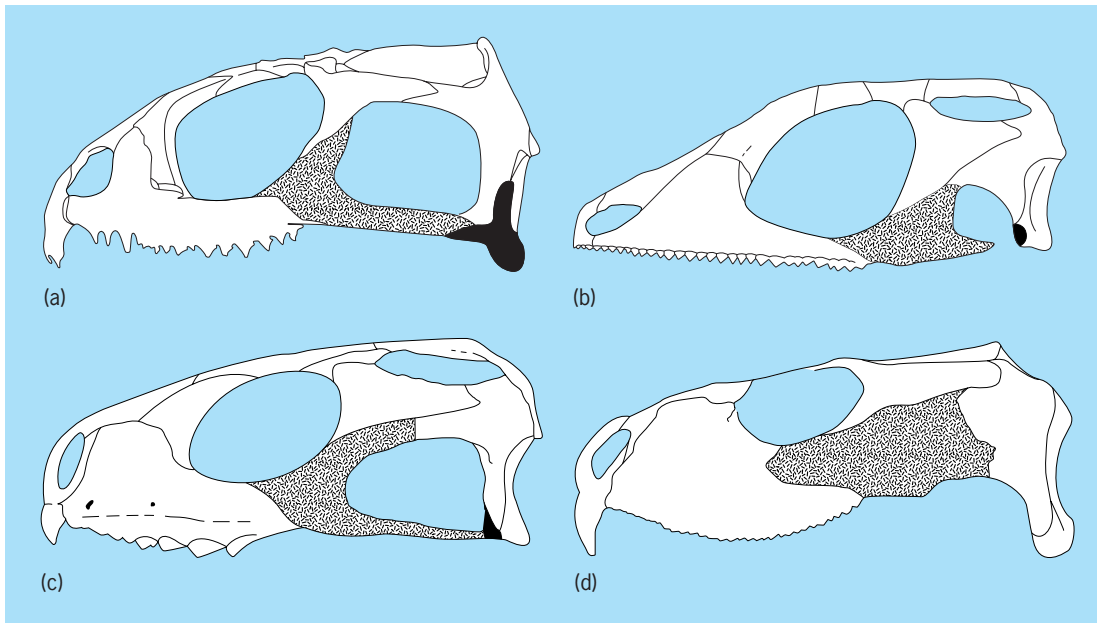
Fig. 1. Rhynchocephalian skulls in simplified lateral view. In each skull, the jugal is shown in stipple and the quadratojugal in black. (*a*) *Sphenodon* (Recent, New Zealand). (*b*) *Gephyrosaurus* (Lower Jurassic, Britain) [*redrawn from S. E. Evans, At the feet of the dinosaurs: The early history and radiation of lizards, Biolog. Rev., 78:513–551, 2003*]. (*c*) *Clevosaurus* (Upper Triassic, Britain) [*redrawn and simplified from N. C. Fraser, The osteology and relationships of Clevosaurus (Reptilia: Sphenodontida), Phil. Trans. Roy. Soc. Lond., B321:125–178. 1988*]. (*d*) The herbivorous *Priosphenodon* (Upper Cretaceous, Argentina) [*redrawn and simplified from S. Apesteguia and F. E. Novas, Large Cretaceous sphenodontian from Patagonia provides insight to lepidosaurian evolution in Gondwana, Nature, 425:609–612, 2003*].

male copulatory organs that characterize lizards and snakes) and an extended, tortoiselike lifespan. It also possesses a well-developed parietal eye, a light-sensitive organ on the top of the head that helps in the control of seasonal cycles. By day, tuatara live in shallow burrows excavated by seabirds; at night they emerge to feed. The typical diet consists of worms, snails, and insects, most notably large local crickets called wetas. However, seasonally the larger males will feed on bird chicks, small lizards, and young tuatara. Fertilization is internal. The eggs are buried and take 12–14 months to develop.

Gephyrosaurus. The earliest known rhynchocephalians date from the beginning of the Upper Triassic in Europe, but their relatively derived skulls suggest a sustained period of evolution before this. The primitive rhynchocephalian morphology is represented by *Gephryosaurus*, a small superficially lizardlike reptile from the Lower Jurassic of Britain (it was already a relict, or living fossil, by that time). The basal position of *Gephyrosaurus* is recognized by the fact that it is the only rhynchocephalian that is not included within the higher group Sphenodontia (that is, *Gephyrosaurus* plus Sphenodontia form Rhynchocephalia).

Other fossil groups. By the end of the Triassic, rhynchocephalians had spread globally and are known from India, Madagascar, South Africa, Brazil, North America, China, and Europe. They seem to have become extinct in Asia in the Early to Middle Jurassic, but elsewhere they survived into the Cretaceous. These survivors show greater anatomical diversity than their insectivorous ancestors, giving rise to derived long-bodied swimmers (for example, *Palaeopleurosaurus* and *Pleurosaurus*) and

herbivores with specialized dentitions (for example, *Priosphenodon*: see Fig. 1*d*). Despite this, the group seems to have become extinct in Europe and North America by the end of the Lower Cretaceous (approximately 110 million years ago). This created something of a paradox, given their survival in New Zealand to the present day. Recent finds of rhynchocephalians in Late Cretaceous deposits in South America (*Priosphenodon*, Argentina) have helped to

Fig. 2. Young *Sphenodon* (tuatara) at Chester Zoo, U.K. (*Photo by Susan Evans*)

bridge the gap, hinting at the possibility of a long, un-recorded post-Cretaceous history in southern conti-nents. *See* REPTILIA.   Susan E. Evans

Bibliography. M. J. Benton (ed.), *The Phylogeny and Classification of the Tetrapods*, vol. 1: *Amphibians, Reptiles and Birds*, Clarendon Press, Oxford, 1988; T. R. Halliday and K. Adler (eds.), *The New Encyclopedia of Reptiles and Amphibians*, Oxford University Press, Oxford, 2002; G. R. Zug, L. J. Vitt, and J. P. Caldwell, *Herpetology*, Academic Press, San Diego, 2001.
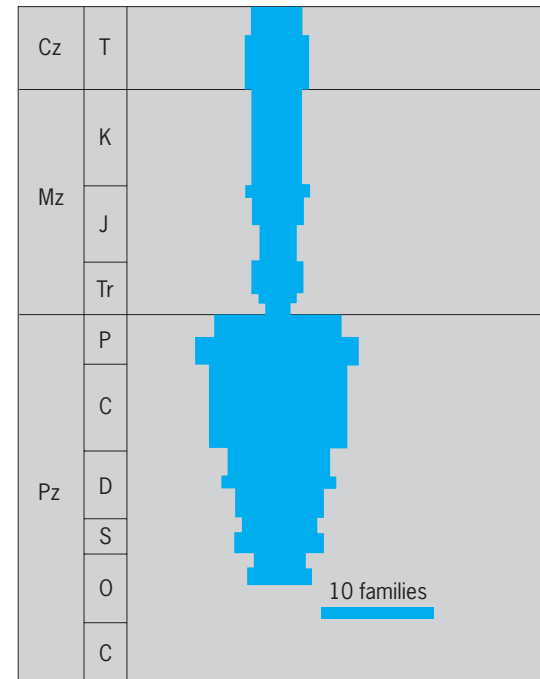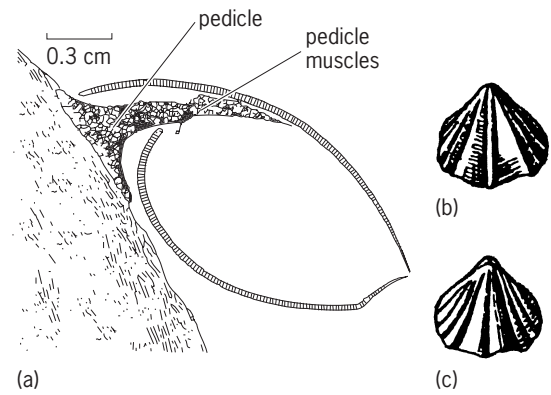
Rhynchonellida

An extant order of brachiopods that has been an important component of marine benthic communities since the Ordovician. As one of only two Paleozoic orders in the subphylum Rhynchonelliformea that survives to the present, the rhynchonellids exemplify the fate of the Paleozoic evolutionary fauna. Despite their stratigraphic longevity, rhynchonellids have been a remarkably conservative group morphologically.

Rhynchonellids possess unequally biconvex valves typically with a fold and sulcus; many species contain strong radial ribs that produce deflections in the commissure, the line of junction between the two valves. Their shells are generally impunctate and also typically lack a hinge line parallel to the hinge axis, resulting in a pointed beak or umbo when viewed in lateral profile (**illus.** *a–c*). Internally, rhynchonellids possess calcareous processes (crura) that in extant species provide support for the lophophore. The presence of crura is a key feature of rhynchonellid evolution. Rhynchonellids were derived from a group of pentamerid brachiopods that also possessed crura. Rhynchonellids are also thought to be the ancestral group for the spire-bearing brachiopods and then ultimately the terebratulids by elaboration of the crura into spiralia and then into a complex loop.

Rhynchonellids are sessile, attached, epifaunal suspension feeders. They have a functional pedicle that they use to attach to the substrate (illus. *a*). Although rhynchonellids were never diverse compared to other brachiopod orders, they were commonly important members of local communities. They achieved a diversity peak in the Devonian and again in the Jurassic (illus. *d*).

Rhynchonellids have shifted their habitat preference in the oceans since their origin in the Middle Ordovician when they originated in shallow low-latitude seas; however, presently they are more common in deep-water habitats from middle and high latitudes and are rare members of benthic communities in low-latitude shallow seas. This major shift in habitat preference represents a common pattern among marine orders over the last 550 million years and especially in the last 250 million years. Marine orders have tended to originate in shallow-water habitats, have expanded offshore, and in many cases have ultimately become restricted to deep-water en-

Rhynchonellida. (*a*) Lateral view of the modern species *Notosaria nigricans* with shell removed showing the pedicle system. (*b–c*) Ventral and dorsal views of *Rhynchotrema*, Ordovician. (*d*) Diversity history of rhynchonellid families. (*a* after J. R. Richardson, Brachiopods, Sci. Amer., 255:100–106, 1986; b and c after R. C. Moore, C. G. Lalicker, and A. G. Fischer, Invertebrate Fossils, McGraw-Hill, New York, 1952; d after J. J. Sepkoski, Jr., A Compendium of Fossil Marine Animal Families, 2d ed., Milwaukee Public Mus. Contrib. Biol. Geol., no. 83, 1992)

vironments. In the case of the rhynchonellids, their restriction from low-latitude shallow-water environments apparently occurred as early as the Jurassic and may reflect differential survival from the late Permian mass extinction. An alternative explanation suggests that interactions with other diversifying taxonomic groups dampened the rebound of the rhynchonellids from the late Permian mass extinction in low latitudes. *See* ARTICULATA (ECHINODERMATA); BRACHIOPODA.   Mark E. Patzkowsky

Bibliography. J. A. Walsh, No second chances? New perspectives on biotic interactions in post-Paleozoic brachiopod history, in P. Copper and J. Jin (eds.), *Brachiopods*, A. A. Balkema, Rotterdam, 1996.

Rhynchonelliformea

One of three subphyla currently recognized in the phylum Brachiopoda, a clade of lophophore-bearing protostome metazoan animals. The name is derived from the stratigraphically oldest subclade with extant representatives, the order Rhynchonellida (**Fig. 1**). Rhynchonelliforms constitute over 90% of the more than 4500 named brachiopod genera, most of which (over 95%) are extinct. Rhynchonelliformea includes
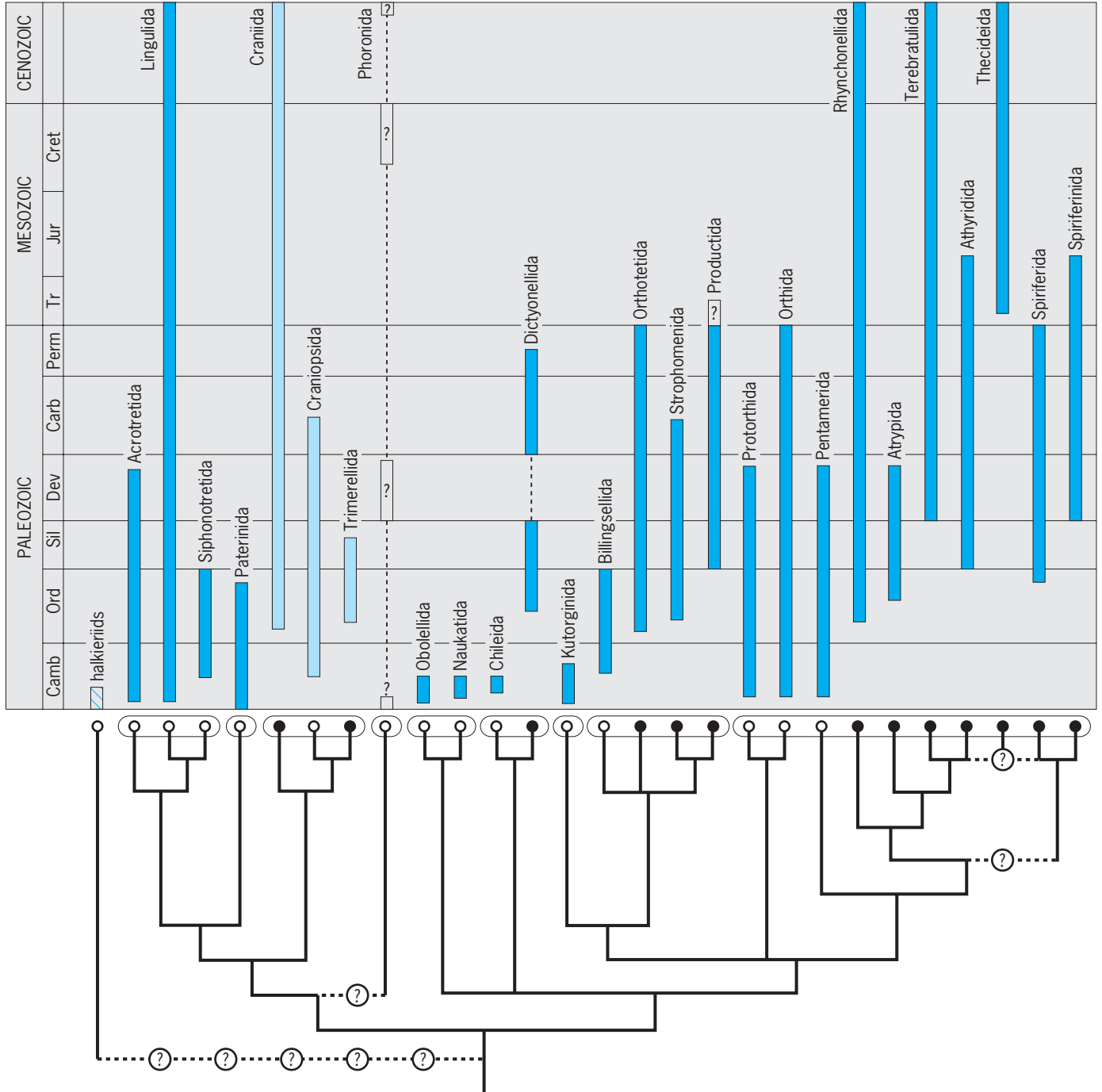

Fig. 1. Stratigraphic ranges and consensus cladogram of Brachiopoda, illustrating one hypothesis of phylogenetic relationship among the 26 orders of brachiopods currently recognized, and phoronids, constructed from data derived primarily from analyses of Cohen and Weydmann (2005) using molecular sequence data from living brachiopods and phoronids; Williams et al. (2000), Holmer and Popov (in Williams et al., 2000), and Popov et al. (in Williams et al., 2000), using morphological data from mostly Cambrian and Ordovician taxa; Carlson (1995) using morphological data from Recent brachiopods; and Carlson and Leighton (in Brunton et al., 2001) using morphological and stratigraphic data together for all rhynchonelliform suborders. Dark color of stratigraphic ranges indicates rhynchonelliforms; light color, linguliforms; and cross-hatched bar at far left, craniiforms. Open circles immediately below stratigraphic ranges indicate those orders first appearing in the Cambrian; closed circles, those first appearing in the post-Cambrian; elongated ellipses surrounding the circles identify the eight brachiopod classes and phoronids; encircled question marks indicate uncertainties in topology. (*Adapted from A. Williams and S. J. Carlson, in press, in R. L. Kaesler, ed., Treatise on Invertebrate Paleontology, pt. H, vol. 6, Geological Society of America and University of Kansas*)
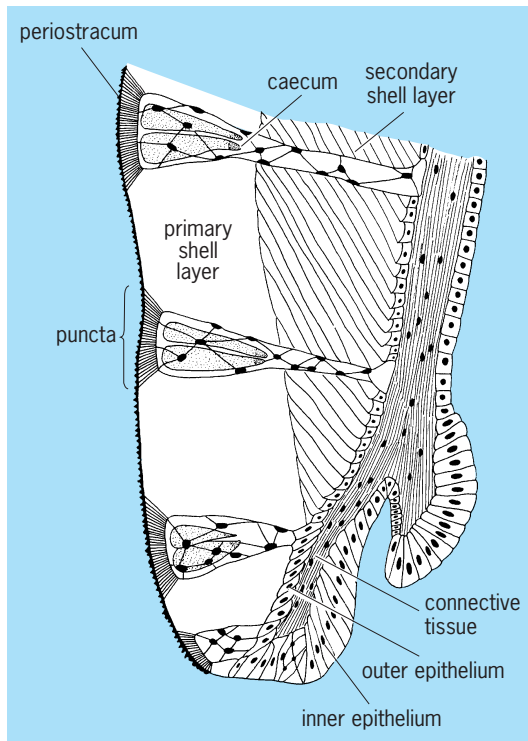
Fig. 2. Generalized section of a punctate rhynchonelliform shell and epithelium at anterior mantle margin. (*After R. L. Kaesler, ed., Treatise on Invertebrate Paleontology, pt. H, vol. 1, Geological Society of America and University of Kansas, 1997*)

all taxa formerly referred to the class Articulata, in addition to several groups newly discovered as fossils, formerly placed in class Inarticulata, or of uncertain taxonomic affiliation. Living Rhynchonelliformea are monophyletic, as confirmed by molecular and mor-

phological data. Phylogenetic relationships among the basal, extinct rhynchonelliforms (Fig. 1) are currently being evaluated.

All rhynchonelliforms have an organic-rich, bivalved calcite shell and can be distinguished from other brachiopods by the following shared, derived characters: the presence of a fibrous secondary shell layer (that transforms to other shell fabrics in some derived rhynchonelliform clades) (**Fig. 2**), a pedicle without a coelomic core (**Fig. 3**), a diductor (valve opening) muscle system (**Fig. 4**), and the presence of articulatory structures between the two valves (**Fig. 5**).

The rhynchonelliform shell is composed of a thin, outer, cryptocrystalline primary layer and a thicker, inner secondary layer of membrane-bound fibers (anvil-shaped in cross section), both mineralized by mantle epithelium. Regularly arranged cylindrical perforations of the shell called punctae (Fig. 2) have evolved more than once among the rhynchonelliforms (as in Terebratulida and Spiriferinida), and contain papillose mantle-tissue extensions called caeca. An unmodified, impunctate fibrous shell structure appears to be primitive for the subphylum (as in Rhynchonellida). A cross-bladed laminar shell structure evolved secondarily from a fibrous microstructure within the strophomenates; it commonly contains conical deflections of the secondary shell layer called pseudopunctae, which point inwardly and appear as tubercles on the valve interior.

The pedicle is a tough but flexible, nonmineralized, cuticle-covered, stalklike appendage by which the animal attaches to the substrate (Fig. 3). It emerges through an indentation of the posterior margin of the ventral valve called a delthyrium (or foramen in derived rhynchonelliforms). The
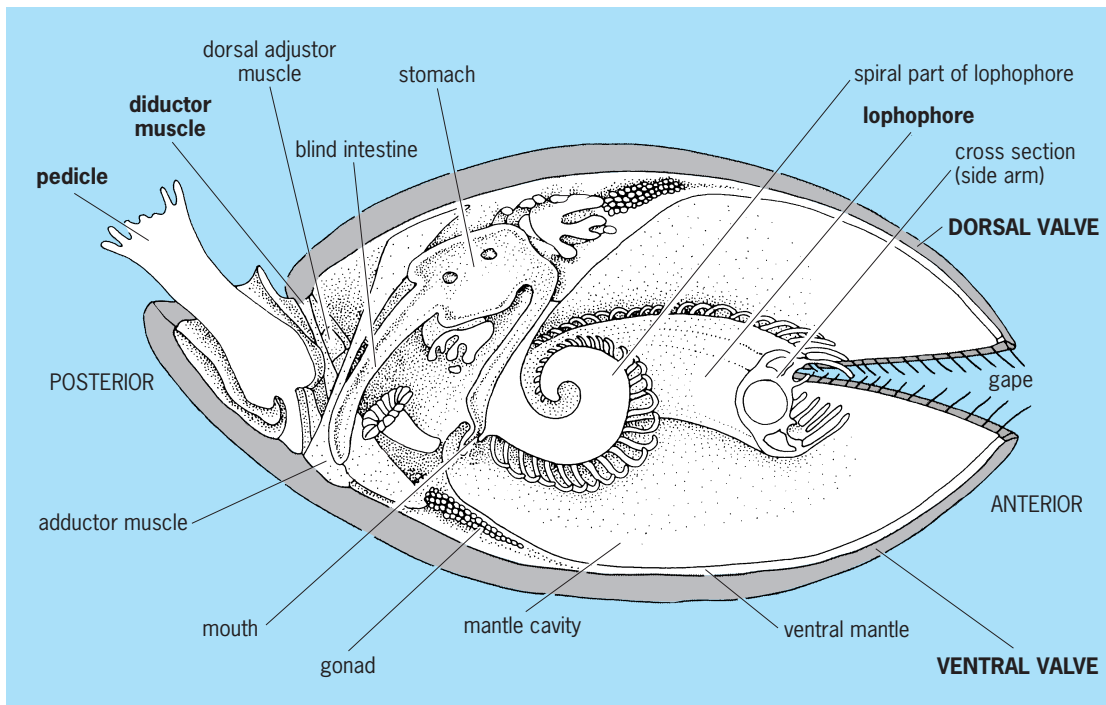

Fig. 3. Diagrammatic representation of the principal organs of *Terebratulina caputserpentis*. (*After R. L. Kaesler, ed., Treatise on Invertebrate Paleontology, pt. H, vol. 1, Geological Society of America and University of Kansas, 1997*)
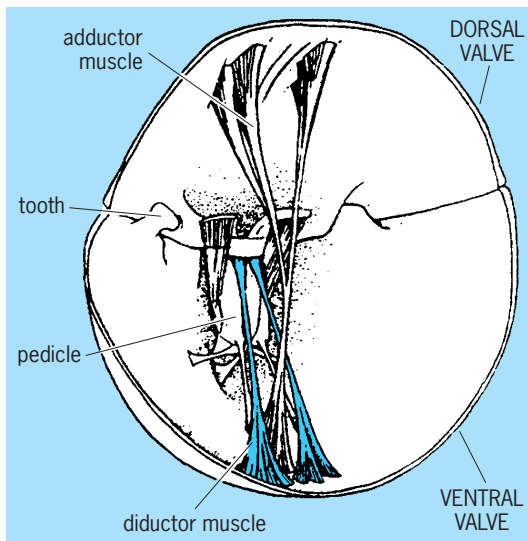
Fig. 4. Musculature of *Gryphus vitreus* in anterolateral view, dorsal valve above. (*After R. L. Kaesler, ed., Treatise on Invertebrate Paleontology, pt. H, vol. 1, Geological Society of America and University of Kansas, 1997*)

Disagreements about relationships still remain, but disparate sources of information are converging on a general pattern of evolutionary relationships (Fig. 1). Rhynchonelliformea includes 19 orders organized in five classes.

Phylum Brachiopoda
    Subphylum Linguliformea
            Craniiformea
            Rhynchonelliformea
    Class Chileata
      Order: Chileida
          Dictyonellida
    Class Obollellata
      Order: Obolellida
          Naukatida
    Class Kutorginata
      Order: Kutorginida
    Class Strophomenata
      Order: Strophomenida

pedicles of all extant rhynchonelliforms contain a core of connective tissue rather than a coelomic cavity, develop from a larval rudiment (not as an extension of the ventral body wall), and can be moved relative to the shell by external adjustor muscles. It is inferred that most extinct rhynchonelliforms possessed similar pedicles. The productides and thecideides, however, lost the pedicle as adults as evidenced by the lack of any delthyrial opening.

Diductor muscles (most commonly in two pairs) contract to open the valves by rotation about a hinge axis defined either by a straight, strophic or curved, astrophic hinge line (Figs. 4 and 5). They typically originate at a posterior projection of the dorsal valve called the cardinal process and insert broadly on either side of the paired adductor (closing) muscles on the ventral valve interior.

The oldest rhynchonelliforms possessed a rudimentary valve articulation (as in Kutorginida). The valves were in contact and rotated about a hinge axis, but dentition was poorly developed. Articulation among derived rhynchonelliforms is developed as two ventral teeth that fit into dorsal sockets on either side of the delthyrium. Tooth-and-socket articulation may be either noninterlocking (the primitive deltidiodont state, as in Orthida) or interlocking (the derived cyrtomatodont state, as in all extant rhynchonelliforms). Some strophomenates possess a row of denticles rather than a single pair of teeth, and most productides have lost dentition entirely. *See* ORTHIDA; STROPHOMENIDA.

Recent revision of the classification of brachiopods has attempted to incorporate more interpretable phylogenetic structure. Phylogenetic analyses have been completed on the extant taxa (with a broad range of characters, but very few taxa), the extinct taxa (with shell characters only, but many more taxa), and a combination of both.
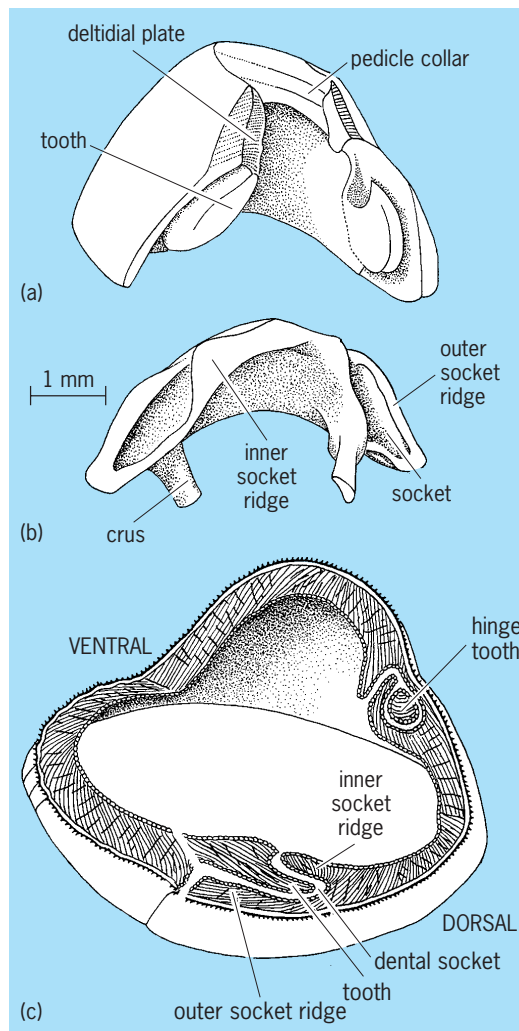

Fig. 5. Articulatory features of partial, posteriormost (*a*) ventral valve and (*b*) dorsal valve of *Terebratulina caputserpentis*; (*c*) view of combined transverse and submedian sections through valves. (*After R. L. Kaesler, ed., Treatise on Invertebrate Paleontology, pt. H, vol. 1, Geological Society of America and University of Kansas, 1997*)

Rhynchonelliform morphologies in the Cambrian are already rather diverse, as evidenced by the presence of eight different orders shortly after their first stratigraphic appearance. Molecular phylogeny research provides strong evidence for a clade of craniiforms and linguliforms together as the sister group to rhynchonelliforms.

Living rhynchonelliforms are distributed globally, and enjoy their greatest abundance and diversity in temperate latitudes. They live from rather deep intertidal to abyssal depths, reaching greatest abundance and diversity in the shallow to relatively deep (few hundred meters) subtidal. Extinct rhynchonelliforms, particularly in the Paleozoic Era, were far more abundant and diverse than today in virtually all geographic realms, and lived at nearly all depths. Fossil rhynchonelliforms have a continuous stratigraphic record from the Early Cambrian Period (Atdabanian Stage), over 540 million years ago, up to the present day. Cambrian diversity was low but significant. After a dramatic diversification in the Ordovician Period, generic diversity remained high through the Paleozoic Era. Rhynchonelliform diversity plummeted in the Permian extinction event, and has remained low (but very slowly increasing) ever since. *See* BRANCHIOPODA; BRACHIOPODA; PENTAMERIDA; SPIRIFERIDA; TEREBRATULIDA.

Sandra J. Carlson

Bibliography. C. H. C. Brunton, L. R. M. Cocks, and S. L. Long (eds.), *Brachiopods Past and Present*, Systematics Association Spec. Vol. Ser. 63, Taylor and Francis, London, 2001; S. J. Carlson, Phylogenetic relationships among extant brachiopods, *Cladistics*, 11:131–197, 1995; B. L. Cohen and A. Weydmann, Molecular evidence that phoronids are a subtaxon of brachiopods (Brachiopoda: Phoronata) and that genetic divergence of metazoan phyla began long before the Early Cambrian, *Organisms, Diversity Evolut.*, 5(3), 2005 (in press); M. A. James et al., Biology of living brachiopods, *Adv. Mar. Biol.*, 28:175–387, 1992; M. J. S. Rudwick, *Living and Fossil Brachiopods*, Hutchinson, London, 1970; A. Williams et al., Part H: *Brachiopoda*, revised, vols. 2–5: Linguliformea, Craniiformea, and Rhynchonelliformea (part), in R. L. Kaesler (ed.), *Treatise on Invertebrate Paleontology*, Geological Society of America and University of Kansas, Boulder and Lawrence, 2000–(in press) 2005.

Rhynchosauria

An order of herbivorous diapsid reptiles in the infraclass Archosauria, limited to the Triassic System but with a worldwide distribution. Rhynchosaurs were pig- or sheep-sized quadrupedal reptiles that were common in the Middle and Upper Triassic Series of India, South America, Europe, and eastern Africa. Except for the earliest South African genus, *Mesosuchus*, they are characterized by multiple rows of teeth on both the upper and lower jaws that were fused into deep sockets. Teeth were not replaced as they were worn but were added posteriorly as the jaw grew. In most genera, the premaxillae are devoid of teeth and overhang the front of the lower jaw like a beak. The external nostril is median rather than paired (see **illus.**).
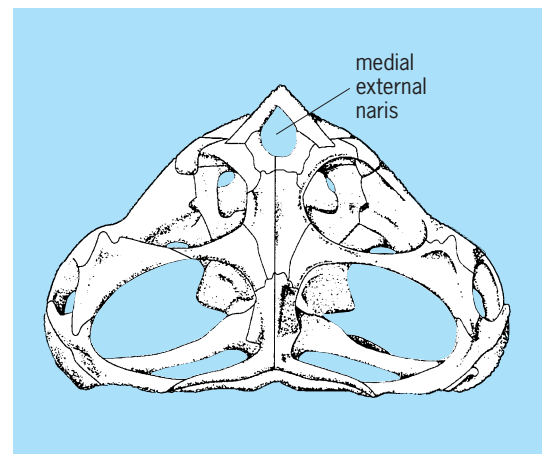
Skull (dorsal view) of the rhynchosaur *Paradapedon*.

Rhynchosaurs were elements of the early archosauromorph radiation that included the protorosaurs and primitive archosaurs. The structure of the ankle is nearly identical in the early members of these groups, but later rhynchosaurs enlarged the centrale which contributed to a simple hinge joint between the lower leg and the tarsals.

It was long thought that the Triassic rhynchosaurs were related to the primitive living genus *Sphenodon*, the New Zealand tuatara, which also has a beaklike premaxilla and a diapsid pattern of the temporal region. However, the teeth of *Sphenodon* differ significantly in being fused to the margin of the jaw rather than being set in sockets, and the beak is formed by premaxillary teeth, not by the bones of the premaxillae. The nature of rhynchosaurian ankle structure and dentition indicates affinities with archosaurs and protorosaurs. *See* PROTOROSAURIA.

Rhynchosaurs were locally common in the Late Triassic but are unknown in the Jurassic. Their rapid extinction may have resulted from changes in the vegetation on which they fed, or from predation from the expanding community of predaceous archosaurs, including early dinosaurs. *See* ARCHOSAURIA; DIAPSIDA; REPTILIA.

Robert L. Carroll

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1987.

Rhyniophyta

The first division under the subkingdom Embryobionta, in the classificatory system of A. Cronquist, A. Takhtajan, and W. Zimmermann. The bryophytes and vascular plants are included in this subkingdom. The category Rhyniophyta was devised for the relatively simple Silurian-Devonian vascular plants long held to be ancestral to other groups of vascular plants and usually referred to as Psilophytales. These plants have leafless stems and lack roots; their general morphological structure is not complex. Nevertheless, studies have shown that they fall into distinct groups on the basis of such characteristics as maturation of primary xylem, complexity of primary xylem in the stem, position of sporangia (terminal versus lateral), shape of sporangia (fusiform versus reniform), dehiscence of sporangia (longitudinal versus distal), and presence or absence of extensive lateral branch systems bearing large masses of terminal sporangia. The three classes of Rhyniophyta currently recognized are Rhyniopsida, Zosterophyllopsida, and Trimerophytopsida. *See* EMBRYOBIONTA; PSILOPHYTALES; RHYNIOPSIDA; TRIMEROPHYTOPSIDA; ZOSTEROPHYLLOPSIDA.   Harlan P. Banks

Rhyniopsida

The earliest demonstrable vascular land plants, appearing in Silurian (mid-Ludlovian) time. Their small, leafless axes were usually branched dichotomously in three planes; adventitious and perhaps pseudomonopodial branching also occurred. Sporangia were usually terminal on the main axes. Some terminated in lateral branches, and some were subtended by adventitious branches. The sporangia were globose, almost reniform, elongate-ellipsoidal, or cylindrical; some were columellate and branched. Dehis-
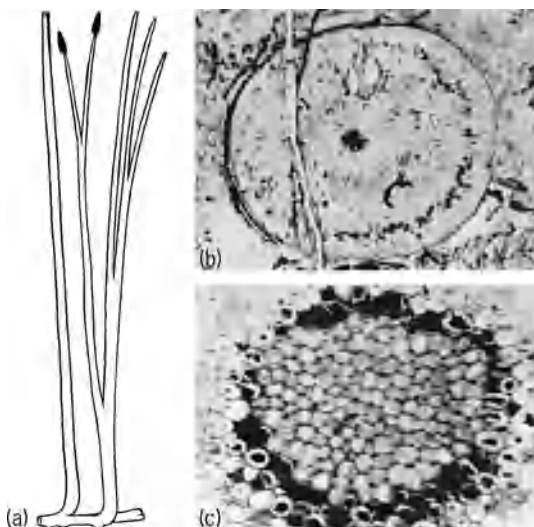
Rhynia. (*a*) Reconstruction showing leafless axis, basal rhizome with rhizoids (hairs), and fusiform, terminal sporangia. (*b*) Photomicrograph of transverse section of petrified stem. (*c*) Central xylem (vascular) strand. (*After Walton, An Introduction to the Study of Fossil Plants, A. and C. Black, 1940*)

cence was unspecialized; some opened by an apical slit, and some by a modified distal region. The spores were simple and produced in tetrads. A possible gametophyte plant, associated with *Rhynia*, was erect, leafless, and terminated by a gametangiophore bearing antheridia and possible archegonia. The basal parts, where known, were rhizomatous, with rhizoids but no true roots. Stomata have been found on some axes and sporangia. The xylem, where known, was centrarch, and consisted of annular elements. Evidence of phloem sieve cells is unconvincing.

Rhynia (see **illus.**) and *Horneophyton* are best known because of their cellular preservation. *Cooksonia*, the oldest genus, is numerically the most abundant. The group flourished in Late Silurian–Early Devonian time and may then have disappeared, although one genus is reported in Upper Devonian strata. *Cooksonia*, *Rhynia*, *Horneophyton*, and *Renalia* are the best-established genera. *Steganotheca*, *Salopella*, *Eogaspesiea*, *Dutoitea*, and some *Psilophytites* and *Hostinella* are probably rhyniophytes. *Taeniocrada*, *Hedeia*, and *Yarravia*, when better known, may belong elsewhere. Green algae were probable precursors, and the Trimerophytopsida appear to represent a higher evolutionary level. *See* PALEOBOTANY; RHYNIOPHYTA.   Harlan P. Banks

Rhyolite

A very light-colored, aphanitic (not visibly crystalline), volcanic rock that is rich in silica and broadly equivalent to granite in composition. Migration of rhyolitic magma through the Earth's crust, which causes much of the Earth's explosive and hazardous volcanic activity, represents a major process of chemical fractionation by which continental crust grows and evolves. *See* GRANITE.

Structure and eruptive style. Rhyolites are formed by the process of molten silica-rich magma flowing toward the Earth's surface. Small differences in this process, notably those related to the release of gas from the magma at shallow depth, produce extremely diverse structural features. Air-fall tephra eruptions form when molten rock foams as it rises to the surface, as a result of the gas being released; the foam breaks into a spray that is ejected into the air at high speeds. Large porous fragments making up the tephra layer are known as pumice; very small particles are known as ash. Some of the magma may not become airborne during an explosive event. Instead, the magmatic spray of ash particles, larger pumice fragments, and gas forms a dense, rapidly moving, ground-hugging fluid. Rhyolitic ash flow sheets formed in this way exhibit a number of characteristic features; these include glass shards, collapsed pumices, densely welded basal vitrophyres, welded devitrified interiors, and vapor-phase-altered zones. If the gas either has a low concentration or can escape easily in the shallow environment, rhyolitic magma sometimes remains intact and flows along the surface as a lava flow. The very viscous and slow-moving magma forms very thick flows and

domes that typically exhibit flow banding. *See* LAVA; MAGMA; PUMICE; TUFF; VOLCANIC GLASS; VOLCANO.

Composition, mineralogy, and texture. The high silica content gives rhyolitic lava a correspondingly high viscosity; this hinders crystallization and often causes young rhyolite to be a mixture of microcrystalline aggregates and glassy material. Rhyolites are generally porphyritic with numerous large crystals (phenocrysts) of sanidine (high-temperature form of potassium feldspar) and quartz disseminated throughout the aphanitic matrix. The crystalline portion of this groundmass consists of quartz intergrown with alkali feldspar. Rhyolite also contains minor amounts of dark minerals, including biotite, hornblende, or pyroxene. Many lavas of rhyolitic composition occur in the glassy state as obsidian or pitchstone. Obsidian is a black natural glass with a conchoidal fracture, whereas pitchstone has a dull appearance. Pitchstone contains up to 10% water and owes its appearance to secondary hydration and devitrification, the conversion of glass to a crystalline aggregate. Because of the glassy nature of most rhyolites, they are best characterized by chemical analysis. They typically have 70–75 wt % silicon dioxide (SiO_2) and more potassium oxide (K_2O) than sodium oxide (Na_2O). *See* FELDSPAR; OBSIDIAN; PITCHSTONE; PORPHYRY; QUARTZ.

Occurrence and origin. Rhyolite is one of the most common volcanic rocks in continental regions; it is virtually absent in the ocean basins. The rock often occurs in large quantities associated with andesite and basalt. It is common in environments ranging from accretionary prisms at continental margins to magmatic arcs related to subduction zones. Rhyolite is also prevalent in extensional regions and hot spots in continental interiors.

Noteworthy are the large and explosive caldera-forming ash flow eruptions. Although no eruption of this magnitude has been observed in historical times, they were quite common in the recent geological past, especially during the middle Tertiary time when vast areas of North, Central, and South America were blanketed with ash-flow sheets covering thousands of square kilometers with thicknesses of hundreds of meters.

The size and scope of these very explosive eruptions have been documented with studies at Yellowstone National Park (northwest-central United States), where three such eruptions occurred on a fairly regular schedule starting roughly 2×10^6 years ago. The first eruption was the largest. Within a few days, 2500 km^3 (600 mi^3) of dust, ash, and lava erupted, an amount 15,000 times greater than the amount blown out by Mount Saint Helens, Washington, in 1980. Similar but smaller eruptions occurred again 1.2×10^6 years ago and 600,000 years ago. Long Valley Caldera, California, and Valles Caldera, New Mexico, are the sites of other relatively recent, similar devastating eruptions. *See* CALDERA.

A variety of theories have been developed to explain the origin of rhyolite. Because it is largely confined to continental crust, many geologists believe that its origin is related to partial melting of pre-existing continental crust. In addition, because it is often associated with basalt and andesite, its origin may be tied to the evolution of these mantle-derived magmas, perhaps reflecting differentiation of the parent magma. More complicated models involving melting of the continental crust by intrusion of hot basaltic magma, fractional crystallization of the basalt, magma mixing, and assimilation of crustal material have been developed to explain the origin of rhyolite and associated rocks. *See* ANDESITE; BASALT; EARTH CRUST.

Research drilling. The area around Mammoth Lakes in northern California contains some of the youngest and most interesting volcanic features in the United States. The largest of these is the Long Valley Caldera, which formed 700,000 years ago during a catastrophic eruption of pyroclastic pumice and ash. The Inyo Domes chain is a much smaller, younger, and simpler feature. It is a line of rhyolitic lava domes, craters, and fractures that runs north-south across the rim of the caldera. The Inyo Domes chain last erupted 600 years ago with an event that rivaled the Mount Saint Helen's eruption of 1980. Six hundred years is a very short time with respect to geological processes; volcanoes that were active 600 years ago have changed very little, so that this region is an ideal place to study how rhyolite magma is formed and erupts.

As part of the Continental Scientific Drilling Program, research holes were drilled into the volcanoes at Inyo to sample the actual intrusions that fed these young eruptions. A special ring-shaped drill bit cut out a core of rock from the hole. Analyses of this core have provided information for constructing chemical and mathematical models for the intrusion and eruption of rhyolitic magmas. The drilling program has elucidated the shallow plumbing systems that feed young eruptions. The evidence collected also suggested that the nature of volcanic eruptions, quiet or explosive, may be controlled by the manner in which gas escapes at shallow depths, not by the original gas content of the magma, as was previously believed. Finally, identification of striking chemical zonation in the conduit to the dome may ultimately reveal a great deal about how magma flows in the Earth's crust. *See* IGNEOUS ROCKS.   Leland W. Younker

Bibliography. F. Barker, Granites and rhyolites, *J. Geophys. Res.*, special issue, 86(B11):10131–10638, 1981; I. S. E. Carmichael, F. J. Turner, and J. Verhoogen, *Igneous Petrology*, 1974; J. H. Fink (ed.), *The Emplacement of Silicic Domes and Lava Flows*, Geolog. Soc. Amer. Spec. Pap. 212, 1987; A. R. McBirney, *Igneous Petrology*, 2d ed., 1997; C. S. Ross and R. L. Smith, *Ash-Flow Tuffs: Their Origin, Geologic Relations, and Identification*, USGS Prof. Pap. 366, 1961.

Riboflavin

Also known as vitamin B_2, riboflavin is widely distributed in nature, and is found mostly in milk, egg white, liver, and leafy vegetables. It is a water-soluble

yellow-orange fluorescent pigment with the structural formula shown below. It is stable to acid and

$$CH_2 — (CHOH)_3 — CH_2OH$$
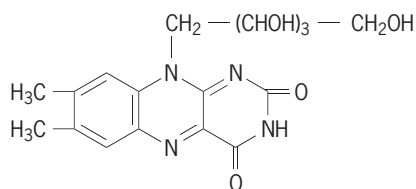
oxidation, but is rapidly destroyed by alkali at elevated temperatures and by light. Although little riboflavin is usually destroyed during cooking, half of the riboflavin in bottled milk exposed to the sun may be destroyed in 2 h.

Riboflavin is determined by microbiological methods using the lactic acid–forming organism *Lactobacillus casei*, which requires it for growth. Riboflavin is also determined by chemical methods in which its fluorescent properties are used.

Dietary requirement. A dietary source of riboflavin is required by all nonruminant animal species studied. Riboflavin deficiency results in poor growth and other pathologic changes in the skin, eyes, liver, and nerves. Riboflavin deficiency in humans is usually associated with a cracking at the corners of the mouth called cheilosis; inflammation of the tongue, which appears red and glistening (glossitis); corneal vascularization accompanied by itching; and a scaly, greasy dermatitis about the corners of the nose, eyes, and ears.

Riboflavin is found in all tissues with its concentration usually paralleling metabolic activity. The vitamin functions biochemically in two coenzymes, flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD). Enzymes containing these coenzymes are called flavoproteins. In general, these enzymes participate in oxidation-reduction reactions by accepting hydrogen ions (protons) and electrons from one substrate and transferring them to another. Over 14 riboflavin-containing enzymes have been studied. *See* COENZYME; ENZYME.

Most of the energy-yielding oxidations of food to carbon dioxide and water occur through the cytochrome system, in which the flavoprotein cytochrome reductase is essential. L-Amino oxidase, another flavoprotein, is important in the interconversion of amino acids to nonnitrogenous metabolities of carbohydrate and fat metabolism. *See* CYTOCHROME.

Riboflavin requirements appear to be related to caloric requirements and muscular activity, and are affected by heredity, growth, environment, age, and health. Evidence in animals suggests a need for increased riboflavin in low-protein diets because of a decreased ability of the liver to retain the vitamin. There is also evidence that as a result of intestinal synthesis by bacteria, less riboflavin is required on diets high in carbohydrate than on those high in fat. Human requirements for riboflavin are based primarily on urinary excretion data and studies in which riboflavin deficiency has been experimentally produced. Riboflavin allowances of the National Re-

search Council have been set at 0.07–0.1 mg/0.75 kg of body weight. Former allowances using a factor of 0.025 mg/g of protein allowance provided similar amounts of the vitamin. Stanley N. Gershoff

Industrial production. Riboflavin is produced commercially by either direct chemical synthesis or fermentation.

Chemical synthesis. Industrial syntheses of riboflavin generally proceed along the lines of the original Karrer method with modifications by individual producers. A typical manufacture starts with D-ribose, which is condensed with 1,3,4-xylidine and simultaneously hydrogenated to form *N*-D-ribitylxylidine. The ribitylxylidine is coupled with diazotized aniline to produce 1,2-dimethyl-4-D-ribitylamino-5-phenyl azo benzene. This azo compound is hydrogenated to 1,2-dimethyl-4-D-ribitylamino-5-aminobenzene, and then is condensed with a mixture of alloxan and alloxantin to riboflavin. By another method, the azo compound is condensed directly with barbituric acid. Leo A. Flexser

Fermentation. Riboflavin produced commercially by fermentation utilizes the synthetic ability of bacteria, yeasts, or fungi. The outstanding organisms for production of riboflavin are the two closely related ascomycete fungi, *Eremothecium ashbyii* and *Ashbya gossypii*. The inoculum is started from slants or from spores dried on sand, and after one or two flask stages is carried through one or two tank inoculum stages.

The final fermentation is carried out in tanks with a capacity of 10,000–100,000 gal (40,000–400,000 liters) with a medium suitable for the organism being used. For *Eremothecium* usually stillage (still slops) from the ethyl alcohol fermentation with skim milk, soybean meal, or casein added is used as a proteinaceous source; a carbohydrate source, such as maltose, sucrose, or glucose, is added. For *Ashbya*, commercially used media may contain corn-steep liquor, and usually also some animal protein such as crude peptones, animal-stick liquor, or fish-stick liquor; the carbohydrate sources are similar to those used for *Eremothecium*. The medium is aerated and usually agitated during fermentation for 96–120 h; optimum titers may be 3–6 g of riboflavin or more per liter. *See* ASCOMYCOTA; CARBOHYDRATE; CASEIN; ETHYL ALCOHOL; INDUSTRIAL MICROBIOLOGY; SOYBEAN.

Riboflavin may be recovered for animal feed supplements by evaporation of the whole broth in multiple-effect evaporators, followed by drum or spray drying. For drug and fine foods uses, pure crystalline riboflavin is isolated by heating the fermentation broth, filtration, and precipitation of the riboflavin with dithionite (hydrosulfite) followed by several purification steps, including crystallization. Ralph E. Bennett

Bibliography. G. F. Combs, Jr., *The Vitamins: Fundamental Aspects in Nutrition and Health*, 1997; H. J. Peppler and D. Perlman (eds.), *Microbial Technology*, vol. 2: *Fermentation Technology*, 2d ed., 1979; B. Volesky and J. Votruba, *Modeling and Optimization of Fermentation Processes*, 1992.

Ribonuclease

A group of enzymes, widely distributed in nature, which catalyze hydrolysis of the phosphodiester bonds in ribonucleic acid (RNA). The sites of hydrolysis may vary considerably, depending upon the specificity of the particular enzyme. Ribonuclease isolated from the pancreas appears to cleave RNA in two steps (**Fig. 1**). Cleavage occurs predominantly at P-O bonds which are attached to the 3′-OH groups of pyrimidine nucleotides. A number of plant ribonucleases catalyze similar reactions but are relatively nonspecific with regard to the nature of the base moiety. On the other hand, a ribonuclease from *Bacillus subtilis* appears to require purine rather than pyrimidine bases as part of the susceptible nucleotides. Even more stringent specificity requirements are found in two ribonucleases from *Aspergillus oryzae*, ribonuclease T_1 and ribonuclease T_2, which specifically hydrolyze the phosphodiester bonds adjacent to 3′-guanylate and 3′-adenylate groups, respectively. Differences in specificity for the site of cleavage have led to the use of these various ribonucleases as tools in determining the structure and chemistry of RNA.

The most thoroughly studied of the ribonucleases has been that isolated from bovine pancreas. This enzyme is the subject of the following presentation.

The systematic name for pancreatic ribonuclease proposed by the Enzyme Commission is polyribonucleotide 2-oligonucleotidotransferase (cyclizing). The small size and great stability of ribonuclease, coupled with the ease of isolation of pure enzyme in large quantities, have contributed notably to its successful characterization. Since the elucidation of the amino acid sequence of ribonuclease, much information has been compiled with regard to the three-dimensional structure of the enzyme and to specific regions of the molecule that are catalytically important. *See* ENZYME; NUCLEIC ACID.

Discovery, isolation, and purification. Pancreatic ribonuclease was discovered in 1920 by W. Jones, who described the hydrolysis of yeast nucleic acid by a heat-stable pancreatic extract. It has now been established that the enzyme is synthesized on the microsomes of the acinar cells in the pancreas and is then distributed throughout the cytoplasm. The techniques employed in the isolation of ribonuclease take advantage of the enzyme's unique stability to extreme conditions of temperature and pH; indeed, this stability has largely facilitated subsequent chemical studies of the protein. M. Kunitz was the first to crystallize ribonuclease (1940), and most commercial preparations of the enzyme have been obtained by means of his procedure or modifications thereof. These crystalline preparations contain at least six distinguishable, enzymatically active components. The major constituent, ribonuclease A, accounts for about 90% of the total activity and may be prepared from the crystalline material by chromatography. *See* CHROMATOGRAPHY; PANCREAS.

Covalent structure. The determination of the primary structure of ribonuclease A represents a milestone in protein chemistry, as ribonuclease was the first enzyme for which such information was obtained. This work was accomplished in 1962 largely through the efforts of W. H. Stein and Stanford Moore and their collaborators, using techniques which have since become routine tools in the structural analysis of a great many proteins. A two-dimensional representation of the covalent structure of ribonuclease A is shown in **Fig. 2**. This drawing indicates only the sequence of amino acids and the positions of the disulfide cross-links; it has no three-dimensional significance. The enzyme consists of a single polypeptide
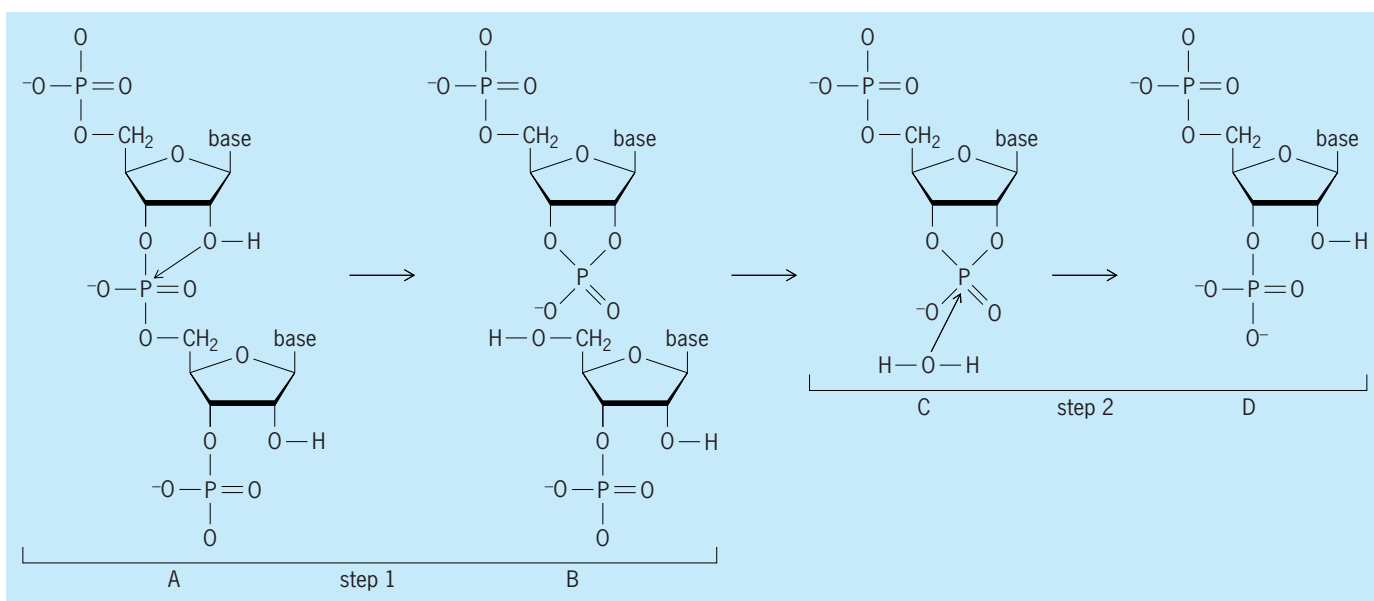
Fig. 1. Cleavage of RNA by pancreatic ribonuclease. In the first step, the oxygen atom of the 2′-OH group of the original, uncut RNA strand (species A) attacks the phosphorus atom of the adjacent phosphate group, cleaving the RNA strand and generating the cyclic 2′,3′-phosphate intermediate (species B). Part of the former RNA strand then detaches from the enzyme, leaving the remaining part of the RNA strand with the 2′,3′-phosphate cyclic end still bound to the enzyme (species C). In the second step, the oxygen atom of a water molecule attacks the phosphorus atom of the cyclic 2′,3′-phosphate group to form a noncyclic 3′-phosphate end (species D).

Ala Ala Lys Phe Glu Arg Glu (His) Met Asp Ser Ser Thr Ser Ala Ala Ser Ser
Ala   10   20   Ser
Thr   Glu Thr Asp Cys Tyr Glu Ser Tyr Ser Thr Met Ser Ile Thr   Asp
                        70   80   Asp   Tyr
Glu   Gly   Cys Cys
                Asp Lys Cys   Arg   Asp
Lys   Ala   COOH   Glu Glu
NH₂   Val   Val   Thr   Met
Ala Val Cys Ser Glu Lys Asp   Ser
                  60   Ala   Gly Met 30
Glu
        Ala Cys Glu Gly Asp Pro Tyr Val Pro Val (His) Phe Asp   Lys
Val Val   110   120   Ser
Asp Ile
        Ile His Lys Asp Ala Glu Thr Thr Lys Tyr Ala Cys Asp Pro Tyr Lys Ser   Ser
Ala   100   90   Arg
    Leu Ser Glu His Val Phe Thr Asp Val Pro (Lys) Cys Arg Asp Lys Thr Leu Asp
        50   40

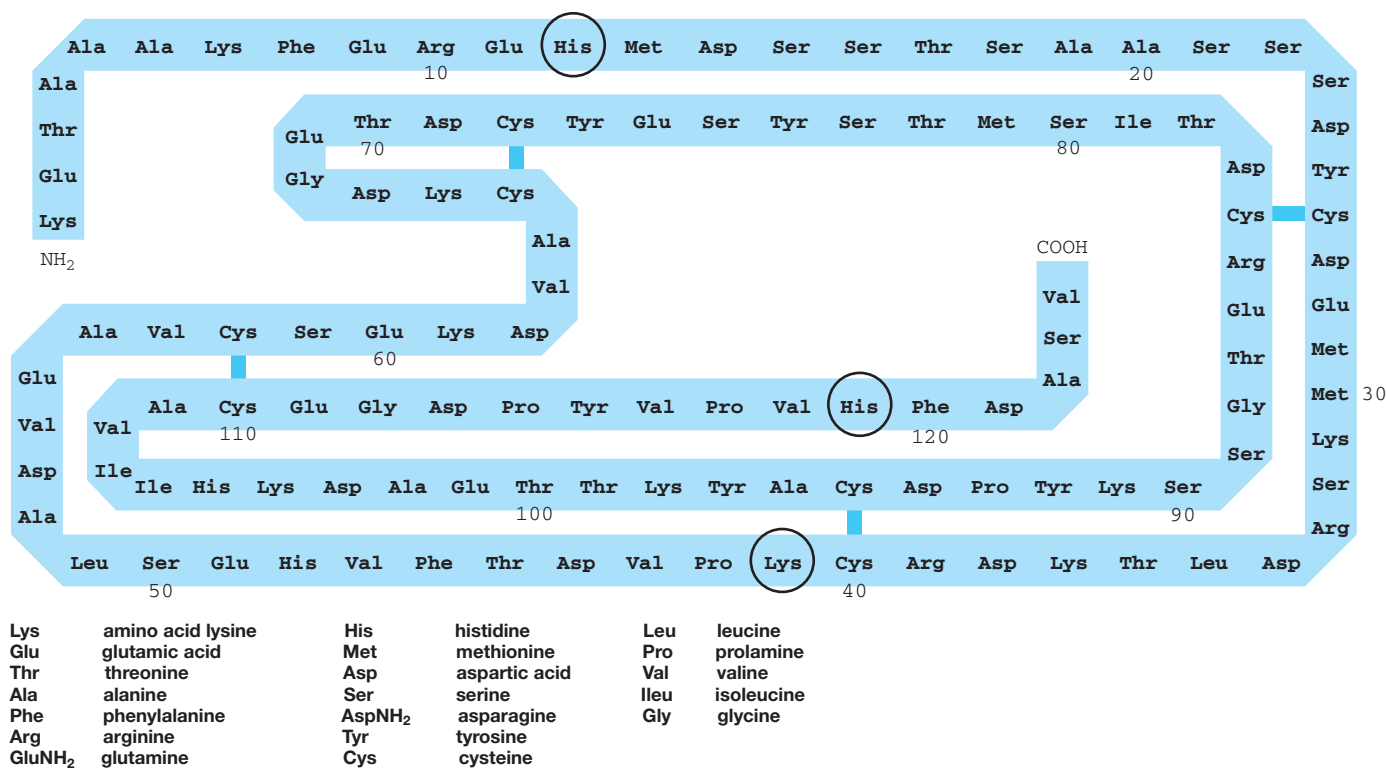| Lys | amino acid lysine | His | histidine | Leu | leucine |
Glu	glutamic acid	Met	methionine	Pro	prolamine
Thr	threonine	Asp	aspartic acid	Val	valine
Ala	alanine	Ser	serine	Ileu	isoleucine
Phe	phenylalanine	AspNH₂	asparagine	Gly	glycine
Arg	arginine	Tyr	tyrosine		
GluNH₂	glutamine	Cys	cysteine		

Fig. 2. Covalent structure of bovine pancreatic ribonuclease A.

chain of 124 amino acid residues containing 8 disulfide bridges between its 8 half-cystine residues. There are no free sulfhydryl groups or tryptophan residues in the enzyme. Particularly noteworthy are the histidines at positions 12 and 119 and lysine-41, all of which are encircled in Fig. 2. Evidence suggests that these residues form part of the active site of ribonuclease. The molecular weight of ribonuclease A calculated from its amino acid composition is 13,683, and this figure is in good agreement with values obtained by physical methods.

Research on ribonuclease has played a prime role in advancing the understanding of protein structure and function; also, it was the first protein to be totally synthesized from its component amino acids.

Tertiary structure. A chemical approach has been utilized with some success in pinpointing various aspects of the three-dimensional conformation of ribonuclease. The amino groups of all 10 lysine residues appear to lie at the surface of the enzyme, although the molecule is resistant to tryptic hydrolysis. Moreover, the enzyme in its native state is not hydrolyzed by a number of proteinases, suggesting that the peptide bonds may participate in hydrogen bonding of some sort which contributes to their stability to enzymatic attack. Tyrosine residues 25 and 97 are buried in the molecule, as are all four methionine residues, the side chains of which probably contribute to hydrophobic binding.

The most definitive information concerning the tertiary structure of the enzyme has been obtained by x-ray crystallography. Ribonuclease S is a fully active enzyme formed from ribonuclease A by hydrolysis with subtilisin, a proteinase which cleaves a single peptide bond in the enzyme between alanine-20 and serine-21. The tertiary structures of the two are essentially the same except that, in ribonuclease S, the residues involved in the subtilisin cleavage are separated by 10–15 angstroms (1–1.5 nanometers). A computer-generated illustration of the polypeptide chain folding in the model of ribonuclease A is shown in **Fig. 3**. Thus far, the x-ray models are compatible with, and confirm most of the data obtained by, chemical means. The positions of the four disulfide cross-links are the same as in the covalent representation (Fig. 2). Histidine-12, histidine-119, and lysine-41
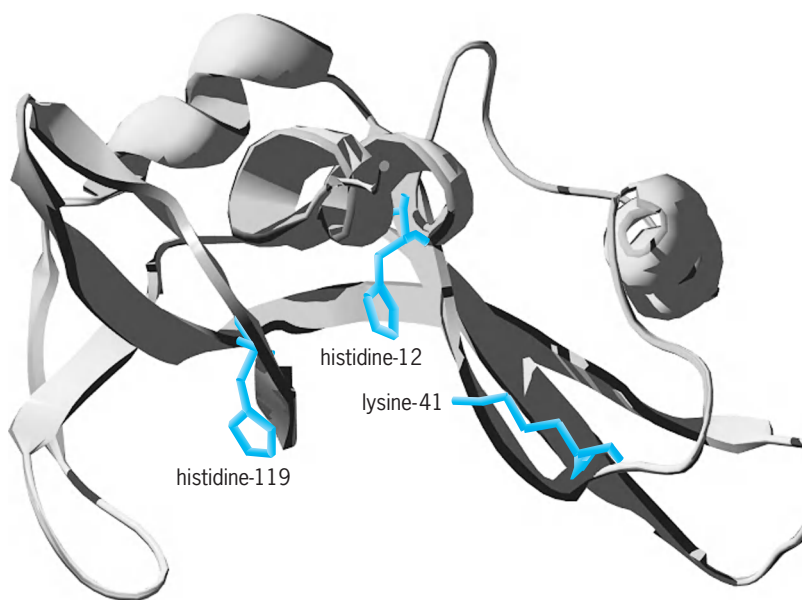

histidine-12
lysine-41
histidine-119

Fig. 3. Computer-generated illustration of the main chain folding in the ribonuclease A molecule.

are close together, in the region of the molecule where substrate is bound. Ribonuclease S contains 15% helix, in close agreement with the value predicted by optical methods (17%). The ribonuclease A molecule is kidney-shaped, with approximate dimensions of $38 \times 28 \times 22$ Å ($3.8 \times 2.8 \times 2.2$ nm). The disposition of charged side chains is toward the surface of the enzyme, while neutral groups are directed toward the interior. Much of the structure is exposed, and there is no part of the molecule which is shielded from the surrounding medium by more than a single layer of the main chain.

Activity. It has been postulated that the cleavage of RNA by ribonuclease occurs in two steps (Fig. 1), and activators and inhibitors of the enzyme may exert their effects on either phase or both phases of the overall reaction. Most of the kinetic studies of the enzyme have been carried out with pyrimidine nucleoside-2′,3′-cyclic phosphates as substrates, in which cases only the second step of the reaction has been assayed. Ribonuclease is active over a broad range of pH from 5 to 9, with optimal activity near 7. The enzyme is irreversibly inhibited by a number of alkylating agents and reversibly inhibited by Zn^{2+}, Cu^{2+}, phosphate, pyrimidine 2′- and 3′-phosphates, and a number of polyanions.

Active site. The active site is that region of the enzyme that interacts with the substrate and at which catalysis takes place. Definitive evidence has been presented which suggests that two of the four histidine residues in ribonuclease are at the active site and that they participate in the function of the enzyme. Different chemical groups may be introduced at a number of positions in ribonuclease without destroying, or even altering, its activity. However, if the histidine residues at positions 12 or 119 are chemically modified by a variety of alkylating agents, catalytic activity is abolished. The chemical modification of these highly reactive histidines is blocked by substances that bind to the active site of the enzyme. Moreover, if histidine-12 is chemically modified, it is no longer possible to react at histidine-119 and vice versa. On the basis of such chemical evidence it was postulated that histidine-12 and histidine-119 are close together even though widely separated in the polypeptide chain, and this has since been confirmed by x-ray analysis (Fig. 3). Evidence based upon chemical modification studies and x-ray analysis indicates that the lysine residue at position 41 is also near the active site, although it is not known whether its amino group participates in catalysis.

The active site would thus appear to consist of a cationic cluster involving the side chains of amino acid residues 119, 12, and 41. The positively charged site undoubtedly plays some role in the electrostatic attraction of binding of the negatively charged substrate. *See* PROTEIN; X-RAY CRYSTALLOGRAPHY.

Robert L. Heinrikson; M. Todd Washington

Bibliography. R. H. Abeles, P. A. Frey, and W. P. Jencks, *Biochemistry*, Jones and Bartlett, Boston, 1992; T. M. Devlin (ed.), *Textbook of Biochemistry with Clinical Correlation*, 5th ed., Wiley, New York, 2001; D. Voet and J. G. Voet, *Biochemistry*, 3d ed., Wiley, New York, 2004.

Ribonucleic acid (RNA)

One of the two major classes of nucleic acid, mainly involved in translating the genetic information carried in deoxyribonucleic acid (DNA) into proteins. Various types of ribonucleic acids function in protein synthesis: transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) function in the synthesis of all proteins, while messenger RNAs (mRNAs) are a diverse set, each member of which acts specifically in the synthesis of one protein. Messenger RNA is the intermediate in the usual biological pathway DNA → RNA → protein. Ribonucleic acid is a very versatile molecule, however. Other types of RNA serve other important functions for cells and viruses, such as the involvement of small nuclear RNAs (snRNAs) in mRNA splicing. In some cases, RNA performs functions typically considered DNA-like, such as serving as the genetic material for certain viruses, or roles typically carried out by proteins, such as RNA enzymes or ribozymes. *See* DEOXYRIBONUCLEIC ACID (DNA).

Structure and synthesis. RNA is a linear polymer of four different nucleotides (**Fig. 1**). Each nucleotide is composed of three parts: a five-carbon sugar known as ribose, a phosphate group, and one of four bases attached to each ribose, adenine (A), cytosine (C), guanine (G), or uracil (U). The structure of RNA is basically a repeating chain of ribose and phosphate moieties, with one of the four bases attached to each ribose. The structure and function of the RNA vary depending on its sequence and length. *See* NUCLEOTIDE; RIBOSE.

In its basic structure, RNA is quite similar to DNA. It differs by a single change in the sugar group (ribose instead of deoxyribose) and by the substitution of uracil for the base thymine (T). Typically, RNA does not exist as long double-stranded chains as does DNA, but as short single chains with higher-order structure due to base pairing and tertiary interactions within the RNA molecule. Within the cell, RNA usually exists in association with specific proteins in a ribonucleoprotein (RNP) complex.

The nucleotide sequence of RNA is encoded in genes in the DNA, and it is transcribed from the DNA by a complementary templating mechanism that is catalyzed by one of the RNA polymerase enzymes. In this templating scheme, the DNA base T specifies A in the RNA, A specifies U, C specifies G, and G specifies C.

Transfer RNA. These small RNAs (70–90 nucleotides) act as adapters that translate the nucleotide sequence of mRNA into a protein sequence. They do this by carrying the appropriate amino acids to ribosomes during the process of protein synthesis. Each cell contains at least one type of tRNA specific for each of the 20 amino acids, and usually several types. The tRNAs exist in a precisely folded, three-dimensional L-shaped structure (**Fig. 2**). Complementary base-pairing interactions between different parts of the tRNA play a major role in specifying this structure, with additional contributions from higher-order interactions. At one end of the L shape is the three-nucleotide anticodon sequence that binds

its complementary sequence in the mRNA, and at the other end is the amino acid specified by the mRNA codon. The base sequence in the mRNA directs the appropriate amino acid–carrying tRNAs to the ribosome to ensure that the correct protein sequence is made. About 25% of tRNA bases are unusual in being modifications (such as methylations) of the four standard bases. Among other roles, these modified bases are thought to facilitate recognition of the mRNA codon by tRNA. In the tRNA tertiary structure, a variable-length extra arm is accommodated within the hinge region that joins the two parts of the L-shaped structure. Few functions have been ascribed to the extra arm, although in a few cases it is known to contribute essential identity elements (in the form of its overall structure and/or specific nucleotide residues within it) for correct amino acid acceptor activity. *See* PROTEIN; RIBOSOMES.

Ribosomal RNA. Ribosomes are complex ribonucleoprotein particles that act as workbenches for the process of protein synthesis, that is, the process of linking amino acids to form proteins. Each ribosome is made up of several structural rRNA molecules and more than 50 different proteins, and it is divided into two subunits, termed large and small. A bacterial ribosome contains one copy each of 5S rRNA (~120 nucleotides) and 23S rRNA (~2900 nucleotides) in its large subunit, and one copy of 16S rRNA (~1540 nucleotides) in its small subunit. These rRNAs are also present in somewhat larger versions in the ribosomes of higher organisms, which also contain an additional rRNA of 5.8S (~160 nucleotides) in their large subunit.

The RNA components of the ribosome account for more than half of its weight. Like tRNAs, rRNAs are stable molecules and exist in complex folded structures. Each of these rRNAs is closely associated with ribosomal proteins and is essential in determining the exact structure of the ribosome. In addition, the rRNAs, rather than the ribosomal proteins, are likely the basic functional elements of the ribosome. They are actively involved in most, if not all, stages of protein synthesis. In this process, they make direct and sometimes transient contacts, via short stretches of base pairing, with the mRNA, tRNA, and each other. Recently, atomic-scale three-dimensional structures have been determined for the large and small ribosomal subunits. This work has shown the absence of any protein moiety within 18 Å of the site of peptide bond formation (within the peptidyltransferase center of the large subunit), confirming that the ribosome is indeed an RNA enzyme (ribozyme).

The nucleotide sequences of many rRNAs from different organisms have been characterized. The secondary stem-loop structures adopted by the RNAs are more highly conserved than the overall primary sequence, indicating that the higher-order structure is the functional one. Because these RNAs are present and perform the same central function in all known organisms, they are well suited for defining evolutionary relations. Sequence comparisons of 16S-type RNAs have been largely responsible for establishing the existence of the third major branch in the evolutionary tree, that of the archaebacteria, and in show-
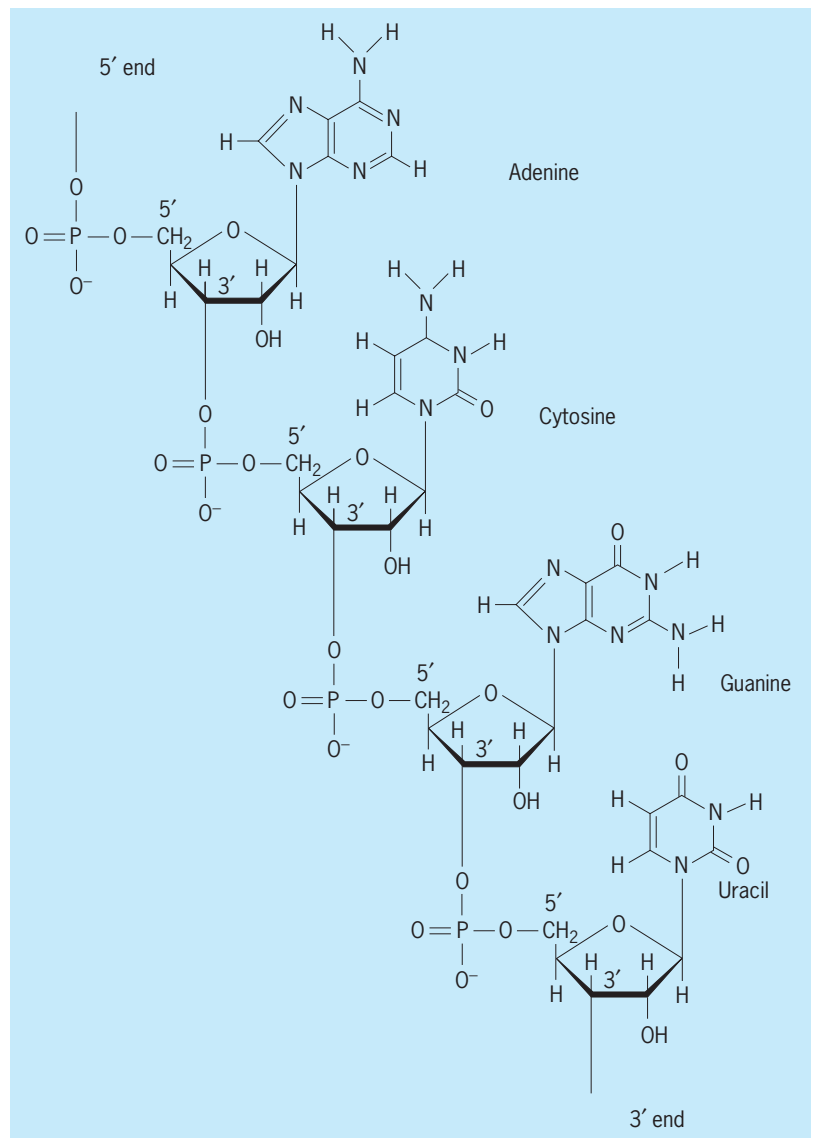

Fig. 1. Portion of an RNA molecule showing four nucleotides, each representing one of the four bases present in RNA.

ing that archaebacteria, eubacteria, and eukaryotes represent independent lines of descent from a common ancestor. *See* ARCHAEBACTERIA; EUKARYOTAE.

Messenger RNA. Whereas most types of RNA are the final products of genes, mRNA is an intermediate in information transfer. It carries information from DNA to the ribosome in a genetic code that the protein-synthesizing machinery translates into protein. Specifically, mRNA sequence is recognized in a sequential fashion as a series of nucleotide triplets by tRNAs via base pairing to the three-nucleotide anticodons in the tRNAs. There are specific triplet codons that specify the beginning and end of the protein-coding sequence. Thus, the function of mRNA involves the reading of its primary nucleotide sequence rather than the activity of its overall structure. A mammalian cell contains up to 20,000 different mRNAs, some of which are present in only a few copies and others in many thousands of copies, depending on how much of the protein product is needed. Most mRNAs range in length between 700
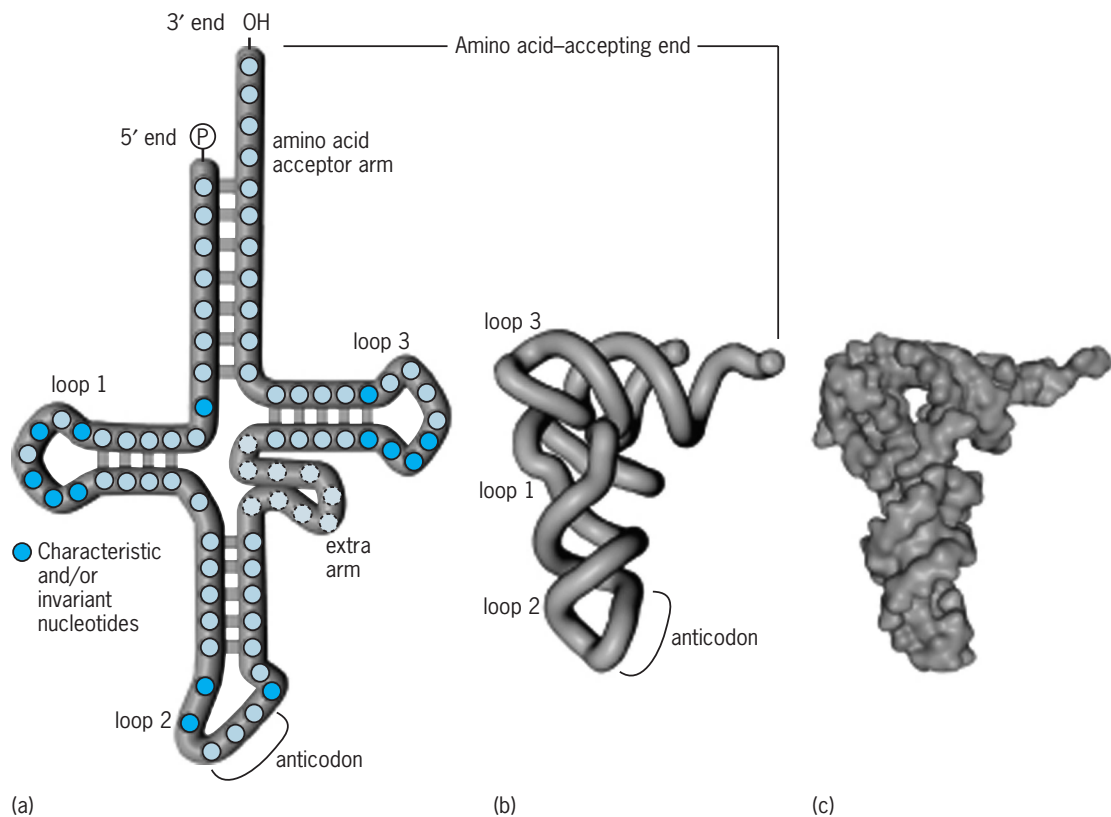
Fig. 2. Structural representations of a generalized tRNA molecule. (*a*) Secondary structure model showing base-pairing interactions within the tRNA. The characteristic and/or invariant nucleotides are the same regardless of the type of tRNA; the remainder vary. The extra arm varies in size among tRNAs. (*b*) Outline of the overall three-dimensional conformation of the tRNA. (*c*) Surface model of the tRNA.

and 10,000 nucleotides, with the length determined primarily by the size of the protein encoded. Messenger RNAs are typically shorter-lived than the more stable structural RNAs, such as tRNA and rRNA. *See* GENETIC CODE.

In eukaryotes, mRNAs are derived from larger precursor RNAs by a series of RNA processing steps that take place within the cell nucleus. Both ends of the pre-mRNA are modified: the 5′ end by addition of a methylated guanine nucleotide cap and the 3′ end by the addition of 50–200 adenosine residues in a non-templated reaction to form a poly(adenosine) tail. Additionally, the informational, or coding, sequences (exons) of most genes are interrupted by long stretches of noncoding sequences known as introns. A typical mRNA precursor consists of alternating exons and introns, with most of the RNA being intron. These long pre-mRNAs must be spliced to remove noncoding intervening sequences. The mRNA splicing process, which consists of RNA cleavage at junctions between exons and introns and ligation of exons, is mediated by another class of cellular RNAs, the small nuclear RNAs (snRNAs). *See* EXON; INTRON.

Small nuclear RNA. Small RNAs, generally less than 300 nucleotides long and rich in uridine (U), are localized in the nucleoplasm (snRNAs) and nucleolus (snoRNAs) of eukaryotic cells, where they take part in RNA processing. These RNAs have characteristic folded structures and function as ribonucleoproteins (snRNPs and snoRNPs), in association with specific proteins. Five snRNPs (U1, U2, U4, U5, U6) participate in intron removal during eukaryotic mRNA splicing, whereas a sixth (U7) has a role in 3′-end formation in histone mRNAs. Several snoRNPs (in particular, U3) are involved in specific steps of precursor rRNA (pre-rRNA) cleavage, whereas many others participate in the extensive posttranscriptional modification that occurs during production of mature rRNA. *See* CELL NUCLEUS.

snRNA. Intron splicing occurs in the spliceosome, a particle assembled from precursor mRNA, snRNPs, and proteins (splicing factors). Intron sequences up to many thousands of nucleotides long are precisely removed, with the flanking exon (coding) sequences being joined in the correct reading frame for translation. The precision and efficiency of splicing are achieved through complementary base pairing between short conserved sequences in the snRNAs and at the intron-exon boundaries (splice junctions). For example, U1 snRNA pairs with the 5′ splice junction, whereas U2 snRNA recognizes and binds to a sequence in the intron just upstream of the 3′ splice junction and encompassing the so-called "branchpoint A" residue. Base-pairing interactions between U2 and U6 snRNAs in the spliceosome lead to activation of the 3′-hydroxyl of the branchpoint A residue so that it functions as a nucleophile in an attack on the 5′ splice junction, thereby initiating splicing. Two successive cuts in the pre-mRNA chain (first at the 5′-splice junction, then at the 3′-splice junction) are

required for intron removal, which is followed by joining of the flanking exons. Although not definitive, mounting evidence indicates that splicing is catalyzed by snRNA rather than by protein, so that the spliceosome (like the ribosome) may be fundamentally an "RNA machine."

snoRNA. One class of modification-guide snoRNP, box C/D, mediates ribose methylation through base pairing with pre-rRNA, with different snoRNAs selecting different residues within a targeted region of pre-rRNA for 2′-O-methylation. The sequence in the snoRNA that is complementary to the modification target in pre-rRNA is immediately upstream of the 5′-end of box D, one of the short conserved sequence motifs that characterize a box C/D snoRNA. The precise site of 2′-O-methylation is specified by the "N + 5" rule, such that the pre-rRNA residue that is modified is always the one that is paired to the fifth snoRNA nucleotide upstream of the D box, within the helical region formed by base pairing between pre-rRNA and snoRNA. Box C/D snoRNPs contain a common set of proteins, one of which (fibrillarin) is the methyltransferase that carries out the formation of 2′-O-methylribose.

A second class of modification-guide snoRNP (box H/ACA) is responsible for formation of pseudouridine (5-ribosyluracil) residues in rRNA. The site of pseudouridine formation is also determined by complementary base pairing between a box H/ACA snoRNA and its target sequence in pre-rRNA. In this case, however, the U residue selected for modification remains unpaired within a pseudouridylation pocket flanked by short pre-rRNA:snoRNA helices. Box H/ACA snoRNPs also contain a common set of proteins, different than the box C/D set, one of which (Cbf5p; dyskerin in humans) is the pseudouridine synthase that carries out the modification.

Although most snoRNAs function in cleavage and modification of pre-rRNA in eukaryotic cells, a subset directs formation of 2′-O-methylnucleosides and pseudouridine residues in spliceosomal snRNAs. This subset of small nuclear RNAs, designated "scaRNAs," is localized in a nucleoplasmic structure known as the Cajal body. Except for the fact that they function in a nonnucleolar milieu, scaRNAs are structurally and functionally indistinguishable from box C/D and box H/ACA snoRNAs.

Catalytic RNA. RNA enzymes, or ribozymes, are able to catalyze specific cleavage or joining reactions either in themselves or in other molecules of nucleic acid. For example, one of the ribonucleases involved in tRNA processing, RNase P, is composed of one protein and one RNA molecule, and under certain conditions the enzymatic activity can be supplied entirely by the RNA moiety. Another example is self-splicing introns found in various genes, such as the rRNA gene of the protozoon *Tetrahymena*. These introns mediate their own removal from an RNA molecule in a reaction that is dependent on the specific higher-order structure in the intron RNA. Presumably the three-dimensional surface of the RNA results in an arrangement of atoms such that certain covalent bonds are strained and unusually reactive,

becoming susceptible to cleavage catalyzed by a bound divalent metal ion [usually magnesium (Mg^{2+})]. Several other types of catalytic RNA are categorized by specific conserved features of primary and secondary structure. *See* CATALYSIS; RIBOZYME.

Viral RNA. While most organisms carry their genetic information in the form of DNA, certain viruses, such as polio and influenza viruses, have RNA as their genetic material. The viral RNAs occur in different forms in different viruses. For example, some are single-stranded and some are double-stranded; some occur as a single RNA chromosome while others are multiple. In any case, the RNA is replicated as the genetic material, and either its sequence, or a complementary copy of itself, serves as mRNA to encode viral proteins. The RNA viruses known as retroviruses contain an enzyme that promotes synthesis of complementary DNA in the host cell, thus reversing the typical flow of information in biological systems. Many transforming viruses, which convert healthy cells to tumor cells, are retroviruses, as is human immunodeficiency virus (HIV), the causative agent of acquired immune deficiency syndrome (AIDS). *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME; ANIMAL VIRUS; RETROVIRUS; VIRUS.

As is evident from the above discussion, RNA molecules may function both as carriers of genetic information and as enzymes. The discoveries of RNA catalysis and of the central role of rRNA in protein synthesis have led to an enhanced appreciation of RNA as the probable original informational macromolecule, preceding both the more specialized DNA and protein molecules in evolution. *See* MOLECULAR BIOLOGY; NUCLEIC ACID.

Other types of RNA. RNA molecules serve other important and diverse cellular functions. For example, a ribonucleoprotein enzyme (telomerase) is responsible for replication of chromosome ends, with the RNA component functioning as a template for the reverse transcriptase activity of telomerase. An essential RNA component is also present in the signal recognition particle (SRP), a ribonucleoprotein complex that directs ribosomes synthesizing membrane and secreted proteins to the endoplasmic reticulum, where synthesis and posttranslational processing are completed. Finally, transfer-messenger RNA (tmRNA) functions during translation to rescue stalled ribosomes, targeting aberrant, partially synthesized proteins for proteolytic degradation.

Editing. RNA editing describes several different posttranscriptional processes that alter the nucleotide sequence of a mature RNA species compared to that of the gene that encodes it. In the case of mRNA, RNA editing results in codon substitutions that change the final protein sequence and that may also create new translation initiation and termination signals. Noncoding RNAs (rRNA, tRNA) sometimes undergo editing, as do certain viral RNAs.

Several different types of RNA editing have been described, such as insertions and deletions of uridine residues in trypanosome mitochondria, cytidine to uridine changes in plant mitochondria and chloroplasts, and adenosine to inosine substitutions in

animal nuclei. Editing ranges from a single cytidine deamination to generation of more than 50% of the final sequence by uridine addition. In trypanosome mitochondria, small guide RNAs (gRNAs) base-pair with the pre-mRNA in such a way as to determine the positions at which insertion or deletion of uridine residues occurs, thereby creating translatable open reading frames that did not exist prior to editing. Through approaches involving biochemical and genetic characterization, considerable progress has been made recently in identifying the RNA and protein components of editing complexes (editosomes) in several editing systems.

Antisense RNA: micro RNAs and small interfering RNAs. Antisense RNA is an RNA transcript (or portion of one) that is complementary to another nucleic acid, usually another RNA molecule. Naturally occurring examples of antisense RNA include, as mentioned earlier, the snoRNAs that mediate site-specific cleavage and modification of rRNA precursors, and gRNAs that supply the information for the uridine insertion/deletion type of RNA editing. In both bacterial and eukaryotic cells, other naturally occurring antisense RNAs function to regulate gene expression by binding to a target RNA molecule and interfering with its normal function. In eukaryotic cells, introduction of antisense RNA (for example, through activation of an appropriate transgene) has been shown to modulate expression of the complementary gene. This observation has encouraged attempts to use antisense RNA to manipulate gene expression ("gene silencing") for therapeutic and other purposes.

An unexpected and exciting consequence of antisense RNA studies has been the discovery of small interfering RNAs (siRNAs) that mediate the phenomenon of RNA interference (RNAi). An enzyme called Dicer generates siRNA via endonucleolytic cleavage of long double-stranded RNA molecules, which yields siRNA products approximately 22 nucleotides in length. Small interfering RNAs that are complementary to a long RNA molecule are able to target that molecule for degradation, thereby affecting RNA stability. Introduction or expression of mRNA-specific siRNAs in target cells is proving to be a powerful means of selectively and comprehensively inactivating gene expression. *See* GENE ACTION.

Micro RNAs (miRNAs) are another class of small RNA species that constitute a natural counterpart of siRNAs. They are the same size as siRNAs and are also generated by Dicer, although miRNAs are processed from longer single-strand precursors that are capable of assuming a hairpin configuration. Micro RNAs specifically regulate the activity of target genes, usually through translational arrest as a result of complementary base pairing with a sequence within the 3′ untranslated region of the corresponding mRNA. They appear to be key components of a novel RNA-based system of gene regulation that is evolutionarily conserved among at least multicellular eukaryotes, and that controls timing of development, stem cell maintenance, and other developmental and physiological processes in plants and animals. *See* RNA INTERFERENCE.   Ann L. Beyer; Michael W. Gray

Bibliography. B. Alberts et al., *Molecular Biology of the Cell*, 4th ed., 2002; T. R. Cech, RNA as an enzyme, *Sci. Amer.*, 255(5):64–75, 1986; R. F. Gesteland, T. R. Cech, and J. F. Atkins (eds.), *The RNA World*, 2d ed., Cold Spring Harbor Laboratory Press, 1999; H. Grosjean and R. Benne (eds.), *Modification and Editing of RNA*, American Society for Microbiology, Washington, DC, 1998; A. Rich and S. H. Kim, The three-dimensional structure of transfer RNA, *Sci. Amer.*, 238(1):52–62, 1978; J. A. Steitz, "Snurps," *Sci. Amer.*, 258(6):58–63, 1988.

Ribosomes

Small particles, present in large numbers in every living cell, whose function is to convert stored genetic information into protein molecules. In this synthesis process, a molecule of messenger ribonucleic acid (mRNA) is fed through the ribosome, and each successive trinucleotide codon on the messenger is recognized by complementary base-pairing to the anticodon of an appropriate transfer RNA (tRNA) molecule, which is in turn covalently bound to a specific amino acid. The successive amino acids become linked together on the ribosome, forming a polypeptide chain whose amino acid sequence has thus been determined by the nucleic acid sequence of the mRNA. The polypeptide is subsequently folded into an active protein molecule. Ribosomes are themselves complex arrays of protein and RNA molecules, and their fundamental importance in molecular biology has prompted a vast amount of research, with a view to finding out how these particles function at the molecular level. *See* GENETIC CODE; PROTEIN; RIBONUCLEIC ACID (RNA).

Components. Ribosomes are composed of two subunits, one approximately twice the size of the other. In the bacterium *Escherichia coli*, whose ribosomes have been the most extensively studied, the smaller subunit (named 30S, according to its sedimentation coefficient) contains 21 proteins and a single 16S RNA molecule. The larger (50S) subunit contains 32 proteins, and two RNA molecules (23S and 5S). The proteins vary in size from about 50 to 500 amino acids; their amino acid sequences have all been determined. Each protein is present in a single copy per ribosome, with the exception of one of the large subunit proteins, for which there are four copies. The RNA molecules have also been sequenced and have lengths of 120 nucleotides (5S RNA), about 1540 nucleotides (16S), and 2900 nucleotides (23S). The overall mass ratio of RNA to protein is about 2:1, and the combined molecular weight of the two subunits is about 2.5×10^6 daltons. Cations, in particular magnesium and polyamines, play an important role in maintaining the integrity of the ribosomal structures.

Comparative studies on the ribosomal components from other prokaryotic species have shown

that the sequences of both proteins and RNA have been highly conserved during evolution. Furthermore, the secondary structure of the RNA, that is, the way in which the RNA is folded into loops by complementary base pairing, has remained virtually identical in all the species studied. The ribosomes are considerably larger in the cytoplasm of higher organisms (eukaryotes). They have a higher protein content, and the two principal RNA molecules are up to 1900 and 4700 nucleotides in length for the small and large subunits, respectively. In contrast, ribosomes from organelles such as mitochondria are much smaller, mammalian mitochondrial ribosomes having RNA molecules of only about one-half the size of their bacterial counterparts. Nevertheless, it has become clear that, despite these gross differences in overall size, all ribosomal RNA molecules have a central core of conserved structure, which presumably reflects the universality of the ribosomal function. *See* CELL (BIOLOGY); ORGANIC EVOLUTION.

Arrangement of components. An understanding of ribosomal function requires a knowledge of the spatial arrangement of the individual ribosomal components; many techniques have been applied with this end in view, mostly using the ribosome of *E. coli* as a substrate. Electron microscopy of the ribosomal subunits has revealed that both subunits have a characteristic morphology or shape (**Figs. 1** and **2**), and that this property can be exploited by using antibodies to the individual proteins; the antibody connects two subunits together, forming a dimer, and the site of antibody attachment can be observed in the electron microscope. In this way a number of ribosomal proteins have been mapped on the subunit surface (Fig. 2). Antibodies to modified nucleotides in the RNA (the RNA contains a few such modifications) can be used in the same way. Other mapping techniques involve the use of physicochemical methods to measure the distance between two selected components, which can be labeled with fluorescent
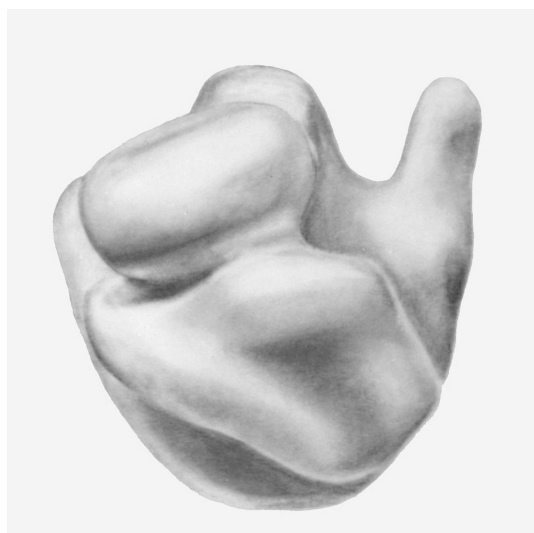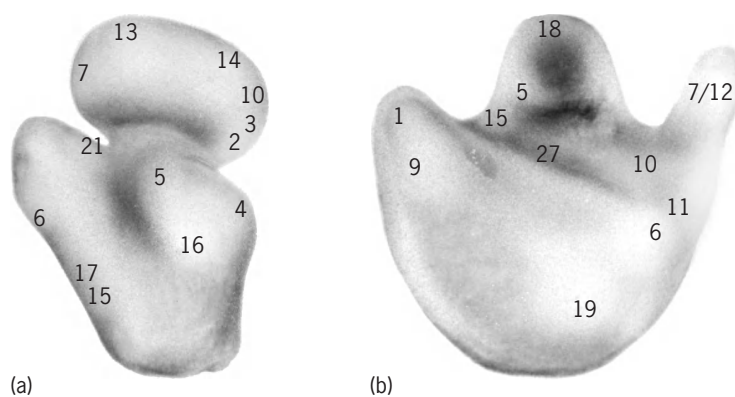

Fig. 2. Models, from electron microscopic data, of (*a*) 30S ribosomal subunit and (*b*) 50S subunit. Numbers identify some proteins that have been localized on subunit surface.

markers or with deuterium. In the latter case, neutron scattering measurements are made to determine the distance between the two components under consideration. Crystals have also been obtained from isolated ribosomal proteins or even whole subunits, which opens the possibility of making detailed structural analyses via x-ray diffraction techniques. *See* ANTIBODY; X-RAY DIFFRACTION.

Chemical cross-linking techniques have been widely applied to investigate "neighborhoods" (as opposed to absolute distances) between different parts of the ribosome. Here the subunits are treated with a suitable bifunctional reagent, and the subsequent analysis determines which parts of the ribosome are close enough to become cross-linked to one another. The technique can be applied equally to the study of protein-protein neighborhoods, protein-RNA neighborhoods, or neighborhoods between different regions of the RNA. It also lends itself well to the study of the ribosomal interface—the region of contact between the large and small subunits. Chemical modification or susceptibility to enzyme digestion can be used in a similar manner to determine which regions of the ribosome are accessible on the surface of the particles.

A number of the methods just outlined rely on the fact that functionally active bacterial ribosomes can be reconstituted from their isolated RNA and protein components. These reconstitution techniques have proved to be a powerful tool in both structural and functional studies, since one or more components can be either omitted from the reconstitution mixture or replaced by a suitably modified counterpart. This type of approach can be complemented by genetic studies, since many ribosomal mutants have been isolated in which individual proteins are either altered or even absent.

In 2000, high-resolution structures of the 30S and 50S subunits were solved using x-ray diffraction. Remarkably, the structure of the 50S subunit indicates that the 23S RNA is responsible for peptidyl transferase activity of the ribosome.

Biosynthesis. Ribosomal RNA is produced in the cell by transcription of the appropriate deoxyribonucleic acid (DNA) gene, and the primary transcription

Fig. 1. Model of the *Escherichia coli* ribosome, based on electron microscopic data, showing the two subunits.

product is a large precursor molecule containing all three ribosomal RNA species (although in eukaryotes the 5S RNA is not contained in this molecule). The precursor is subsequently cleaved in several stages to yield the mature ribosomal RNA molecules. In prokaryotes the ribosomal proteins begin to be added while the RNA is still being transcribed; in eukaryotes the RNA is synthesized in the cell nucleolus and the completed precursor is then transported to the cytoplasm. The ribosomal proteins are themselves synthesized via mRNA on ribosomes in the cytoplasm, and in the case of *E. coli* it is known that the majority of the genes coding for ribosomal proteins are grouped in a few large clusters on the genome. Modification of the RNA and proteins (for example, by methylation) occurs at various stages during the assembly process. *See* GENE; GENE ACTION.

Role in protein synthesis. Once the assembly of the ribosomal subunits is complete, they are ready to take part in protein biosynthesis. This process is essentially very similar in both prokaryotes and eukaryotes; what follows is a brief summary of what happens in *E. coli*.

The first step (initiation of protein synthesis) is that an initiator tRNA molecule attached to the amino acid *N*-formyl methionine recognizes its appropriate codon on a mRNA molecule, and binds with the mRNA to the 30S subunit. One end (the 3′ end) of the 16S RNA is also involved in the formation of this complex. A 50S subunit then joins the complex, forming a complete 70S ribosome. A number of proteins (initiation factors, which are not ribosomal proteins) are also involved in the process. At this stage, the initiator aminoacyl tRNA occupies the P-site (peptidyl site) on the ribosome, while a second tRNA binding site (the A-site, or aminoacyl site) is free to accept the next coded aminoacyl tRNA molecule.

In the subsequent steps, the elongation process (**Fig. 3**), aminoacyl tRNA molecules are brought to the A-site as ternary complexes together with guanosine triphosphate (GTP) and a protein factor (elongation factor Tu). Once an aminoacyl tRNA is in the A-site, the initiator amino acid (or at later stages the growing polypeptide chain) is transferred from the P-site tRNA to the A-site aminoacyl tRNA. It is not clear whether this peptidyl transferase activity requires the active participation of a ribosomal component, or whether merely the correct placement of the two tRNA molecules is sufficient for the process to occur spontaneously.

After peptide transfer has taken place, the peptide is attached to the A-site tRNA, and an "empty" tRNA molecule is at the P-site. The peptidyl tRNA complex must now be translocated to the P-site, in order to free the A-site for the next incoming tRNA molecule. Here again protein factors and GTP are involved, and it is probable that the empty tRNA occupies a third ribosomal site (exit or E-site) before finally leaving the complex. Studies show that elongation factor G hydrolyzes GTP to bring about translocation of the tRNA molecules.

The protein chain is completed by the appearance of a "stop" codon in the mRNA. This is recognized by yet another set of protein factors, which causes the completed polypeptide chain to be released from the ribosome. At any one time a number of ribosomes are engaged in the reading of a single mRNA molecule, which leads to the appearance of polyribosomes (polysomes). Each step of the protein biosynthetic process can be disrupted by various antibiotics, and these (or their analogs) can be used for studying different aspects of the ribosomal function, as well as for locating functional sites on the ribosomal subunits by utilization of electron microscopy. *See* MOLECULAR BIOLOGY.

R. Brimacombe; H. G. Wittmann; S. Joseph

Bibliography. R. Green and H. Noller, Ribosomes and translation, *Annu. Rev. Biochem.*, 66:679-716, 1997; W. E. Hill (ed.), *The Ribosome: Structure, Function, and Evolution*, 1990; J. D. Puglisi, S. C. Blanchard, and R. Green, Approaching translation at atomic resolution, *Nature Struct. Bio.*, 7(10): 855–861, 2000; G. Spedding (ed.), *Ribosomes and Protein Synthesis: A Practical Approach*, 1990; H. G. Wittmann, Architecture of prokaryotic ribosomes, *Annu. Rev. Biochem.*, 52:35–65, 1983; R. A. Zimmerman and A. E. Dahlberg, *Ribosome RNA: Structure, Evolution, Processing and Function in Protein Biosynthesis*, 1995.
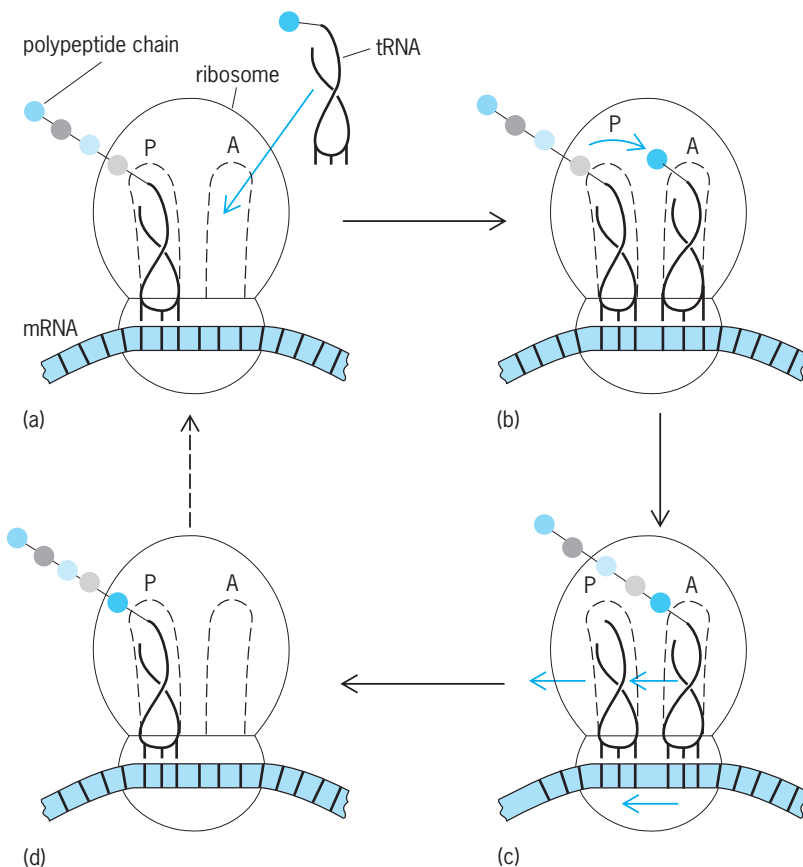

Fig. 3. Simplified diagram of a ribosome synthesizing a protein. It shows the messenger RNA (ribbon), the transfer RNA molecules binding to the A- and P-sites, and the growing polypeptide chain emerging from the large subunit. The steps are: (*a*) Aminoacyl tRNA (as ternary complex) enters A-site. (*b*) Peptide is transferred to the new amino acid at the A-site. (*c*) mRNA and peptidyl tRNA are translocated, with subsequent release of "empty" tRNA. (*d*) Ribosome, with polypeptide one amino acid longer, is ready for the next elongation cycle. Colored arrows indicate movement.

Ribozyme

A ribonucleic acid (RNA) molecule that, like a protein, can catalyze specific biochemical reactions. *See* RIBONUCLEIC ACID (RNA).

Self-splicing rRNA. The first example of RNA molecules with catalytic activity came from observations of RNA processing in the protozoon *Tetrahymena thermophila*. Processing reactions, that is, the biochemical reactions that convert a newly synthesized RNA molecule to its mature form, may result in the removal of terminal sequences of an RNA molecule or in the removal of internal sequences that interrupt the mature RNA segments. The interruptions are called intervening sequences and are removed by an RNA processing reaction called splicing. Splicing involves bond breakage to excise the intervening sequence and bond formation to religate the mature RNA segments. *See* INTRON.

The intervening sequence of *T. thermophila* ribosomal RNA (rRNA) belongs to a large family of functionally related intervening sequences that were isolated from phylogenetically diverse organisms. The self-splicing reaction is inhibited by biochemicals that denature the native RNA structure, thereby providing evidence for a role of a specific folded RNA structure in functional activity. This folded structure is important to the ribozyme's ability to efficiently bind a substrate molecule in the active site, to accelerate the conversion of the substrate molecule to a product molecule by lowering the activation energy for chemical bond cleavage or religation, and to enable release of the reaction product from the ribozyme's active site. Furthermore, specific mutated-nucleotide residues play important roles in binding substrates in the active site, and in stabilizing the folded three-dimensional structure of the ribozyme. Similar mutations within the primary amino acid sequence of a protein enzyme can also destroy its structure and function. Thus, biological catalysts composed of RNA behave very much like protein enzymes.

RNase P. Ribonuclease P (RNase P) is the RNA processing enzyme responsible for the maturation of the $5'$ ends of precursor transfer RNA (tRNA) molecules. The intact enzyme from bacteria and eukaryotes is composed of both RNA and protein. Both moieties are required for function if the enzyme's activity is measured under physiological conditions. The protein component RNase P is thought to facilitate the ability of the ribozyme to fold into the appropriate three-dimensional structure that is necessary for substrate binding and catalysis.

The RNase P RNA molecules from divergent bacteria differ considerably in the nucleotide sequence composition. In contrast to self-splicing RNA molecules, no single nucleotide in the RNase P ribozyme is critical to the function of RNase P. Thus, the RNA processing reaction catalyzed by RNase P appears less dependent on the identity of specific nucleotide residues in the ribozyme than is evident for RNA-catalyzed self-splicing. The notion that radically different ribozyme structures catalyze quite distinct RNA processing reactions is not new: protein enzyme families that are composed of different structures catalyze different types of biochemical reactions as well.

Ribozymes versus enzymes. Ribozymes share many similarities with protein enzymes, as assessed by two parameters that are used to describe a biological catalyst. The Michaelis-Menten constant K_m relates to the affinity that the catalyst has for its substrate. Ribozymes possess K_m values in the micromolar range, which is comparable to K_m values of protein enzymes. The catalytic rate constant describes how efficiently a catalyst converts substrate into product. The values of this constant for ribozymes are markedly lower than those values observed for protein enzymes. Nevertheless, ribozymes accelerate the rate of chemical reaction with specific substrates by 10^{11} compared with the rate observed for the corresponding uncatalyzed, spontaneous reaction. Therefore, ribozymes and protein enzymes are capable of lowering to similar extents the activation energy for chemical reaction. *See* ENZYME.

Primordial catalyst. Before the discovery of ribozymes, biochemists debated which biological molecule could behave as the primordial catalyst. Deoxyribonucleic acid (DNA), which is a stable, information-rich molecule, was previously not believed to encode any functional activity. Proteins are functionally rich molecules, however, they may not have readily lent themselves to being reproduced in the primordial environment because of their diverse amino acid composition. Because RNA molecules can encode both information and catalytic activity, many biochemists now believe that RNA was the likely primordial biological catalyst. In that role, RNA molecules may have catalyzed the elementary biochemical reactions that were essential to the early forms of life in the primordial biosphere. Polypeptides may have supplanted RNA-based enzymes as a greater need for more stable and functionally diverse biological catalysts evolved.

Daniel W. Celander

Bibliography. D. W. Celander and T. R. Cech, Visualizing the higher order folding of a catalytic RNA molecule, *Science*, 251:401–407, 1991; C. Guerrier-Takada et al., The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme, *Cell*, 35:849–857, 1983; D. Herschlag and T. R. Cech, Catalysis of RNA cleavage by the *Tetrahymena thermophila* ribozyme: 1. Kinetic description of the reaction of an RNA substrate complementary to the active site, *Biochemistry*, 29:10159–10171, 1990; K. Kruger et al., Self-splicing RNA: Autoexcision and auto-cyclization of the ribosomal RNA intervening sequence of *Tetrahymena, Cell*, 31:147–157, 1983; F. Michel and E. Westhof, Modelling of the three-dimensional architecture of Group I catalytic introns based on comparative sequence analysis, *J. Mol. Biol.*, 216:585–610, 1990; H. F. Noller, V. Hoffarth, and L. Zimniak, Unusual resistance of peptidyl transferase to protein extraction procedures, *Science*, 256:1416–1419, 1992; D. S. Waugh and N. R. Pace, Gap-scan deletion analysis of *Bacillus subtilis* RNase P RNA, *FASEB J.*, 7:188–195, 1993.

Rice

The plant *Oryza sativa* is the major source of food for nearly one-half of the world's population. In China, Japan, Korea, the Philippines, India, and other countries of Asia, rice is far more important than wheat as a source of carbohydrates. In some countries of the Orient the consumption of rice per capita is estimated at 200–400 lb (90–180 kg) per year. In contrast, the yearly per capita consumption of rice in the United States is only about 8 lb (3.6 kg). The most important rice-producing countries are mainland China, India, and Indonesia, but in many smaller countries rice is the leading food crop. In the United States, rice production is largely concentrated in selected areas of Arkansas, California, Louisiana, and Texas. *See* CARBOHYDRATE; CYPERALES; WHEAT.

Use. Over 95% of the world rice crop is used for human food. Although most rice is boiled, a considerable amount is consumed as breakfast cereals. Rice starch also has many uses. Broken rice is used as a livestock feed and for the production of alcoholic beverages. The bran from polished rice is used for livestock feed; the hulls are used for fuel and cellulose. The straw is used for thatching roofs in the Orient and for making paper, mats, hats, and baskets. Rice straw is also woven into rope and used as cordage for bags. This crop serves a multitude of purposes in countries where agriculture is dependent largely upon rice.

Origin and description. Rice apparently originated more than 7000 years ago in Southeast Asia, in the areas that are now eastern India, Indochina, and southern China. Rice was introduced to North America as early as 1609, and became established in South Carolina about 1690. Until about 1890, rice was grown mainly in the southeastern states. Louisiana became an important rice-producing state in the late nineteenth century. Early in the twentieth century, rice production spread to southeastern Texas, eastern Arkansas, and north-central California. Rice is a comparatively new crop in its present areas of greatest production in the United States.

Rice is unlike many other cereal grains in that all cultivated varieties belong to the same species and have 12 pairs of chromosomes, as do most wild types. The extent of variation in morphological and physiological characteristics within this single species is greater than for any other cereal crop. Although the chromosome number is the same, many of the ancient types have become so widely differentiated that hybrids between them are only partially fertile. *See* GENETICS.

Rice is an annual grass plant varying in height from 2 to 6 ft (0.6 to 1.8 m). Plants tiller, that is, develop new shoots freely, the number depending upon spacing and soil fertility. Among the many types grown, some mature in 80 days; others require over 200 days. The inflorescence is an open panicle (**Fig. 1**). Flowers are perfect and normally self-pollinated, with natural crossing seldom exceeding 3–4%. A distinct characteristic of the flower is the six anthers rather than the customary three of other grasses. Spikelets

Fig. 1. Panicles of (*a*) short-grain and (*b*) long-grain rice. Each spikelet has a single caryopsis enclosed in the lemma and palea.

have a single floret, lemma and palea completely enclosing the caryopsis or fruit, which may be yellow, red, brown, or black. Lemmas may be awnless, partly awned, or fully awned. Threshed rice, which retains its lemma and palea, is called rough rice or paddy. *See* FLOWER; FRUIT; GRASS CROPS; INFLORESCENCE; REPRODUCTION (PLANT).

Varieties. In the United States, only about 25 varieties are in commercial production. Cultivated rices are classified as upland and lowland. Upland types, which can be grown in high-rainfall areas without irrigation, produce relatively low yields. The lowland types, which are grown submerged in water for the greater part of the season, produce higher yields. In contrast to most plants, rice can thrive when submerged because oxygen is transported from the leaves to the roots. All rice in the United States is produced under lowland or flooded conditions. Rice varieties are also classified as long- or short-grain (**Fig. 2**). Most long-grain rices have high amylose content and are dry or fluffy when cooked, while most short-grain rices have lower amylose content
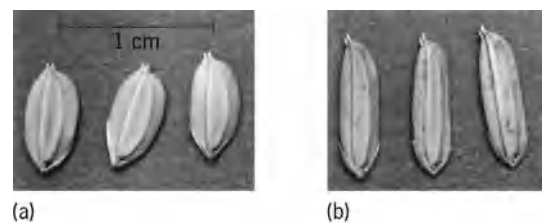

Fig. 2. Rice grains range from (*a*) short to (*b*) long.

and are sticky when cooked. In the United States a third grain length is recognized: medium-grain. The medium-grain rices have cooking qualities similar to those of short-grain varieties.

Plant breeders have been successful in developing high-yielding varieties that combine the vigor and disease resistance of typical tropical rices with the short stature of Taiwan varieties. The "architecture" of these plants ensures their ability to stay erect until harvesting, thus preventing crop loss due to lodging (falling over) into the water.

Cultural practices. Because of the peculiar conditions of growing a crop submerged in water, rice land seldom becomes a part of a regular rotation system. Often two or more successive crops of rice are grown, and the land is then pastured or fallowed to control weeds. Rice is grown on heavy soils underlain by an impervious subsoil to prevent seepage of water. *See* SOIL.

In Asian countries and elsewhere, rice fields are established by transplanting seedlings from beds when the plants are 20–50 days old. Fields to be transplanted are flooded and worked into a soft mud. Clumps of three to four seedlings are pushed into the mud in rows to permit hand cultivation for control of weeds. This system of transplanting seedlings saves irrigation water and permits the field in which they are to be established to grow another crop while the smaller-sized seedling bed is being grown.

In the United States, transplanting is impractical because of high labor costs. In California, almost all rice seed is sown from airplanes directly into fields flooded with 4–6 in. (10–15 cm) of water. In the southern states, rice is seeded either by airplane or by grain drill. In the latter case the rice is seeded into dry soil, which is gradually flooded to a depth of 4–6 in. (10–15 cm) as the rice emerges and grows. In all rice areas of the United States the water is kept at the 4–6 in. (10–15 cm) depth until the land is drained shortly before harvest.

The rice crop in Asian countries and some other countries is harvested by hand (**Fig. 3**). In the United States, rice is harvested with large self-propelled combines similar to those used for wheat and other grain crops. The best stage for combine harvesting is at grain moisture content of 23–28%. Rice dried to a lower moisture content while standing in the field may break up in milling, resulting in lower grain quality. When harvested at this high moisture content, the grain is dried artificially to 14% moisture, care being taken to keep the drying temperature below 110°F (43°C) to avoid damage to the grain. *See* AGRICULTURAL MACHINERY; AGRICULTURAL SCIENCE (PLANT); GRAIN CROPS.   J. N. Rutger

Cultivars and genetics. The genus *Oryza* has unusual diversity and great antiquity because it originated in the Gondwana supercontinent and is now distributed over Africa, Oceania, Asia, and Latin America. The genus consists of about 20 wild species and two cultigens (*O. sativa* of Asia and *O. glaberrima* of western Africa). The enormous diversity in *O. sativa* results from ecogenetic diversification following its wide dispersal by human populations.

At the time of the Green Revolution, which for rice began in the late 1960s, probably more than 100,000 cultivars existed in different parts of the world. These numerous unimproved cultivars (land races) in irrigated areas were displaced by the high-yielding modern varieties, and many disappeared in the absence of continuous cultivation. Even the wild rices are increasingly threatened by development projects. Since 1972, however, massive international efforts in field collection have netted more than 40,000 samples and saved many cultivars from extinction. The base collection maintained at the International Rice Research Institute (IRRI) in the Philippines totals 86,000 accessions. Holdings of national and regional gene banks total over 100,000 but include many duplicates.

Genetic conservation of rice germ plasm has relied almost entirely on preserving seed or plants in gene banks. Most of the major rice-growing countries have established refrigerated storage facilities, but genetic erosion inside gene banks in the humid tropics continues when cold, dry storage conditions are not maintained or when seeds are not frequently regenerated. The IRRI collection is backed up by a duplicate storage site at the U.S. National Seed Storage Laboratory in Fort Collins, Colorado. Conservation in natural habitats has been fostered, but it is difficult to implement.

In spite of some setbacks due to pest epidemics, rice germ plasm is the most intensively evaluated and used crop collection in the world. The potential genetic vulnerability in *O. sativa* is considerably higher than that of other major cereals: About 1000 improved varieties bred since the 1960s and grown on nearly 60% of the world's rice lands largely share the same Chinese semidwarfing gene (sd_1); many tropical semidwarfs have a common cytoplasm; and more than 90% of the 3.2×10^7 acres (1.3×10^7 hectares) of hybrid rice grown in China have a common

Fig. 3. Hand harvesting is common in many countries.

male-sterile cytoplasm. Moreover, farmers in tropical Asia tend to use the same modern variety twice in a year in staggered plantings, adding to pest incidences and quick breakdown of genetic resistance. Although no major disaster has occurred, farmers using the above practices and heavy fertilization have suffered substantial losses in a cyclic manner.   Te-Tzu Chang

Disease. A comprehensive disease control program is used in rice culture to ensure against yield and quality losses in fields and deterioration of quality during drying and storage. Key elements in this program are preventive measures such as crop rotation and sanitation to suppress carryover of pathogens from crop to crop in the soil or crop residues, use of good-quality seeds treated with an effective seed protectant fungicide, selection of adapted varieties with greatest degree of resistance to potentially epidemic diseases of the area, weed

Fig. 4. Rice blast disease.

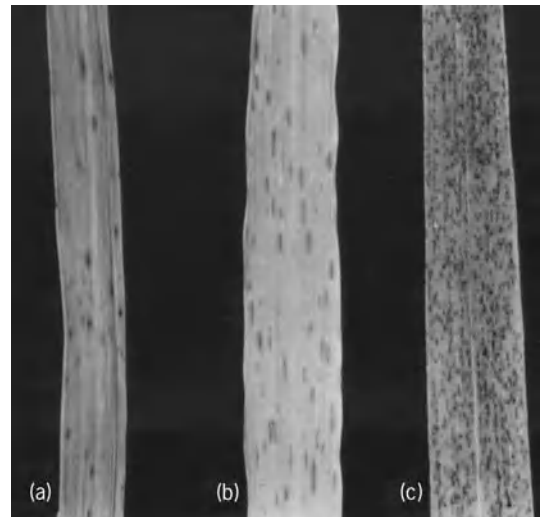Fig. 5. Brown bordered sheath spot disease.

Fig. 6. Fungal diseases of rice. (*a*) Brown spot. (*b*) Narrow brown leaf spot. (*c*) Leaf smut.

control, timely water management, and fertilization practices that consider both proper balance and timely application. In some disease situations, applications of foliar fungicides may be used to protect against buildup of disease to economically important levels. *See* FUNGISTAT AND FUNGICIDE.

Control measures. The variety of measures required for disease control is necessitated by the diversity of pathogens affecting the crop. Pathogens may be classed as fungi, bacteria, nematodes, viruses, or physiologic agents. There are numerous unique disease control strategies within each class. It is essential, therefore, to make correct identification to select the most appropriate control measures.

It is not possible to predict with much certainty what diseases are likely to be most prevalent in a given season, and their intensity may vary in different areas or years. However, certain conditions are known to favor certain diseases.

Fungi. Rice blast (**Fig. 4**), one of the most widely distributed and serious rice diseases, is caused by the airborne fungus *Pyricularia oryzae* and is favored by high levels of native or applied nitrogen such as are found in newly cleared land or following fish culture. In the seedling stage the disease is also favored by delayed flooding. Host plant resistance is a major means of control. Two culm rotting diseases, bordered sheath spot (**Fig. 5**) and stem rot, are caused by the soil-invading fungi *Rhizoctonia solani* and *Magnaporthe salvanii*, respectively, and are favored by continuous cropping of rice and very thick stands that hold moisture on the vegetation for several days near heading time. Stem rot is also favored by soils deficient in potassium. Rotation, proper fertilization, and proper seeding rates are the major means of controlling these two diseases. Brown leaf spot, caused by *Helminthosporium oryzae* (**Fig. 6**), is another serious and widespread airborne fungal disease; it is most serious on upland rice or where plants are stressed by cold or inadequate water, low nitrogen levels, or damaging herbicide applications.

Physiologic disorder. Straighthead (**Fig. 7**), a physiologic disorder, is likely to occur in low, poorly drained

Fig. 7. Straighthead, a physiologic disorder of rice.

Fig. 8. White tip disease of rice.

spots in the field or on highly fertile, newly cleared land not properly drained.

Nematodes. White tip disease (**Fig. 8**) is caused by a seed-borne nematode that overseasons almost exclusively in the rice kernel; it builds up to damaging levels if seeds are used from successive crops, season after season. It can be controlled by a seed certification program that excludes this pathogen.

Bacteria. Bacterial blight, caused by *Xanthomonas oryzae*, is most serious in low, wet, poorly drained locations that are easily flooded during the seedling stages and are often subject to mist.

Virus. Viral diseases are mostly transmitted by leaf hoppers and plant hoppers and thus build up in areas where two or three successive crops are grown each year.

Identification. Proper identification of diseases and the circumstances that favor each of them permits selection of varieties, cultural practices, or other control measures to prevent their occurrence at economic levels. *See* PLANT PATHOLOGY.

George E. Templeton

Processing. The rice kernel has four primary components: the hull or husk, the seedcoat or bran, the embryo or germ, and the endosperm. The main objective of milling rice is to remove the indigestible hull and additional portions of bran to yield whole unbroken endosperm. Rice milling involves relatively uncomplicated abrasive and separatory procedures which provide a variety of products dependent on the degree of bran removal or the extent of endosperm breakage.

Rice milling. For rice milling (**Fig. 9**), rough rice, or paddy, which is dried grain with the husk intact and with no adhering stalk, is cleaned by selective screening, disk separators, and air aspiration systems to remove dirt, stones, chaff, and other extraneous material. Clean paddy may be converted by parboiling and drying prior to milling.

Hull removal. The coarse husk of paddy is loosened from the intact kernel by either disk or rubber
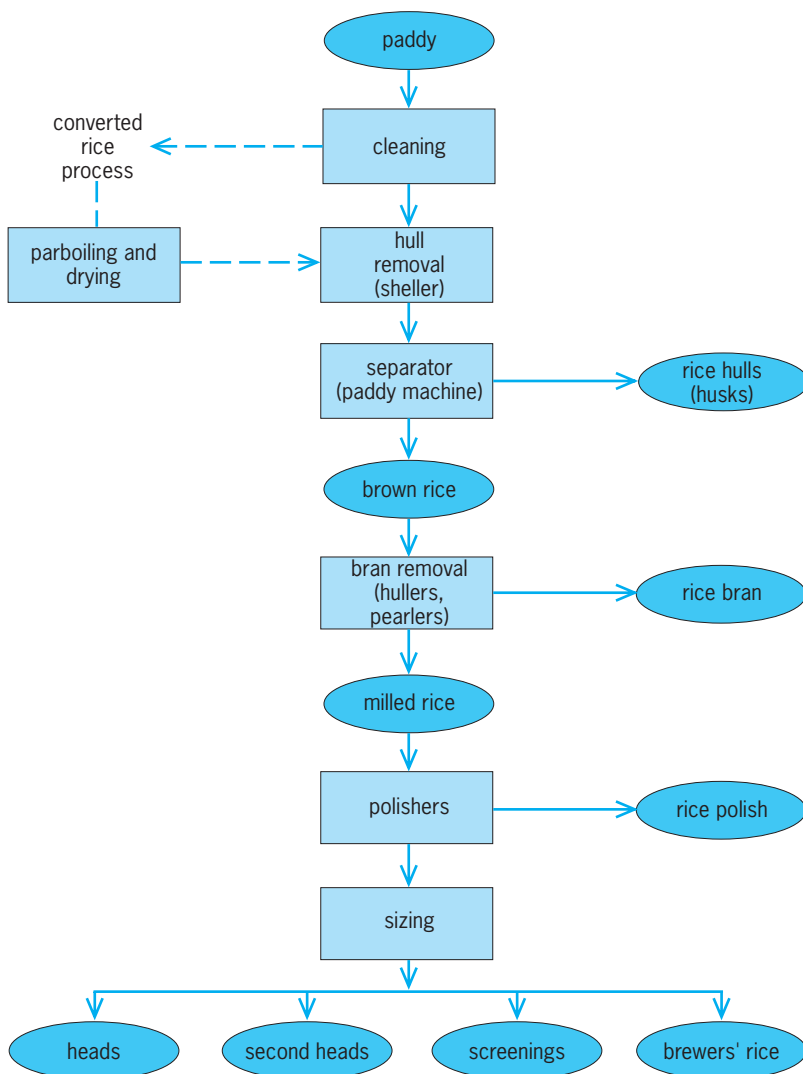

Fig. 9. General rice milling operations and products.

roll shellers. Disk shellers comprise two large horizontal emery-faced abrasive plates (disks) positioned with parallel surfaces. The upper disk is stationary and possesses a center hole into which a stream of paddy enters, while the lower disk is solid and can be rotated at designated speeds. The distance between these disks is adjusted to a width greater than that of the rice kernel length. Paddy passes from the center of the sheller to its periphery by a tumbling centrifugal action. This abrasive action removes the husk from about 80% of the kernels in a single pass. Disk shellers are adjusted to minimize kernel damage; however, they are being replaced increasingly by more efficient rubber roll shellers. Rubber-surfaced rolls are positioned in parallel at a specified gap and rotate toward each other at differential speeds. Rough rice fed between the rolls results in a gentle removal of husk by tearing action. This method yields greater quantities of hulled and undamaged kernels than do disk shellers.

Following the sheller operation, the low-density freed hulls are lifted and separated by air aspiration. Rice hulls comprise about 17–20% of the paddy and contain about 20% ash, primarily silica. Hulls have been used as boiler feed stock and as a press aid in fruit juice extraction.

The remaining paddy and dehulled kernels are conveyed to a separator which is commonly referred to as a paddy machine. This unit separates the less dense paddy from dehulled kernels on a vibrating, inclined gravity platform in which paddy migrates upward through a series of vertical baffle plates while the dehulled kernels migrate down the platform. Segregated paddy is returned to additional shellers. The dehulled kernel with seedcoat intact is termed brown rice.

Bran removal. Brown rice is further milled to remove the bran and thus to produce a whitened product by gentle and progressive abrasion in a series of machines referred to as hullers or pearlers. Actually the term huller is misleading, as the hull has already been removed. Hullers are constructed with an abrasive roller cylinder rotating inside a stationary cylindrical sieve. As this cylinder rotates, brown rice passes between the abrasive surface and the screen and is scoured to remove the outer bran layer. The fine bran particles are removed through the sieve. The adjustment, sizing, speed, and rice feed rates are critically controlled to minimize kernel breakage. Typically successive movement through two or three hullers, each adjusted to an increasingly finer setting, is desired. Pearlers are also used to remove bran by passing kernels between a cone-shaped rotor and a screen. Abrasive action is obtained among kernels and bran separated by screening. Heat generated during bran removal must be dissipated by air currents to avoid damage to rice cooking characteristics. Further whitening of rice may be accomplished by brushing or polishing to remove the inner white-bran, aleurone layer, and small portions of endosperm. This procedure develops light-colored fines termed rice polish which are frequently used as an ingredient in formulated baby foods.

Sizing of milled rice. Milled kernels are separated by size by using a variety of systems. Finest fragments are removed by reel screening, and remaining broken kernels by disk separation. Brewers' rice is the smallest kernel fragment screened and is used as fermentation feed stock. Medium-sized pieces (less than one-third kernel) are termed screenings, second heads include large pieces up to three-fourths of a kernel, while the whole unbroken kernels are classified as heads. Polished rice may be surface-coated with glucose by tumbling to improve its gloss, sheen, and uniformity.

Parboiled or converted rice. The soaking and cooking of paddy prior to milling has been practiced in many cultures. The greatest concentration of water-soluble B-complex vitamins (thiamine, riboflavin, and niacin) is found in the bran layers of the kernel, which are lost during dry-milling operations. During parboiling, the soluble vitamins diffuse into the endosperm and greatly enhance the nutritive value of the milled rice. Soaked and boiled paddy is dried prior to conventional milling. Parboiled rice starch is gelatinized, and the kernel acquires a darker, creamier color upon milling. The endosperm of parboiled rice is firmer and denser, and requires additional cooking time compared to regular rice; however, since the surface starch is removed, the cooked product resists sticking and clumping. *See* NIACIN; RIBOFLAVIN; THIAMINE.

Many industrial systems have been developed to accomplish the parboiling of paddy. Variations include long-time steeping, heating in live steam, and the use of vacuum infusion. The converted rice produced in the United States is made from cleaned paddy under an exacting process. Paddy is placed in large chambers and subjected to high vacuum to evacuate air from the kernel tissue. Hot water is directly percolated into the evacuated tank, and external pressure is applied to infuse the soluble nutrients of the hull and bran into the endosperm. Following sufficient hot water treatment, the steep water is drained and the rice is subjected to live steam. Drying of the parboiled paddy may include vacuum or conventional air treatment and a tempering period to assure uniformity of kernel moisture, which is maintained at about 12%.

Prepared rice products. Instant rice is made from whole grain rice by pretreating under controlled cooking, cooling, and drying conditions to impart the quick-cooking characteristic. Typically, instant rices require less than 5 min preparation time, as compared to 20–25 min for regular or parboiled rice products. A variety of processing schedules, including use of soaking, heating, cooking, vacuum or steam pressure, and controlled dehydration, are employed to render milled rice suitable for rapid preparation. Use of these techniques enables hydration and gelatinization of starch and yields an open, porous kernel surface with minimum kernel damage. This precooked product readily rehydrates and softens to a palatable grain in boiling water.

Ready-to-eat breakfast rice cereals are prepared

from milled rice as flakes or puffs. Rice is frequently precooked under steam pressure, conditioned to uniform kernel moisture, and passed through high-pressure, smooth flaking rolls, and toasted. Vacuum puffing of cooked grains or flaked rice is common. Instant hot baby cereal is produced by drum-drying a slurry of rice flour. Thin sheets of cooked dehydrated cereal are removed by surface-scraping, rotating, steam-heated drums. The sheets are ground to yield thin flakes which readily hydrate and form a characteristic, soft, pasty porridge. *See* CEREAL; FOOD MANUFACTURING.

Rice bran oil. This product was developed as a result of increased extraction of lipids from rice bran. Edible oils obtained from rice bran must be treated to inactivate lipase enzymes that will hydrolyze the lipids to form free fatty acids and rancid odors during storage. The crude lipid content of rice bran is about 10% by weight, and has been commercially extracted in several Asian countries, including India, China, and Japan. Traditionally, rice bran oil was obtained through hydraulic pressing, which provided a relatively low yield (10–12%). This procedure has been replaced by solvent extraction, which yields 16–18%. Bran is typically preheated by using steam pressure to sufficiently inactivate hydrolytic lipases prior to hexane extraction. The crude rice bran oil possesses a dark greenish-brown to light yellow pigmentation, which is removed during subsequent refining processes. Refined rice bran oil possesses a low content of linolenic acid (highly unsaturated) and a high content of tocopherols (having antioxidant properties). Both constituents contribute to increased oxidative stability. These properties enable rice bran oil to be utilized as an edible-grade oil in a variety of applications as well as an industrial feedstock for soap and resin manufacture. Improved utilization of rice bran oil has been achieved through improved extraction, refining, and formulation in food products; thus, this by-product utilization has the potential for increasing the value added to rice as a food and food ingredient. *See* FAT AND OIL (FOOD); SOLVENT EXTRACTION.

Mark A. Uebersax

Bibliography. T. T. Chang, Conservation of rice genetic resources: Luxury or necessity?, *Science*, 224:251–256, 1984; D. G. Dalrymple, *Development and Spread of High-Yielding Rice Varieties in Developing Countries*, 1986; S. K. DeDatta, *Principles and Practices of Rice Production*, 1987; H. C. Dethloff, *A History of the American Rice Industry, 1685–1985*, 1988; S. Gangopadhyay and S. Y. Padmanabhan, *Breeding for Resistance in Rice*, 1987; T. H. Hargrove et al., Twenty years of rice breeding, *BioScience*, 38:675–681, 1988; B. O. Juliano (ed.), *Rice Chemistry and Technology*, 1985; W. E. Marshall and J. I. Wadsworth, *Rice Science and Technology*, 1993; A. Prakash et al., *Rice Storage and Insect Pest Management*, 1987; B. M. Shepard et al., *Helpful Insects, Spiders, and Pathogens: Friends of the Rice Farmer*, 1987; S. Tsunoda and N. Takahashi (eds.), *Biology of Rice*, 1984; H. Waibel, *The Economics of Integrated Pest Control in Irrigated Rice*, 1986.

Ricinulei

An order of extremely rare arachnids, also known as the Podogona, with a body less than 1 in. (25 mm) in length. Superficially, they resemble ticks in general appearance and movement, and are found only in tropical Africa and in the Americas, from the Amazon to Texas. The two anterior pairs of appendages are chelate. The terminal segments of the third legs of the male are modified as copulatory structures. Less than 25 modern species are known. The occurrence of several fossils from Carboniferous time suggests that the group was formerly more common. Virtually nothing is known about the reproduction, growth, and ecology of the Ricinulei. *See* ARACHNIDA.

C. Clayton Hoff

Rickettsioses

Often severe infectious diseases caused by several diverse and specialized bacteria, the rickettsiae and rickettsialike organisms. The best-known rickettsial diseases infect humans and are usually transmitted by parasitic arthropod vectors (see **table**).

Agents. Rickettsiae and rickettsialike organisms are some of the smallest microorganisms visible under a light microscope. Although originally confused with viruses, in part because of their small size and requirements for intracellular replication, rickettsiae and rickettsialike organisms are characterized by basic bacterial (gram-negative) morphologic features. Their key metabolic enzymes are variations of typical bacterial enzymes. These organisms apparently exploit the normally rich pool of host intracellular nutrients for their own, perhaps abbreviated, synthetic pathways. This, at least partial, dependence on host-cell molecules helps explain why rickettsiae (with two exceptions) replicate only in living eukaryotic cells.

The genetic material of rickettsiae and rickettsialike organisms likewise seems to conform to basic bacterial patterns. The genome of all rickettsialike organisms consists of double-stranded deoxyribonucleic acid (DNA). In addition to the genomic DNA, *Coxiella burnetii*, the causative agent of Q fever, may also have extrachromosomal, circular DNA plasmids.

Rickettsiae enter host cells by phagocytosis and reproduce by simple binary fission. The site of growth and reproduction varies among the various genera. Rickettsiae that cause spotted fever appear to move easily through the membranes of infected cells and can often be seen in infected cell nuclei as well as in the cytoplasm. Other rickettsiae, for example, the typhus-group rickettsiae, do not appear able to pass easily out of the host cell through the cytoplasmic membranes. Instead of being budded and released to the extracellular environment, they may pack the host cytoplasm until the host cell is destroyed. Members of the rickettsialike genus *Ehrlichia* replicate within cytoplasmic vacuoles instead of replicating directly within the cytoplasm. Similarly, *C. burnetii* replicates only within the phagolysosomes of the

Causes and modes of transmission for some common rickettsial diseases		
Disease	Causative agent	Mode of transmission
Epidemic typhus	*Rickettsia prowazekii*	Louse vector
Murine typhus	*Rickettsia typhi*	Flea vector
Trench fever	*Rochalimaea quintana*	Louse vector
Rocky Mountain spotted fever	*Rickettsia rickettsii*	Tick vector
Q fever	*Coxiella burnetii*	Tick vector, airborne transmission
Scrub typhus	*Rickettsia tsutsugamushi*	Mite vector

host cells under acidic conditions that otherwise would be considered quite hostile to microbial growth.

Although once grouped within the genus *Rickettsia*, the causative agent of trench fever and one other related organism isolated only from rodents have been grown on artificial media. Thus, they have been assigned to the genus *Rochalimaea*, despite apparently significant DNA homologies with typhus-group rickettsiae and very similar epidemiologic characteristics.

How other rickettsiae are related is less clearly defined. *Rickettsia tsutsugamushi*, the causative agent of scrub typhus, replicates in the cytoplasm of infected cells, is transmitted by arthropod vectors, and is included in the *Rickettsia* genus. However, *R. tsutsugamushi* shares no significant common protein antigens with the spotted fever or typhus-group rickettsial agents that would indicate close evolutionary histories; several other features of *R. tsutsugamushi* are equally distinct.

Human disease. Clinically, the rickettsial diseases of humans are most commonly characterized by fever, headache, and some form of cutaneous eruption, often including diffuse rash, as in epidemic and murine typhus and Rocky Mountain spotted fever, or a primary ulcer or eschar at the site of vector attachment, as in Mediterranean spotted fever and scrub typhus. Signs of disease may vary significantly between individual cases of rickettsial disease. Q fever is clinically exceptional in several respects, including the frequent absence of skin lesions.

All of the human rickettsial diseases, if diagnosed early enough in the infection, can usually be effectively treated with the appropriate antibiotics. Tetracycline and chloramphenicol are among the most effective antibiotics used; they halt the progression of the disease activity, but do so without actually killing the rickettsial organisms. Presumably, the immune system, even that of patients who have been treated with antibiotics, is ultimately responsible for ridding the body of infectious organisms. Penicillin and related compounds are not considered effective. *See* ANTIBIOTIC.

Most rickettsial diseases are maintained in nature as diseases of nonhuman vertebrate animals and their parasites. Although human disease may be extremely serious, human infection may usually be regarded as peripheral to the normal natural infection cycles, and human-to-human transmission is not the rule. However, two noticeable exceptions to the zoonotic pattern of rickettsial disease can occur:

the organism responsible for epidemic typhus (*Rickettsia prowazekii*) and the agent responsible for trench fever (*Rochalimaea quintana*) have the potential to spread rapidly within louse-ridden human populations. *See* ZOONOSES.

Historically, epidemic typhus has been an important factor in shaping Western civilization. Throughout most of European recorded history, epidemic typhus has been associated with famine and war and has often played a dominant role in determining the outcome of military campaigns. Epidemic typhus is transmitted between persons by the human body louse. Poor sanitation and inadequate living conditions create an ideal environment for human louse infestations, and hence for potential human epidemic typhus transmission. One estimate suggests that approximately 25 million persons suffered from epidemic typhus infections in Russia alone from the beginning of World War I through the Bolshevik Revolution; the mortality rate prior to the advent of appropriate antibiotics is estimated to have ranged from 10 to 66%. With the advent of higher standards of personal hygiene, the use of insecticides to control lice, and the recognition of the cycle of human epidemic typhus infections, the dominance of this disease over human life and political affairs has largely receded. However, human epidemic typhus still persists in some isolated human populations.

Even in louse-free human environments, recrudescent human typhus (Brill-Zinsser disease) may occur in individuals who have experienced a primary infection many years in the past. This disease complicates attempts at typhus eradication. In addition, an organism essentially identical to the causative agent of human epidemic typhus (*Rickettsia prowazekii*) exists in an almost totally separate, nonhuman cycle among North American flying squirrels. Sporadic human disease, usually linked to close human association with flying squirrels, occurs every year, and the disease syndrome is compatible with classic human epidemic typhus. Like most Brill-Zinsser disease in developed countries, secondary spread of the flying squirrel-associated disease between humans has not been documented; however, in the absence of an effective human epidemic typhus vector (the body louse), human-to-human transmission would not be expected.

The causative agent of another human disease, murine typhus (*Rickettsia typhi*), is closely related to the agent responsible for epidemic typhus. Murine typhus persists as a worldwide disease of rats and is sporadically vectored from rats to humans by fleas.

Human-to-human spread of the disease does not occur.

During World War I and, to a lesser extent, World War II, trench fever (caused by *Rochalimaea quintana*), like epidemic typhus, was a serious problem among military troops. Large numbers of troops were affected, although the clinical signs of the disease were relatively mild. Trench fever is infrequently diagnosed today.

The spotted fever group of rickettsiae represents an interesting collection of closely related organisms. All known spotted fever group organisms are transmitted by ticks. Despite a global distribution in the form of various diseases, nearly all spotted fever group organisms share close genetic, antigenic, and certain pathologic features. Examples of human diseases include Rocky Mountain spotted fever (in North and South America), fièvre boutonneuse or Mediterranean spotted fever (southern Europe), South African tick-bite fever (Africa), Indian tick typhus (Indian subcontinent), and Siberian tick typhus (northeastern Europe and northern Asia). If appropriate antibiotics are not administered, Rocky Mountain spotted fever, for example, is a life-threatening disease. As many as 1000 cases of Rocky Mountain spotted fever are diagnosed each year in the United States. Several other well-documented and rather common spotted fever group rickettsiae appear not to be pathogens of humans, although they can often be found infecting the same tick species that may transmit virulent human spotted fever disease.

Coxiella burnetii, the agent responsible for Q fever, commonly infects certain domestic animals, especially cattle, sheep, and goats, although disease is often inapparent. The organism is present only at low levels in these animals, except during parturition. This organism can also often be found in the mammary glands of these animals and is frequently shed in milk as well as in urine and feces. The organism is especially resistant to inactivation and can form highly infective aerosols. Thus, Q fever is a common disease of slaughter-house workers and others in close contact with the animals and animal products in which the organism is present. The acute stage of the disease in humans is usually self-limiting; however, if chronic endocarditis develops as a result of Q fever infection, the disease is often fatal.

Scrub typhus disease occurs throughout Asia, Australia, and the Pacific islands. The natural reservoirs of scrub typhus are native rats, and the disease is transmitted by chiggers. Considerable variation occurs among the protein antigens of *Rickettsia tsutsugamushi*, the causative agent; unlike other rickettsiae, this species appears to consist of a constellation of related disease-causing organisms.

Human disease associated with infection by an organism identical or very similar to *Ehrlichia canis*, a disease agent first recognized in dogs, has been recognized in the United States. The disease is believed to be transmitted by ticks, and the basic ecology of the disease may be similar to that of the spotted fever organisms. *See* INFECTIOUS DISEASE.

Russell L. Regnery

Bibliography. E. H. Lennette, P. Halonen, and F. A. Murphy (eds.), *Laboratory Diagnosis of Infectious Diseases: Principles and Practice*, vol. 2: *Viral, Rickettsial, and Chlamydial Diseases*, 1988; K. Maeda et al., Human infection with *Ehrlichia canis*, a leukocytic rickettsia, *N. Engl. J. Med.*, 316(14):853–856, 1987; D. H. Walker (ed.), *Biology of Rickettsial Diseases*, vols. 1 and 2, 1988.

Riemann surface

A generalization of the complex plane that was originally conceived to make sense of mathematical expressions such as $\sqrt{z}$ or log z. These expressions cannot be made single-valued and analytic in the punctured plane $\mathbf{C}\backslash\{0\}$ (that is, the complex plane with the point 0 removed). The difficulty is that for some closed paths the value of the expression when reaching the end of the path is not the same as it is at the beginning. For example, the closed path can be chosen to be the unit circle centered at $z = 0$ and followed counterclockwise from $z = 1$. If $\sqrt{z}$ is assigned the value $+1$ at $z = 1$, its value at the end of the circuit is -1. Similarly, if log z is assigned the value 0 at $z = 1$, at the end of the circuit, allowing the values to change continuously, the value is $2\pi i$. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

Construction. The construction of an abstract surface for $\sqrt{z}$ on which $\sqrt{z}$ has a single value at each point and on which the values of $\sqrt{z}$ vary continuously around each closed path, always ending with the starting value, proceeds as follows. From the complex plane $\mathbf{C}$ the negative half of the real axis removed. The slit $\{z = x + iy : y = 0, 0 \leq x\}$ is thought of as having two edges: an upper or $+$ edge and a lower or $-$ edge ($z = 0$ is not counted). Another copy of this slit plane is then placed over, and parallel to, the original, as the first and second floors in a building. Next, the two sheets are connected by attaching the $-$ edge of the bottom sheet to the $+$ edge of the top sheet, and the $+$ edge of the bottom sheet to the $-$ edge of the top sheet. The abstract surface thus constructed is not embedded in euclidean 3-space because the attachments cannot be done in 3-space without creating self-intersections. The expression $\sqrt{z}$ is defined at the point over z on the upper sheet to have positive imaginary part and on the lower to have negative part. If a point is imagined traveling counterclockwise around the unit circle centered at the origin in the bottom sheet, it hits the negative edge of the slit, ascends to the top sheet, leaves from the positive edge of the top slit, travels counterclockwise around the circle in the upper sheet, and finally returns to the initial point after descending from the negative edge of the top slit to the positive edge of the bottom. The values of $\sqrt{z}$ change continuously around this path and return to the initial value.

The abstract surface constructed is called the Riemann surface for $\sqrt{z}$. It is two-sheeted over the z plane, except that only one point lies over $z = 0$, which is therefore called a branch point. Actually the surface lies not just over $\mathbf{C}$ but over the extended
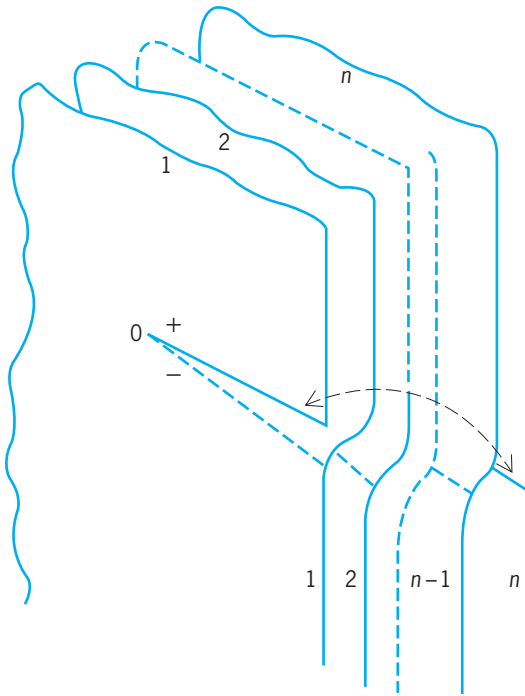
Fig. 1. Riemann surface for $\sqrt[n]{z}$. (**After A. G. Sveshnikov and A. N. Tikhonov, The Theory of Functions of a Complex Variable, Mir Publishers, 1973**)

plane $\mathbf{C} \cup \{\infty\}$ (or the Riemann sphere), formed by attaching the so-called point at infinity (∞) to the complex plane. There is also a branch point of order 2 over $z = \infty$, and the surface constructed is compact. The Riemann surface for $\sqrt[n]{z}$, which has n branches instead of 2, is shown in **Fig. 1**. *See* TOPOLOGY.

Likewise a Riemann surface can be reconstructed for expression (1). It is compact of genus g. A surface

$$[z(z - a_1)\dots(z - a_{2g})]^{1/2} \qquad (1)$$

$$a_i \neq a_j \neq 0 \text{ for } i \neq j$$

can also be constructed for $\log z$. It is a doubly infinite spiral over $\mathbf{C}/\{0\}$, and is not compact.

Construction for algebraic curves. More generally, there is a compact Riemann surface R_w corresponding to any irreducible algebraic curve defined by Eq. (2),

$$w^n + p_1(z)w^{n-1} + \cdots + p_n(z) = 0 \qquad (2)$$

where p_i (z), $1 \leq i \leq n$, are polynomials, $n > 2$. On R_w, $w = w(z)$ becomes a single-valued meromorphic function. Except for a finite number of points of the extended plane $\mathbf{C} \cup \{\infty\}$, for each $z = z_0$ there are n distinct values of w, $w_1(z_0), \dots, w_n(z_0)$, and so R_w has n sheets over the extended plane. These can be fastened together according to the nature of the curve to form a single surface. The global function $w(z)$ has the values $w_1(z_0), \dots, w_n(z_0)$ at the n points over z_0. The construction for algebraic curves was first carried out by B. Riemann. *See* ALGEBRAIC GEOMETRY.

Construction for power series. Alternatively, it is possible to start with a convergent power series about,

for example, $z = 1$, given by Eq. (3) [for example,

$$w = \sum_{n=0}^{\infty} a_n(z - 1)^n \qquad (3)$$

a local solution of an algebraic equation as above]. The power series together with all its analytic continuations determines a Riemann surface R_w which in general is not compact but on which $w = w(z)$ becomes a single-valued meromorphic function.

Topological properties. In the early 1920s, H. Weyl defined a Riemann surface in the context of the newly developing field of topology. This definition does not depend on having a particular function in hand, and does not require the use of such seemingly artificial devices as slits.

A Riemann surface is a two-dimensional connected topological manifold (surface) with a complex structure. This means there is a covering of the surface with open neighborhoods $\{U_i\}$ with the following properties: (1) Each neighborhood U_i corresponds (is homeomorphic to) to a neighborhood V_i in $\mathbf{C}$; this correspondence can be denoted as $z_i: U_i \to V_i$. (2) If two neighborhoods, U_i and U_j, overlap, then the composition $z_j \circ z^{-1}{}_i$, which is defined on $z_i(U_i \cap U_j)$ in $\mathbf{C}$, is a conformal map onto $z_j(U_i \cap U_j)$. *See* MANIFOLD (MATHEMATICS).

For the surface for $\sqrt{z}$ constructed above, the correspondence z_i is given by the projection to the complex plane except in neighborhoods of the branch points over 0 and ∞. About those points the correspondence is described by $\sqrt{z}$ or $\sqrt{z^{-1}}$, which determines a one-to-one mapping of neighborhoods onto neighborhoods of the origin in $\mathbf{C}$. Every Riemann surface has meromorphic functions and therefore is associated with power series.

A Riemann surface can be thought of as a topological surface together with a rule for measuring angles at points. The angles about a point must sum to 2π. An angle-preserving orientation-preserving mapping between Riemann surfaces is called conformal. In the theory, no distinction is made between conformally equivalent surfaces. *See* CONFORMAL MAPPING.

Any smooth surface embedded in euclidean space can be made into a Riemann surface. Angles between intersecting smooth curves on the surface can be measured simply by using the euclidean metric of the surrounding space to measure the angle of intersection of their tangent lines.

Conversely, it is known that any Riemann surface, compact or noncompact, can be conformally mapped onto a surface embedded in euclidean 3-space as a closed subset. In fact, each Riemann surface can be so embedded in infinitely many ways.

Likewise, a polyhedral surface can be made into a Riemann surface. On the flat pieces the angles are as given. But before measuring angles at a vertex, the surface must be flattened out about the vertex, for the angles about each point must add to 2π. This is done by applying a map of the form z^α, where $\alpha = 2\pi/\theta$, θ being the sum of the angles at the vertex.

A Riemann surface that is compact without boundary can be conformally mapped onto the

Riemann surface associated with some algebraic curve.

A 2-dimensional Riemannian manifold can be made into a Riemann surface. Local conformal mappings into the plane are determined by solving the Beltrami equation; these determine the conformal structure. *See* RIEMANNIAN GEOMETRY.

The most familiar Riemann surfaces are the regions of the plane or the sphere (the Riemann sphere), such as the 3-punctured sphere, an annulus, and the complement of the Cantor set. Via steographic projection, the complex plane is mapped onto the unit sphere so that the point at infinity becomes the north pole. Thus, the complex plane is alternately described as the once-punctured sphere.

Uniformization. Perhaps the most important single theorem of complex analysis is the uniformization theorem. It says that a simply connected Riemann surface can be mapped conformally onto exactly one of: the sphere, the complex plane, or the unit disk $\{z : |z| < 1\}$.

If M is a manifold, then, corresponding to each subgroup H of the fundamental group of M, there is another manifold called a covering manifold of M. If H is the trivial group, then the covering is simply connected and is called the universal covering. This is a standard construction in topology, but its application to Riemann surfaces is important because the covering surface of a Riemann surface is also a Riemann surface. In particular, uniformization may be applied to the universal covering surface of a Riemann surface. The universal cover of the Riemann surface R is the unit disk except in the cases that R can be mapped conformally onto the sphere itself (spherical case), or the complex plane, the 1-punctured plane, or a torus (euclidean cases). As a consequence, every Riemann surface has an intrinsic geometry: spherical, euclidean, or hyperbolic. *See* GROUP THEORY.

Almost all Riemann surfaces fall into the third category: their universal covering surfaces can be mapped conformally onto the unit disk, and they have an intrinsic hyperbolic geometry coming from the hyperbolic geometry of the unit disk in the hyperbolic metric given by Eq. (4), which has constant

$$ds = \frac{2|dz|}{1 - |z|^2} \tag{4}$$

gaussian curvature -1. *See* DIFFERENTIAL GEOMETRY; GEOMETRY; NONEUCLIDEAN GEOMETRY.

Put another way, each such Riemann surface R can be "unrolled" onto the unit disk. In the unit disk, R can be modeled as a polygon in the hyperbolic geometry. The sides of the polygon are arranged in pairs. A pair of sides are identified with each other by an orientation-preserving isometry that maps one side onto the other. When the polygon is "rolled up" by identifying each pair of sides, a Riemann surface is obtained which is conformally equivalent to R. **Figure 2** shows an ideal hyperbolic square (it has hyperbolic area 2π) and the first few generations of the tiling of the hyperbolic plane by isometric copies
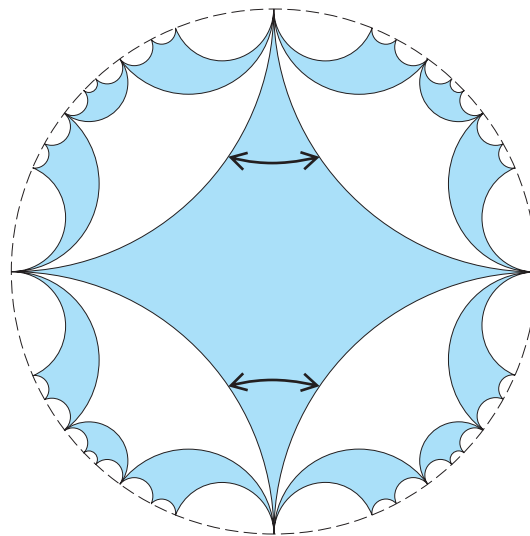

Fig. 2. Rolling up an ideal hyperbolic square. (*After W. Dyck, Ueber Untersuchen und Aufstellung von Gruppe und Irrationalitaet regulaerer Riemannscher Flaechen, Math. Ann., 17:473–509, 1880*)

of itself; "ideal" means its vertices are "at infinity" in terms of the metric and its vertex angles there are zero. (Figure 2 is an approximation; there are really infinitely many white pieces along the edge of the disk, getting smaller and smaller.) When the pairs of sides indicated by double-headed arrows in Fig. 2 are identified, the square "rolls up" to become the 3-punctured sphere. (In the other direction, one can take a 3-punctured sphere and indicate how a tile is placed within it: Take one of the punctures and draw arcs from that puncture to each of the two others. This procedure cuts the 3-punctured sphere into a 4-sided figure that represents the original tile. This operation is analogous to making a torus from a sheet of paper by first rolling the sheet up to get a cylinder and then pasting the two ends of the cylinder together to get a torus.)

Moduli. The conformal equivalence classes of all compact Riemann surfaces of fixed genus $g \geq 0$ with $b \geq 0$ boundary components and $n \geq 0$ punctures such that $d = 6g + 3b + 2n - 6 \geq 1$ can be parametrized: The collection forms a simply connected space $M_{g,b,n}$ with singularities (an orbifold), called moduli space, of real dimension d. Distinct points correspond to surfaces which cannot be mapped conformally onto each other. If a point is imagined moving through $M_{g,b,n}$, the corresponding Riemann surface changes continuously. If $b = 0$, the space $M_{g,0,n}$ has a compact boundary $\partial M_{g,0,n}$. As a point approaches it, the Riemann surface degenerates, that is, gets "pinched." Thus, a 4-punctured sphere can become pinched so as to result in two 3-punctured spheres. A compact surface of genus 2 can become pinched to form two 1-punctured tori, or a single surface of genus 1 with two punctures, or two 3-punctured spheres.

Thus, a Riemann surface can be studied from three different vantage points: its complex analysis, its geometry in the hyperbolic metric (for most surfaces), and for compact surfaces, the algebraic geometry

of its associated polynomial. In addition, Riemann surfaces can be gathered together in spaces $M_{g,b,n}$ of topologically similar surfaces. The study of these spaces is called Teichmüller theory.   Albert Marden

Bibliography. L. V. Ahlfors, *Lectures on Quasiconformal Mappings*, 1966, reprint 1987; L. V. Ahlfors and L. Sario, *Riemann Surfaces*, 1960, reprint, Princeton University Press, 1996; C. Chevalley, *Introduction to the Theory of Algebraic Functions of One Variable*, 1951, reprint, American Mathematical Society, 1987; H. M. Farkas and I. Kra, *Riemann Surfaces*, 2d ed., Springer-Verlag, 1991; O. Forster, *Lectures on Riemann Surfaces*, 1981, reprint, Springer-Verlag, 1993; F. Kirwan, *Complex Algebraic Curves*, Cambridge University Press, 1992; G. Springer, *Introduction to Riemann Surfaces*, 2d ed., Chelsea Publishing, 1981.

Riemannian geometry

The geometry of riemannian manifolds. Riemannian geometry was initiated by B. Riemann in 1854, following the pioneering work of C. F. Gauss on surface theory in 1827. Riemann introduced a coordinate space in which the infinitesimal distance between two neighboring points is specified by a quadratic differential form, given below. Such a space is a natural generalization of euclidean geometry and Gauss's geometry of surfaces in three-dimensional euclidean space, as well as the noneuclidean geometries: hyperbolic geometry (previously discovered by J. Bolyai and N. I. Lobachevsky) and elliptic geometry. A riemannian manifold is a topological space that further generalizes this notion. Riemannian geometry derives great importance from its application in the general theory of relativity, where the universe is considered to be a riemannian manifold. *See* DIFFERENTIAL GEOMETRY; EUCLIDEAN GEOMETRY; NONEUCLIDEAN GEOMETRY; RELATIVITY.

Riemannian spaces. An n-dimensional riemannian space is a space whose points can be characterized by n coordinates u^i ($i = 1, 2, \ldots, n$), and where an infinitesimal distance ds between two points with coordinate differences du^i is given by a quadratic differential form, Eq. (1), to be called the metric. In

$$ds^2 = \sum_{i,j} g_{ij}\, du^i\, du^j \qquad (1)$$

$$g_{ij} = g_{ji} \qquad 1 \le i, j \le n$$

general, the quantities g_{ij} are functions of the coordinates u^i. This allows the definition of the arc length of a curve $u^i = u^i(t)$, $t_0 \le t \le t_1$. Indeed, taking the square root of both sides of Eq. (1) and integrating over t from t_0 to t_1 yields Eq. (2) for the arc length s.

$$s = \int_{t_0}^{t_2} \sqrt{\sum_{i,j} g_{ij} \frac{du^i}{dt} \frac{du^j}{dt}}\, dt \qquad (2)$$

Other geometric notions, such as the angle between intersecting curves and the volume of a domain, can also be defined. Generally the ds^2 is supposed to be positive definite, meaning that it is greater than zero

for all values of the du^i, unless they are all zero; or at least nondegenerate, meaning that inequality (3) is

$$\det(g_{ij}) \ne 0 \qquad (3)$$

satisfied, where $\det(g_{ij})$ is the determinant of the matrix whose elements are the g_{ij}. *See* DETERMINANT; INTEGRATION.

Examples of riemannian spaces include the following:

1. The n-dimensional euclidean space E^n has the metric given by Eq. (4).

$$ds^2 = (du^1)^2 + \cdots + (du^n)^2 \qquad (4)$$

2. More generally, elliptic and hyperbolic spaces are given by the noneuclidean metric of Eq. (5), where K is a constant and r^2 is given by Eq. (6).

$$ds^2 = \frac{(du^1)^2 + \cdots + (du^n)^2}{\left(1 + \dfrac{K}{4} r^2\right)^2} \qquad (5)$$

$$r^2 = (u^1)^2 + \cdots + (u^n)^2 \qquad (6)$$

The space is called elliptic if $K > 0$, and hyperbolic if $K < 0$; K is the curvature of the space.

3. If the assumption of positive definiteness is dropped, the n-dimensional lorentzian space with the metric given by Eq. (7) is a generalization of a

$$ds^2 = (du^1)^2 + \cdots + (du^{n-1})^2 - (du^n)^2 \qquad (7)$$

riemannian space. A four-dimensional lorentzian space characterizes the physical universe in the special theory of relativity. *See* SPACE-TIME.

4. If E^N is the euclidean space of dimension N with the coordinates $x_1, \ldots, x_N$, a submanifold M of dimension $n < N$ is defined by Eq. (8), where the functional

$$x_A = x_A(u^1, \ldots, u^n) \qquad 1 \le A \le N \qquad (8)$$

matrix $(\partial x_A / \partial u^i)$ is of rank n. Then M has the induced riemannian metric given by Eq. (9). The case $N = 3$, $n = 2$, is treated in the surface theory of Gauss.

$$ds^2 = \sum_{i,j,A} \frac{\partial x_A}{\partial u^i} \frac{\partial x_A}{\partial u^i}\, du^i\, du^j \qquad (9)$$

Tensors. Riemannian geometry is local in the sense that it is valid in a neighborhood of the u^i coordinates. It studies properties that are invariant under coordinate transformations of the form of Eq. (10),

$$u'^i = u'^i(u^1, \ldots, u^n) \qquad 1 \le i \le n \qquad (10)$$

where the functions are smooth with the functional determinant $\det(\partial u'^i / \partial u^j) \ne 0$. Thus, in the new coordinates, the metric is given by Eq. (11). This implies that the g'_{ij} are related to the g_{ij} by Eq. (12).

$$ds^2 = \sum_{i,j} g'_{ij}\, du'^i\, du'^j \qquad g'_{ij} = g'_{ji} \qquad (11)$$

$$g'_{ij} = \sum_{k,l} g_{kl} \frac{\partial u^k}{\partial u'^i} \frac{\partial u^l}{\partial u'^j} \qquad (12)$$

Following Einstein, from now on the convention will be adopted that repeated indices mean summation. It will also be assumed that all small Latin indices run from 1 to n, unless otherwise specified. A system of quantities $T^{i_1 \cdots i_r}{}_{j_1 \cdots j_s}$ associated with each coordinate system u^i is called a tensor, contravariant of order r and covariant of order s, if under the transformation of Eq. (10) they are transformed according to the rule given by Eq. (13). It is called a contravariant

$$T'^{i_1 \cdots i_r}{}_{j_1 \cdots j_s} = T^{k_1 \cdots k_r}{}_{l_1 \cdots l_s}$$
$$\times \frac{\partial u'^{i_1}}{\partial u^{k_1}} \cdots \frac{\partial u'^{i_r}}{\partial u^{k_r}} \frac{\partial u^{l_1}}{\partial u'^{j_1}} \cdots \frac{\partial u^{l_s}}{\partial u'^{j_s}} \quad (13)$$

vector if $r = 1$ and $s = 0$; it is called a covariant vector if $r = 0$ and $s = 1$. Thus, the g_{ij}, called the fundamental tensor, is a covariant tensor of order two and is symmetric. *See* TENSOR ANALYSIS.

Covariant differentiation. It is useful to define the matrix (g^{ij}) to be the inverse matrix of (g_{ij}), so that Eq. (14) is satisfied. Another set of useful quantities are the Christoffel symbols, given by Eq. (15). They

$$g_{ij}g^{jk} = \delta_i^k = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases} \quad (14)$$

$$\Gamma_{ijk} = \frac{1}{2}\left(\frac{\partial g_{ij}}{\partial u^k} + \frac{\partial g_{jk}}{\partial u^i} - \frac{\partial g_{ik}}{\partial u^j}\right) \quad (15)$$

can be used to define a covariant differentiation. For $r = 2$, $s = 1$, the covariant derivatives of $T^{i_1 i_2}{}_j$ are given by Eq. (16). A similar formula is valid for arbi-

$$T^{i_1 i_2}{}_{j|m} = \frac{\partial T^{i_1 i_2}{}_j}{\partial u^m} + \Gamma_k{}^{i_1}{}_m T^{k i_2}{}_j$$
$$+ \Gamma_k{}^{i_2}{}_m T^{i^1 k}{}_j - \Gamma_j{}^k{}_m T^{i_1 i_2}{}_k \quad (16)$$

trary values of r and s. A fundamental fact is that $T^{i_1 i_2}{}_{j|m}$ is still a tensor, of one covariant order higher than $T^{i_1 i_2}{}_j$. The Christoffel symbols are characterized by the symmetry condition of Eq. (17) and the prop-

$$\Gamma_{ik}^j = \Gamma_{ki}^j \quad (17)$$

erty that the covariant derivatives of the fundamental tensor are identically zero.

In particular, a contravariant vector T^i has the covariant differential given by Eq. (18), where ω_j^i is given by Eq. (19).

$$DT^i = dT^i + \omega_j^i T^j \quad (18)$$

$$\omega_j^i = \Gamma_{jk}^i \, du^k \quad (19)$$

Parallelism and geodesics. If the vectors $T^i(t)$ are given along a curve $u^i = u^i(t)$, they are parallel along the curve in the sense of T. Levi-Civita if Eq. (20) is

$$DT^i(t) = 0 \quad (20)$$

satisfied. This curve is called a geodesic if its tangent vectors du^i/dt are parallel. Geodesics are therefore the integral curves of the system of ordinary differential equations (21) of the second order.

$$\frac{d^2 u^i}{dt^2} + \Gamma_{jk}^i \frac{du^j}{dt}\frac{du^k}{dt} = 0 \quad (21)$$

Geometrically, geodesics are the shortest curves joining two points. In other words, Eqs. (21) are the Euler equations of the variational problem given by Eq. (22).

$$\delta \int ds = 0 \quad (22)$$

See CALCULUS OF VARIATIONS.

For the euclidean space given in Eq. (4), the fundamental tensor and Christoffel symbols are given by Eqs. (23). In this case, parallelism of a vector means

$$g_{ik} = \delta_{ik} \qquad \Gamma_{ijk} = 0 \quad (23)$$

constancy of its components T^i. Also, the geodesics are the straight lines defined by Eqs. (24). Thus,

$$\frac{d^2 u_i}{dt^2} = 0 \quad (24)$$

the above definitions of parallelism and geodesics are generalizations of standard notions of euclidean geometry. An essential difference is that Eqs. (20), being ordinary differential equations, are defined only along a curve, so that parallel displacement of a vector from one point to another depends on the curve joining them.

Riemann-Christoffel tensor. The properties of a riemannian space are described by its Riemann-Christoffel curvature tensor. It is defined by Eq. (25),

$$R_{ijrs} = \frac{\partial \Gamma_{ijs}}{\partial u^r} - \frac{\partial \Gamma_{ijr}}{\partial u^s}$$
$$+ g^{kb}(\Gamma_{ikr}\Gamma_{jbs} - \Gamma_{iks}\Gamma_{jbr}) \quad (25)$$

and satisfies the symmetry relations of Eqs. (26).

$$R_{ijrs} = -R_{jirs} = -R_{ijsr}$$
$$R_{ijrs} = R_{rsij} \quad (26)$$
$$R_{ijkl} + R_{iklj} + R_{iljk} = 0$$

The tensor has the following geometrical interpretation: If ξ^i and η^j are two vectors, and σ is the plane spanned by them, then the Plücker coordinates of the plane σ are given by Eq. (27), and the quantity K defined by Eq. (28) depends only on the point u

$$p^{ij} = \xi^i \eta^j - \xi^j \eta^i \quad (27)$$

$$K(u, \sigma) = -\frac{R_{ijkl}p^{ij}p^{ij}}{(g_{ik}g_{jl} - g_{il}g_{jk})p^{ij}p^{ij}} \quad (28)$$

and the plane σ through u. It is called the sectional or riemannian curvature at (u, σ). Geometrically, it is equal to the gaussian curvature of the surface generated by the geodesics through u and tangent to the vectors of σ.

The simplest riemannian spaces are those of constant curvature, for which the sectional curvature is independent of (u, σ). Their condition is Eq. (29),

$$R_{ijkl} = -K(g_{ik}g_{jl} - g_{il}g_{jk}) \quad (29)$$

where K is a constant. The metric given by Eq. (5) has this property.

Other important tensors can be constructed from the Riemann-Christoffel tensor. Thus, Eq. (30) de-

$$R_{ik} = g^{jl}R_{ijkl} = R_{ki} \tag{30}$$

fines a tensor called the Ricci tensor, and the quantity R defined by Eq. (31) is a scalar, called the scalar curvature.

$$R = g^{ik}R_{ik} \tag{31}$$

The four-dimensional lorentzian space characterizes the universe of the theory of general relativity. It is restricted by the Einstein field equations (32),

$$R_{ik} - {}^1\!/_2\, g_{ik}R = 8\pi\kappa T_{ik} \tag{32}$$

where κ is a constant and T_{ik} is the energy-stress tensor.

Manifolds and bundles. At the foundation of riemannian geometry is the notion of a riemannian manifold. A differentiable manifold is a topological space with a covering by coordinate neighborhoods, such that in the domain where the u^i coordinates and the $u^{i\prime}$ coordinates are both valid, they are related by the transformation of Eq. (10) with the conditions mentioned there. The manifold is called riemannian if a riemannian metric is defined in each neighborhood that gives the same value of ds^2 in the intersection of any two neighborhoods; that is, Eq. (12) is satisfied. Riemannian geometry is divided between local and global: the former dealing with properties in a neighborhood, the latter with properties of the manifold as a whole. *See* MANIFOLD (MATHEMATICS); TOPOLOGY.

Examples of local properties include the Levi-Civita parallelism defined by Eq. (20) and the Riemann-Christoffel tensor. An example of a global property is the notion of completeness: A riemannian manifold is called complete if every geodesic can be indefinitely extended, with the possibility of a closed geodesic repeating itself (such as a great circle on the sphere) not excluded. On a complete riemannian manifold any two points can be joined by a shortest geodesic.

The tensors have a global formulation, in the notion of tensor bundles. If the riemannian manifold M is of dimension n, the contravariant vectors, or tangent vectors, at a point x in M, form the tangent space T_x and the covariant vectors, or covectors, form the cotangent space T_x^*. They are dual vector spaces in the sense that there is a pairing that associates to each X in T_x and ω in T_x^* a quantity $\langle X, \omega \rangle$, which is a real number and is linear in each of the arguments. In fact, if X has the components X^i and ω has components ω_i in the same local coordinate system, then $\langle X, \omega \rangle$ is given by Eq. (33).

$$\langle X, \omega \rangle = X^i \omega_i \tag{33}$$

More generally, the tensors $T_x(r,s)$ of contravariant order r and covariant order s at a point x in M form a vector space of dimension n^{r+s}, which is given by

Eq. (34). A tensor bundle over M is defined to be the

$$T_x(r, s) = \underbrace{T_x \otimes \cdots \otimes T_x}_{r} \otimes \underbrace{T_x^* \cdots \otimes T_x^*}_{s} \tag{34}$$

union of the vector spaces $T_x(r,s)$ for each point x in M, denoted

$$\bigcup_{x \in M} T_x(r, s)$$

together with the projection mapping π defined by Eq. (35). For $r = 1$, $s = 0$ and for $r = 0$, $s = 1$,

$$\pi: \bigcup_{x \in M} T_x(r, s) \to M\pi\, T_x(r, s) = x \tag{35}$$

these are, respectively, the tangent and cotangent bundles over M. A mapping given by Eq. (36) such

$$s: M \to \bigcup_{x \in M} T_x(r, s) \tag{36}$$

that π o s = identity, that is, the assignment of a tensor to every point of M, is called a section of the tensor bundle or a tensor field.

Space-forms and symmetric spaces. Some results on global riemannian geometry will now be described. The riemannian manifolds of constant curvature are locally characterized by Eq. (29). They are called space-forms, because they are locally the euclidean and noneuclidean spaces. The metric is given by Eqs. (5). For $K = 0$, such a space is a euclidean space E^n with the coordinates $u^1, \ldots, u^n$. For $K < 0$, the space is restricted to the domain of points satisfying Eq. (37), and is called the hyperbolic space

$$(u^1)^2 + \cdots + (u^n)^2 < -\frac{4}{K} \tag{37}$$

H^n. For $K > 0$, E^n can be compactified by the addition of a so-called point at infinity, making it into the n-dimensional sphere S^n; the resulting space is called the elliptic or spherical space. All these spaces, the euclidean, the hyperbolic, and the spherical, are simply connected (that is, any closed curve in the space can be deformed continuously to a point).

A fundamental theorem says that a complete, simply connected riemannian manifold of constant curvature K is isometric to the euclidean space, the hyperbolic space, or the spherical space, according to whether $K = 0$, $K < 0$, or $K > 0$. (An isometry is a one-to-one mapping that preserves the metric ds^2.) It follows that a complete riemannian manifold of constant curvature (which is not necessarily simply connected) is covered by one of the above three spaces. Two important examples are the following:

1. The manifold constructed by identifying in E^n the points $(u^1, \ldots, u^n)$ and $(u^1 + 1, \ldots, u^n + 1)$. The result is an n-dimensional torus, having a metric with $K = 0$.

2. The manifold constructed by identifying the antipodal points of S^n. The result is the real projective space, having a metric with $K > 0$.

A more general family of riemannian spaces consists of the symmetric spaces. They are characterized
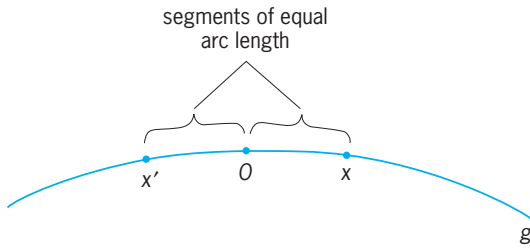
Fig. 1. Configuration of points in definition of a symmetry about a point 0.

by the vanishing of the covariant derivatives of the Riemann-Christoffel tensor, that is, by Eq. (38). Ge-

$$R_{ijkl|m} = 0 \qquad (38)$$

ometrically this can be expressed by the property that the symmetries about the points of the space are isometries. (A symmetry about a point 0 of a riemannian space is a mapping that takes any nearby point x into a point x' such that x, 0, and x' lie on a single geodesic g in the configuration shown in **Fig. 1**, and the arc lengths of the segments $x'0$ and $0x$ of g are equal.) Symmetric riemannian spaces are of great importance in the applications of riemannian geometry.

Gauss-Bonnet formula. In two dimensions, this formula states the following: If D is a domain with a sectionally smooth boundary, then Eq. (39) is satis-

$$\sum (\pi - \alpha) + \int_{\alpha D} k_g \, ds$$
$$+ \int \int_D K \, dA = 2\pi \chi(D) \quad (39)$$

fied, where the first term is the sum of the exterior angles at the corners, the second term is the integral of the geodesic curvature along the sides, the third term is the integral of the gaussian curvature over D, and $(D)_\chi$ is the Euler characteristic of D. The Euler characteristic can be defined by dividing up D into cells (**Fig. 2**); then the Euler characteristic is given by Eq. (40). For a rectilinear triangle in the euclidean

$\chi(D)$ = number of vertices

$\qquad$ − number of edges + number of cells (40)

plane, $k_g = 0$, $K = 0$, and Eq. (39) indicates that the angle-sum is equal to π.

In higher dimensions, for the sake of simplicity, the theorem will be given only for a compact oriented riemannian manifold M of dimension $n = 2m$ without boundary. The so-called curvature form is given by Eq. (41), where the wedge symbol $\wedge$ de-

$$\Omega_{ij} = \sum R_{ijkl} du^k \wedge du^l \qquad (41)$$

notes exterior or wedge multiplication. The pfaffian is defined by Eq. (42), where $\epsilon_{i_1 \ldots i_n}$ is +1 or −1 ac-

$$Pf = \sum \epsilon_{i_1 \ldots i_n} \Omega_{i_1 i_2} \wedge \ldots \wedge \Omega_{i_{n-1} i_n} \qquad (42)$$

cording as its indices form an even or odd permu-

tation of $1, \ldots, n$, and is otherwise zero, and the sum is extended over all indices from 1 to n. Then the Gauss-Bonnet theorem states that Eq. (43) holds,

$$(-1)^m \frac{1}{2^{2m} \pi^m m!} \int_M Pf = \chi(M) \qquad (43)$$

where $\chi(M)$ is the Euler-Poincaré characteristic of M. The Euler-Poincaré characteristic, an important topological invariant, can be defined by dividing up M into cells; then the Euler-Poincaré characteristic is given by Eq. (44), a generalization of Eq. (40). The

$$\chi(M) = \sum_{0 \leq k' \dim M} (-1)^k \, (\text{number of } k\text{-cells}) \quad (44)$$

Gauss-Bonnet theorem gives a relation between local and topological invariants.

DeRham and Hodge theorems. A geometrical idea that leads to important topological invariants is that of bounding or connectivity. For instance, every closed curve on the two-dimensional sphere bounds a domain, while there are closed curves on the two-dimensional torus that do not bound. This idea is developed into the homology theory in algebraic topology. Dual to this is the notion of exterior differential forms, given by Eq. (45), where the $a_{i_1 \ldots i_p}$ are smooth

$$\alpha = a_{i_1 \ldots i_p} du^{i_1} \wedge \cdots \wedge du^{i_p} \qquad (45)$$

functions in the coordinates u_i, and the wedge multiplication of differentials is antisymmetric, that is, Eq. (46) holds. The form α is said to be of degree p.

$$du^i \wedge du^j = -du^j \wedge du^i \qquad (46)$$

The exterior derivative is defined by Eq. (47), and is

$$d\alpha = da_{i_1 \ldots i_p} \wedge du^{i_1} \wedge \cdots \wedge du^{i_p} \qquad (47)$$

of degree $p + 1$. The exterior differentiation, as an operator on exterior differential forms, can be said to be adjoint to the boundary operator, in view of
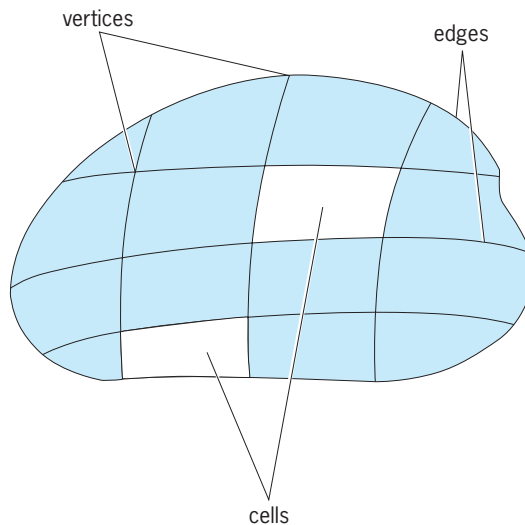

Fig. 2. Division of a domain D into cells in the definition of the Euler characteristic.

Stokes' theorem, which states that Eq. (48) holds,

$$\int_D d\alpha = \int_{\alpha D} \alpha \tag{48}$$

where D is a $(p + 1)$-dimensional domain and αD is its boundary. It is easily verified that $dd\alpha = 0$. *See* STOKES' THEOREM.

All the smooth forms of degree p on a manifold M constitute a vector space A^p (of infinite dimension). The form α is said to be closed if $d\alpha = 0$. If C^p is the space of closed forms of degree p, then dA^{p-1} is a substance of C^p. DeRham's theorem states that if the manifold M is compact, the quotient space C^p/dA^{p-1} (also denoted C^p mod dA^{p-1}) is of finite dimension b^p, which is the pth Betti number of M, an important topological invariant of M.

When M is a riemannian manifold of dimension n, this theorem can be refined. Based on the inner product in the cotangent space T_x^*, x in M, the $*$-operator can be defined, which maps a p-form α into an $(n - p)$-form $*\alpha$. The codifferential δ defined by Eq. (49) maps a p-form into a $(p - 1)$-form. The op-

$$\delta = \pm * d* \tag{49}$$

erator Δ defined by Eq. (50) maps a p-form into a

$$\Delta = d\delta + \delta d \tag{50}$$

p-form, and is called the Laplace operator. This is a generalization, to a riemannian manifold, of the classical Laplace operator in analysis. In analogy to the classical case, a form α is called harmonic if $\Delta\alpha = 0$. *See* LAPLACE'S DIFFERENTIAL EQUATION.

Hodge's theorem states that for every compact riemannian manifold a coset of C^p mod dA^{p-1} contains exactly one p-form that is harmonic.

Generalizations. Riemannian geometry has several interesting generalizations, including finslerian geometry, symplectic geometry, and the complex manifolds.

Finslerian geometry. Riemannian geometry can be considered to be based on the integral in Eq. (2), giving the arc length of a curve. The geometry based on the more general integral of Eq. (51), where F

$$s = \int_{t_0}^{t_1} F\left(u^i, \frac{du^i}{dt}\right) dt \tag{51}$$

is positively homogeneous of degree one, that is, Eq. (52) holds, is called finslerian geometry. Besides

$$F(u^i, \lambda v^i) = \lambda F(u^i, v^i) \qquad \lambda > 0 \tag{52}$$

riemannian geometry, another important special case of finslerian geometry is minkowskian geometry, where $F(du^i/dt)$ is a function of du^i/dt only. In this case, the geodesics are straight lines, and a theory of convex bodies can be developed, which plays an important role in number theory.

Symplectic geometry. Another interesting object is the replacement of the symmetric covariant tensor g_{ij} by an antisymmetric one, that is, a covariant tensor a_{ij}

that satisfies Eq. (53). The matrix (a_{ij}) is taken to be

$$a_{ij} = -a_{ji} \tag{53}$$

of maximum rank, so that the space is of even dimension. The tensor is the same as the exterior quadratic differential form given by Eq. (54). The space is called

$$\alpha = a_{ij} du^i \wedge du^j \tag{54}$$

symplectic if such an α is given, satisfying $d\alpha = 0$. The resulting geometry is called symplectic geometry. It is a fundamental structure in mechanics. *See* MATRIX THEORY.

Unlike riemannian geometry, symplectic geometry has no local invariants, for a theorem of J. G. Darboux states that there are local coordinates where α takes the normal form given by Eq. (55).

$$\alpha = du^1 \wedge du^2 + \cdots + du^{n-1} \wedge du^n \tag{55}$$

Complex manifolds. Finally, there is the important generalization when the coordinates u^i are complex numbers, and the change of coordinates of Eq. (10) is given by holomorphic functions. Such a manifold is called a complex manifold. It is of complex dimension n and even real dimension $2n$. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

Even the case $n = 1$ is of great importance. A complex manifold of (complex) dimension one is called a Riemann surface. It is a basic object in complex function theory and plays a fundamental role in string theory in theoretical physics. *See* RIEMANN SURFACE; SUPERSTRING THEORY.

A complex manifold is called hermitian if there is a hermitian differential form, given by Eq. (56), where

$$ds^2 = b_{ij} du^i d\overline{u}^j \qquad \overline{b}_{ij} = b_{ji} \tag{56}$$

the bar denotes complex conjugation. Hermitian geometry is the counterpart of riemannian geometry for complex manifolds.   S. S. Chern

Bibliography. E. Cartan, *Geometry of Riemannian Spaces*, 1983; I. Chavel, *Riemannian Geometry: A Modern Introduction*, 1995; S. Chern, *Complex Manifolds Without Potential Theory*, 1978; L. P. Eisenhart, *Riemannian Geometry*, 2d ed., 1950; W. Klingenberg, *Riemannian Geometry*, 2d ed., 1995; S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, vol. 1, 1963, vol. 2, 1969; C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, 1973.

Rift valley

One of the geomorphological expressions between two tectonic plates that are opening relative to each other or sliding past each other. The term originally was used to describe the central graben structures of such classic continental rift zones as the Rhinegraben and the East African Rift, but the definition now encompasses mid-oceanic ridge systems with central valleys such as the Mid-Atlantic Ridge. *See* MID-OCEANIC RIDGE; PLATE TECTONICS.
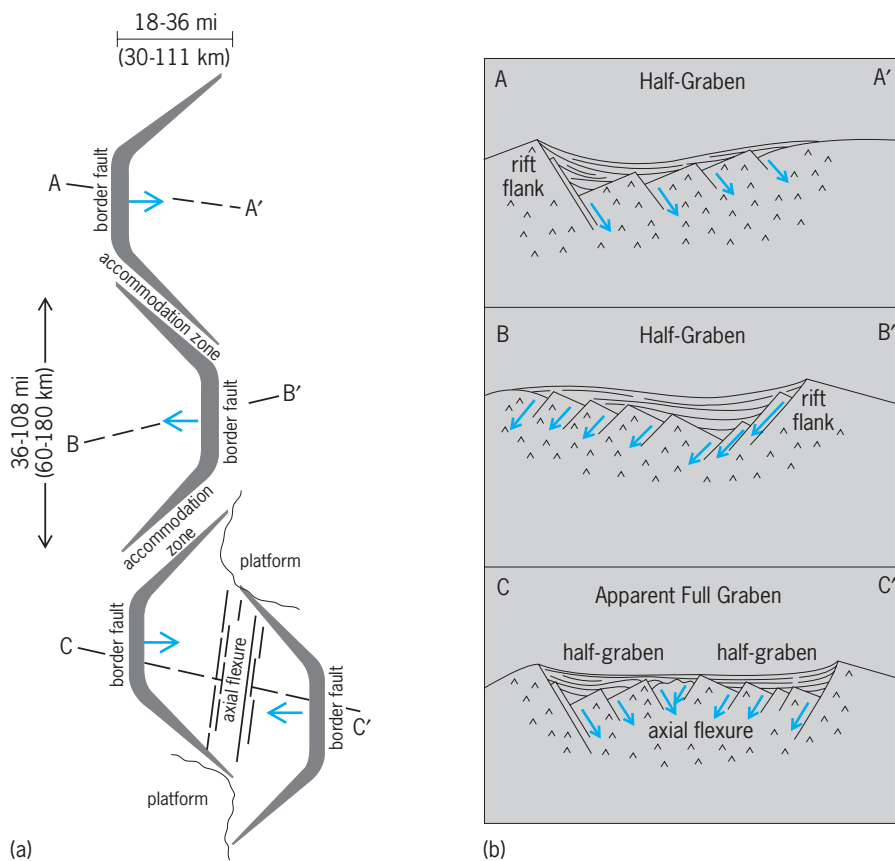
Sketch showing several ways in which half-graben units can link to make rift zones. (*a*) Map view. (*b*) Cross sections. A-A′ and B-B′ show simple half-grabens that alternate polarity along strike. C-C′ shows a typical morphology across an overlapping, facing arrangement of half-grabens.

Continental and oceanic rift valleys are end members in what many consider to be an evolutionary continuum. In the case of continental rift valleys, plate separation is incomplete, and the orientation of the stress field relative to the rift valley can range from nearly orthogonal to subparallel. Strongly oblique relationships are probably the norm. Also, volcanism is not a universal phenomenon in continental rift valleys, and when it does occur, the positioning, expressions, and chemistries of the effluents are complicated and variable. In contrast, oceanic rift valleys mark the place where the trailing edges of two distinctly different plates are separating. The separation is complete, and the spreading is organized and focused, resulting in rift valleys that tend to be oriented orthogonal or suborthogonal to the spreading directions. Basaltic magmas with a tholeiitic chemistry upwell from the asthenosphere into oceanic rift valleys, essentially filling the void that otherwise would develop between the separating plates. This is how new oceanic lithosphere is generated. *See* ASTHENOSPHERE; MAGMA.

The basic cross-sectional form of rift valleys consists of a central graben surrounded by elevated flanks. Considerable debate exists regarding how much of the elevation of the flanks is residual and how much represents actual uplift, but it is almost universally accepted that the central grabens of con-

tinental rift valleys are subsidence features. The crystalline basement floors of some parts of the Tanganyika and Malawi (Nyasa) rift valleys in East Africa lie more than 5 mi (9 km) below elevated flanks. *See* GRABEN.

In continental rift valleys the true cross-sectional form is typically asymmetric, with the rift floors tilted toward the most elevated flank. Most of the subsidence is controlled by one border fault system, and most of the internal faults parallel the dip of the border faults. Sedimentation and volcanism tend to obscure this basic symmetry, especially where deposits are very thick. The polarity of these half-graben structures alternates along the strike of many rift zones (see **illus.**). This creates a kind of segmentation in which half-grabens of opposite dip domains are separated by structures referred to as accommodation zones (or transfer faults). It is possible that these polarity alternations may occur when ancient dislocation systems cross the rift axis. *See* FAULT AND FAULT STRUCTURES.

Oceanic rift valleys are also distinctly separated into segments by structures known as transform faults. The cross-sectional form of oceanic rift valleys can be markedly asymmetric, but this may not be the norm, and it is unclear if any pattern occurs to the variations along strike. Regardless, it is unlikely that the cross-sectional form of oceanic rift valleys is related genetically to that of continental rift valleys,

except in the broadest possible terms. *See* TRANS-FORM FAULT.

Many modern continental rift valleys are marked by spectacular, deep lakes whose floors can be covered with many miles of sediment. Lakes Tanganyika and Malawi (Nyasa) in East Africa and Lake Baikal in Russia are examples. Ancient continental rift valleys are marked usually by long, narrow strips of sedimentary rocks or by volcanic rocks that fill in what were originally basins. The fact that there are many continental rifts argues that the fate of most rift valleys is fossilization prior to successful plate separation. The Mid-Continent Rift in the midwestern United States is an example of a rift that failed more than a billion years ago. The North Sea is an example of a system of rift basins that were subjected to much extension prior to abandonment about 100 million years ago. The North Sea and the Gulf of Suez are examples of failed rifts that were subjected to marine incursion. However, these are not oceanic rift valleys in the true sense because neither reached the stage of complete plate separation (that is, no oceanic lithosphere was generated). The North Sea, Gulf of Suez, and Reconcavo Basin in Brazil demonstrate that rift valley settings can be associated with prolific deposits of hydrocarbons. *See* BASIN; LITHOSPHERE.   Bruce R. Rosendahl

Bibliography. D. W. Burbank and R. S. Anderson, *Tectonic Geomorphology: A Frontier in Earth Science*, 2001; P. M. LeTourneau and P. E. Olsen (eds.), *The Great Rift Valleys of Pangea in Eastern North America, Volume II: Sedimentology, Stratigraphy, and Paleontology*, 2003; B. R. Rosendahl, Architecture of continental rifts with special reference to East Africa, *Annu. Rev. Earth Planet. Sci.*, 15:445–503, 1987; B. R. Rosendahl et al., Structural expressions of rifting: Lessons from Lake Tanganyika, Africa, in *Sedimentation in the African Rifts*, Spec. Publ. GSSP 23, Quart. J. Geol. Soc. Lond., 1986.

Rift Valley fever

An arthropod-borne (primarily mosquito), acute, febrile, viral disease of humans and numerous species of animals, particularly ruminants. Rift Valley fever is caused by a ribonucleic acid (RNA) virus in the genus *Phlebovirus* of the family Bunyaviridae. In sheep and cattle, it is also known as infectious enzootic hepatitis. First described in the Rift Valley of Africa, the disease presently occurs in west, east, and south Africa and has extended as far north as Egypt. Outside of Africa, Rift Valley fever has spread to the island of Madagascar and the Arabian Peninsula (Saudi Arabia and Yemen).

Clinical signs and pathology. In humans, clinical signs of Rift Valley fever are influenzalike, and include fever, headache, muscular pain, weakness, nausea, epigastric pain, and photophobia. Most people recover within 4–7 days, but some individuals may have impaired vision or blindness in one or both eyes; a small percentage of infected individuals develop encephalitis or a hemorrhagic syndrome and die.

The clinical signs of Rift Valley fever in animals depend on the species infected, age, and physiological condition. The most susceptible are neonatal lambs, calves, and puppies, which develop a high fever; become weak, and die 24–48 h after infection. The mortality rate in lambs that are less than a week of age can be over 90%. Mature sheep and cattle develop a high fever; abortion at any stage of gestation is the most prominent sign. Mortality rates in aborting ewes may reach 30%, and 10% in aborting cows.

The prominent lesion in aborted lambs and calves results from the death of many liver cells, causing the organ to become yellow and friable. Lesions in older animals that die from the disease consist of widespread hemorrhages on serous membranes and on intestines, and the localized death of liver cells.

Diagnosis. Rift Valley fever should be suspected when the following observations are made in a disease outbreak: (1) high abortion rates (possibly approaching 100%) in ewes, cows, and female dogs, with lower rates in goats and in other ruminants; (2) high mortality (possibly approaching 100%) in lambs and calves less than 7 days old, and lower rates of disease and mortality in older animals; (3) extensive liver lesions in aborted fetuses and newborn animals; (4) an influenzalike disease in people, particularly in individuals associated with livestock; (5) occurrence of the disease during a period of high insect activity following heavy seasonal rainfall with persistent flooding; (6) rapid spread. The diagnosis is confirmed by isolating the virus from tissues of an infected animal or human. Viral antigen may be detected in the laboratory from tissues or tissue smears using a number of techniques. Antibodies (IgM or IgG) can be detected in serum.

Epidemiology. Historically, outbreaks of Rift Valley fever have occurred at 10–15-year intervals in normally dry areas of Africa subsequent to a period of heavy rainfall. During interepidemic periods the virus is present in the eggs of the mosquito *Aedes lineatopinnis*, which can survive for many years in the dried soil of grassland depressions called dambos. After a period of heavy rain, the dambos fill with water, the mosquito eggs hatch, the larvae mature, and then the infected mosquitoes can feed on a ruminant animal, which becomes infected. A key factor in the epizootic (outbreak of disease among animal populations) of Rift Valley fever is sustained flooding, which leads to the emergence of large numbers of secondary (that is, other species of) mosquitoes. Once the virus proliferates in the blood of the ruminant animal, most species of mosquito feeding on this animal can also become infected and transmit the virus to other animals. Humans can also become infected by an aerosol that can be produced during autopsy; ritual slaughter, as occurred in Saudi Arabia, by contact with an infected animal that has aborted; or by laboratory procedures. Moreover, they can develop a sufficiently high blood infection to be a source of infection for mosquitoes; thus, a sick individual could accidentally transmit the virus to other parts of the world. In Egypt, human

infection has occurred through mosquito bites in an urban peridomestic mosquito cycle.

Prevention and control. Animals should not be moved from an infected area to a Rift Valley–free area if there is any indication of active disease. Control of mosquitoes by spraying insecticides may be of value in reducing human exposure in limited areas. In addition, control of the disease is best accomplished by widespread vaccination of susceptible animals to prevent amplification of the virus and, thus, infection of vectors. Any individual that works with infected animals or live virus in a laboratory should be vaccinated. An inactivated cell-culture virus vaccine is available for use in both animals and humans. This vaccine has the disadvantages of requiring two inoculations over a 30-day period and a large quantity of antigen. A modified live Smithburn strain vaccine is available for use in nonpregnant animals. Other vaccines still under development are based on a mutagen-induced MP12 strain and a Clone 13 strain. *See* ANIMAL VIRUS; VACCINATION.   Charles A. Mebus

Bibliography. H. Caplan, C. J. Peters, and D. H. L. Bishop, Mutagen-directed attenuation of Rift Valley fever virus as a method for vaccine development, *J. Gen. Virol.*, 66:2271–2277, 1985; F. G. Davies, K. J. Linthicum, and A. D. James, Rainfall and epizootic Rift Valley fever, *Bull. WHO*, 63:941–943, 1985; G. Geran (ed.), *CRC Handbook Series in Zoonosis*, 1981; A. Jouan et al., Analytical study of a Rift Valley fever epidemic, *Res. Virol.*, 140(2):175–186, 1989; K. J. Linthicum et al., Climate and satellite indicators to forecast Rift Valley fever epidemics in Kenya, *Science*, 285:397–400, 1999; C. J. Peters and J. M. Meegan, Rift Valley fever, in G. W. Beran and J. H. Steele (eds.), *Handbook of Zoonoses*, Sec. B: *Viral Zoonoses*, 2d ed., pp. 403–420, CRC Press, Boca Raton, 1994; W. C. Reeves et al., The potential effect of global warming on mosquito-borne arboviruses, *J. Med. Entomol.*, 31(3):323–332, 1994; A. Shimshony and R. Barzilai, Rift Valley fever, *Adv. Vet. Sci. Comp. Med.*, 27:347–425, 1983; WHO/FAO Working Group on Rift Valley fever, *Rift Valley Fever: An Emerging Human and Animal Problem*, WHO Publ. 63, 1982; H. Zeller et al., Enzootic activity of Rift Valley fever virus in Senegal, *Am. J. Trop. Med. Hyg.*, 56(3):265–272, 1997.

Rigel

The bright star in the southwest corner of the constellation Orion (apparent magnitude +0.12), also referred to as β Orionis. It is a blue-white supergiant of spectral type B8, one of the most luminous stars known. Its intrinsic brightness is estimated to be more than 50,000 times that of the Sun. Rigel is a very young star by stellar standards, with such an enormous rate of energy output that its life span is only a few million years. By comparison, the Sun is approximately 5×10^9 years old, and is still only half way through its main-sequence evolution. Rigel is the brightest member of the Orion OB1 Association, a large molecular cloud complex in which active star formation is currently taking place. In addition to its high luminosity, Rigel has a photosphere which is so large that if the star were placed where the Sun is, it would almost fill the orbit of Mercury. *See* MOLECULAR CLOUD; SPECTRAL TYPE.

The spectrum of Rigel reveals pronounced absorption lines of neutral helium and hydrogen. There is evidence that the star is loosing mass in the form of strong winds emanating from its surface, at a rate of 10^{-6} solar mass per year, or about 6×10^{13} metric tons (7×10^{13} tons) per second. It rotates around its axis with a projected velocity of 42 km/s (26 mi/s). The effective temperature is approximately 12,000 K (21,100°F). Rigel also displays small irregular light variations and erratic changes in its radial velocity that seem to be quite common among evolved stars and may be due to pulsation. *See* ORION; STELLAR ROTATION; STAR; SUPERGIANT STAR.   David W. Latham

Rigid body

An idealized extended solid whose size and shape are definitely fixed and remain unaltered when forces are applied. Treatment of the motion of a rigid body in terms of Newton's laws of motion leads to an understanding of certain important aspects of the translational and rotational motion of real bodies without the necessity of considering the complications involved when changes in size and shape occur. Many of the principles used to treat the motion of rigid bodies apply in good approximation to the motion of real elastic solids. *See* NEWTON'S LAWS OF MOTION; RIGID-BODY DYNAMICS.   Dudley Williams

Rigid-body dynamics

The study of the motion of a rigid body under the influence of forces. A rigid body is a system of particles whose distances from one another are fixed. The general motion of a rigid body consists of a combination of translations (parallel motion of all particles in the body) and rotations (circular motion of all particles in the body about an axis). Its equations of motion can be derived from the equations of motion of its constituent particles. *See* RECTILINEAR MOTION; ROTATIONAL MOTION.

The location of a mass point m_i can be specified relative to a fixed-coordinate system by a position vector $\vec{r}_i$ with cartesian components (x_i, y_i, z_i). The vector force $\vec{F}_i$ which acts on the mass points has corresponding components (F_{ix}, F_{iy}, F_{iz}). Newton's second law for the motion of m_i is stated in Eq. (1).

$$\vec{F}_i = m_i \ddot{\vec{r}}_i \qquad (1)$$

Here $\ddot{\vec{r}}_i \equiv d^2\vec{r}_i/dt^2 = \vec{a}$ is the acceleration of m_i. *See* ACCELERATION; FORCE; NEWTON'S LAWS OF MOTION.

Translational motion. If Eq. (1) is summed over all particles in the rigid body, the left-hand side becomes

the total force

$$\vec{F} = \sum_i \vec{F}_i$$

acting on the rigid body. If the internal forces satisfy Newton's third law (to each action there is an equal but opposite reaction), the contributions of the internal forces cancel in pairs and $\vec{F}$ is the total external force on the rigid body, $\vec{F}^{\text{ext}}$. The right-hand side can be expressed in terms of the center-of-mass (CM) position vector defined by Eq. (2), where M is the total

$$\vec{R} = \frac{1}{M} \sum_i m_i \vec{r}_i \qquad (2)$$

$$M = \sum_i m_i$$

mass of the body. [In continuous systems the mass elements m_i are replaced by differential mass elements $dm = \rho(\vec{r})dV$. where $\rho(\vec{r})$ is the mass density and dV is the differential volume element. For a rigid body with a continuous mass density, the center-of-mass vector is calculated by integrating dm over the volume of the body as in Eq. (3).] Then the sum of

$$\vec{R} = \frac{1}{M} \int \vec{r} \rho(\vec{r}) \, dV \qquad (3)$$

the equations of motion, Eq. (1), takes the form of Eq. (4). This equation of motion for the center of

$$\vec{F}^{\text{ext}} = M\ddot{\vec{R}} \qquad (4)$$

mass of the rigid body has exactly the form of the equation of motion for a particle of mass M and position $\vec{R}$, under the influence of an external force $\vec{F}^{\text{ext}}$. Consequently, the second law of motion holds, not just for a particle, but for an arbitrary rigid body, if the position of the body is interpreted to mean the position of its center of mass. *See* CENTER OF MASS.

The momentum of a mass point is given by the product of the mass and the velocity, $\vec{p}_i = m_i \dot{\vec{r}}_i$, where $\dot{\vec{r}}_i \equiv d\vec{r}_i/dt$. The total momentum of the center of mass of the rigid body, obtained by summing over the momenta $\vec{p}_i$ of its constituent masses, is given by Eq. (5). In terms of the center-of-mass momentum

$$\vec{P} = M\dot{\vec{R}} \qquad (5)$$

$\vec{P}$, the equation of motion for the center of mass is expressed by Eq. (6). For an isolated rigid body, the

$$\dot{\vec{P}} = \vec{F}^{\text{ext}} \qquad (6)$$

external force is zero and therefore $\vec{P}$ is constant. According to Eq. (5), this implies that the center of mass moves with constant velocity $\vec{V} = \vec{p}/M$. *See* CONSERVATION OF MOMENTUM; MOMENTUM.

In fact, the preceding equations for translational motion hold for any body, rigid or nonrigid.

Rotational motion. The total angular momentum of a rigid body about a point O with coordinate $\vec{r}_O$ is the sum of the angular momenta of its constituent masses, and is given by Eq. (7). Here $\times$ denotes the

$$\vec{L}_O = \sum_i (\vec{r}_i - \vec{r}_O) \times m_i(\dot{\vec{r}}_i - \dot{\vec{r}}_O) \qquad (7)$$

cross-product of the coordinate vector $(\vec{r}_i - \vec{r}_O)$ with the momentum vector $m_i(\dot{\vec{r}}_i - \dot{\vec{r}}_O)$. The time derivative of $\vec{L}$ is given in Eq. (8). Hereafter the point O

$$\dot{\vec{L}}_O = \sum_i (\vec{r}_i - \vec{r}_O) \times m_i(\ddot{\vec{r}}_i - \ddot{\vec{r}}_O) \qquad (8)$$

is taken to be either a fixed point (in which case $\dot{\vec{r}}_O = \ddot{\vec{r}}_O = 0$) or the center-of-mass point. Using the equation of motion (1), $m_i \ddot{\vec{r}}_i$ can be replaced by $\vec{F}_i$. Thus the rotational equation of motion (9) is obtained. The right-hand side of Eq. (9) is known as the

$$\dot{\vec{L}}_O = \sum_i (\vec{r}_i - \vec{r}_O) \times \vec{F}_i \qquad (9)$$

tained. The right-hand side of Eq. (9) is known as the torque, $\vec{N}$. The contribution of the internal forces to the torque vanishes if the "extended third law" holds; namely, action equals reaction and is directed along a line between the particles. In this circumstance the rotational equation of motion is given by Eqs. (10),

$$\dot{\vec{L}}_0 = \vec{N}_0^{\text{ext}} \qquad (10a)$$

$$\vec{N}_0^{\text{ext}} = \sum_i (\vec{r}_i - \vec{r}_O) \times \vec{F}_i^{\text{ext}} \qquad (10b)$$

where $\vec{N}^{\text{ext}}$ is the total torque associated with external forces that act on the rigid body. *See* ANGULAR MOMENTUM; TORQUE.

It is straightforward to show from Eq. (7) that the angular momentum about an arbitrary point O is related to the angular momentum about the center of mass by Eq. (11). The torque about an arbitrary point

$$\vec{L}_O = \vec{L}_{\text{CM}} + (\vec{R} - \vec{r}_O) \times \vec{P} \qquad (11)$$

O can also be easily related to the torque about the center of mass, by Eq. (12).

$$\vec{N}_O^{\text{ext}} = \vec{N}_{\text{CM}}^{\text{ext}} + (\vec{R} - \vec{r}_O) \times \vec{F}^{\text{ext}} \qquad (12)$$

The position of every particle in a rigid body is specified by the position of any one point of the body (such as the center of mass) plus the orientation of the body about that point. Three coordinates are needed to specify one particle in the body. Two angular coordinates can specify the position of a second particle, since it lies at a fixed distance from the first. Only one further coordinate is needed to locate the position of a third particle because its distances from the first and second particles are fixed. The positions of any other particles in the rigid body are completely determined by their distances from the first three particles. Hence six coordinates determine the positions of all particles in a rigid body, and the motion of a rigid body is described by six equations of motion. The translational motion of the center of mass is determined by Eq. (6), and the rotational motion about the center of mass, or a fixed point, is determined by Eq. (10). These six equations, which hold for any system of particles, completely describe the motion of a rigid body.

Motion of an isolated system. The equation $\dot{\vec{L}} = \vec{N}^{\text{ext}}$ has the same form as $\dot{\vec{P}} = \vec{F}^{\text{ext}}$. Both $\vec{L}$ and $\vec{p}$ are constants for an isolated system since $\vec{N}^{\text{ext}} = 0$

and $\vec{F}^{\,\text{ext}} = 0$. Even though the two conditions $\vec{N}^{\,\text{ext}} = 0$ and $\vec{F}^{\,\text{ext}} = 0$ appear similar, there are some important differences for systems in which internal motion is possible. If $\vec{F}^{\,\text{ext}} = 0$, a center of mass which is at rest will remain so, regardless of internal forces or internal motion. If $\vec{N}^{\,\text{ext}} = 0$, the total angular momentum is constant, and if initially zero, will remain zero. However, $\vec{L} = 0$ does not exclude changes in orientation of the system by the use of merely internal forces. There is no rotational analog to the equation $\vec{r}(t) = (\vec{p}/M)t + \vec{r}(0)$ for linear motion of the center of mass.

Static equilibrium. In the design of permanent structures, the conditions under which a rigid body remains in steady motion under the action of a set of forces are of great importance. The six conditions for complete equilibrium of a rigid body are given in Eqs. (13). However, in many circumstances equilib-

$$\begin{aligned}
\vec{F}^{\,\text{ext}} &= \sum_i \vec{F}_i^{\,\text{ext}} = 0 \\
\vec{N}_{\text{CM}}^{\,\text{ext}} &= \sum_i (\vec{r}_i - \vec{R}) \times \vec{F}_i^{\,\text{ext}} = 0
\end{aligned} \qquad (13)$$

rium is desired only for a subset of the six independent directions of motion. To illustrate, the external force in the direction of motion of an accelerating automobile is nonzero, but equilibrium is maintained in all other directions. For rotational equilibrium, the torque about the center of mass of the automobile vanishes. However, by Eq. (12) the torque about another point O may differ from zero. *See* STATICS.

Single-axis rotations. Consider first the case of a rigid body that is constrained to rotate about a fixed axis. For definiteness, the z axis is chosen as the rotation axis, $\vec{\omega} = \omega_z \hat{z}$. The rotational velocity of a mass point m_i at a distance $\mathscr{R}_i = (x_i^2 + y_i^2)^{1/2}$ from the z axis is $v_\phi = \mathscr{R}_i \omega_z$, where ϕ is the azimuthal angle (**Fig. 1**). The angular momentum of this mass point

is given by Eq. (14). Summing over all masses, the

$$L_z = \mathscr{R}_i p_\phi = \mathscr{R}_i (m_i v_\phi) = m_i \mathscr{R}_i^2 \omega_z \qquad (14)$$

angular momentum of the rigid body is given by Eq. (15). The quantity of proportionality between L_z

$$L_z = \left(\sum_i m_i \mathscr{R}_i^2 \right) \omega_z \qquad (15)$$

and ω_z is called the moment of inertia, which will be denoted here by I_z. The moment-of-inertia expression for a continuous mass distribution is given in Eq. (16).

$$I_z = \int (x^2 + y^2) \rho(\mathscr{R}) dV \qquad (16)$$

The rotational motion about the z axis is accelerated by the external torque according to Eq. (10). The moment of inertia I_z is time-independent, since the perpendicular distance $\mathscr{R}$ from the rotation axis of a given mass point is fixed in the rigid body. Hence, equation of motion (17) is obtained.

$$N_z = I_z \dot{\omega}_z \qquad (17)$$

Moments of inertia about an axis through the center of mass of some simple bodies of uniform density are given in the **table**. The moment of inertia I_z' about the parallel axis is related to the moment of inertia I_z^{CM} by the parallel axis rule of Eq. (18), where d is the

$$I_z' = I_z^{\text{CM}} + M d^2 \qquad (18)$$

perpendicular distance between the two axes and M is the total mass of the rigid body. N_Z and L_Z have different values about the parallel axis than about the center of mass, as given by Eqs. (11) and (12).

Since Eq. (17) has the same mathematical structure as the equation for linear motion in one direction, $F_z = M \dot{v}_z$, the techniques for solving one-dimensional linear motion problems can also be applied to solve Eq. (17) for rotations about a fixed axis.

Energy conservation. If the external force on a rigid body is conservative, energy conservation methods

Moments of inertia of some simple bodies

Body	Axis through center of mass	Moment of inertia (I_{CM})
Rod, length l	Perpendicular to rod	$^1/_{12} M l^2$
Rectangular	Parallel to side b	$^1/_{12} M a^2$
plate, sides	Perpendicular to plate	$^1/_{12} M (a^2 + b^2)$
a, b		
Cube, sides a	Perpendicular to face	$^1/_6 M a^2$
Hoop, radius a	Perpendicular to plane	$M a^2$
Disk, radius a	Perpendicular to plane	$^1/_2 M a^2$
	Parallel to plane	$^1/_4 M a^2$
Solid cylinder, radius a	Along cylinder axis	$^1/_2 M a^2$
Spherical shell, radius a	Any axis	$^2/_3 M a^2$
Solid sphere, radius a	Any axis	$^2/_5 M a^2$

Fig. 1. Rotation of a symmetric rigid body around a fixed axis through the center of mass.

may be used in determining the motion. The conserved mechanical energy is $E = T + V$, where the kinetic energy T and the potential energy V are given in Eqs. (19), where $\vec{r}_s$ is a reference point which

$$T = \frac{1}{2} \int v^2 \rho(\vec{r}') \, dV' \qquad (19a)$$

$$V(\vec{r}) = -\int_{\vec{r}_s}^{\vec{r}} \vec{F}(\vec{r}') \cdot d\vec{r}' \qquad (19b)$$

specifies the zero of the potential. The velocity of a mass point in the rigid body is the vector sum of the center-of-mass velocity and the velocity of rotation about the center of mass, Eq. (20). By using this fact, the kinetic energy can be expressed by Eq. (21)

$$\vec{v}_i = \vec{V} + \mathscr{R}_i \omega_z \hat{\phi} \qquad (20)$$

$$T = \tfrac{1}{2} M V^2 + \tfrac{1}{2} I_z^{\text{CM}} \omega_z^2 \qquad (21)$$

as the sum of the translational kinetic energy and the rotational kinetic energy about an axis through the center of mass. The gravitational potential energy of a body near the surface of the Earth is $V(\vec{r}) = Mgh$, where h is the vertical distance of the body's center of mass from the Earth's surface and g is the gravitational acceleration. *See* CONSERVATION OF ENERGY; ENERGY.

Impulsive motion. For forces such as a sharp blow that act only during a very short time, it is convenient to use an integrated form of the laws of motion. The translational motion of the center-of-mass point is determined by Eq. (6). When both sides of this equation are integrated with respect to t over the time interval $\Delta t = t_1 - t_0$, during which the force acts, Eq. (22) is obtained. The time integral of the force

$$\Delta \vec{P} = \vec{P}^1 - \vec{P}^0 = \int_{t_0}^{t_1} \vec{F}^{\text{ext}} \, dt \qquad (22)$$

on the right is called the impulse. For angular motion the integrated form of the equation of motion, Eq. (23), results. The time integral of the torque is

$$\Delta \vec{L} = \vec{L}^1 - \vec{L}^0 = \int_{t_0}^{t_1} \vec{N}^{\text{ext}} \, dt \qquad (23)$$

called the angular impulse. For rotations of a rigid body about a symmetry axis which passes through the center of mass, the angular-impulse equation can be expressed as in Eq. (24), where z is the symmetry axis.

$$\Delta L_z = I_z^{\text{CM}}(\omega_z^1 - \omega_z^0) = \int_{t_0}^{t_1} N_z^{\text{ext}} \, dt \qquad (24)$$

See IMPACT; IMPULSE (MECHANICS).

The dynamics of billiard shots is an example of the usefulness of the impulse formulation of the equations of motion. For simplicity, consider shots in which the cue hits the ball in its vertical median plane in a horizontal direction. The cue imparts an impulse to the stationary ball at a vertical distance h above the
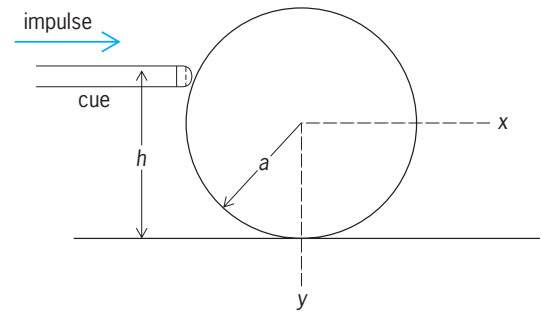

Fig. 2. Impulse imparted to a billiard ball by a cue stick. (*After V. Barger and M. Olsson, Classical Mechanics: A Modern Perspective, McGraw-Hill, 1973*)

table (**Fig. 2**). The linear impulse is given in Eq. (25),

$$M \Delta V_x = M V_x^1 = \int_{t_0}^{t_1} F_x \, dt \qquad (25)$$

where V_x^1 is the velocity of the center of mass just after impact. The angular impulse about the z axis which passes through the center of mass of the ball is given in Eq. (26), where a is the radius of the

$$\Delta L_z = I_z^{\text{CM}} \omega_z^1 = \int_{t_0}^{t_1} (h - a) F_x \, dt \qquad (26)$$

ball. By elimination of the force integral and substitution of the sphere moment of inertia, the relation of Eq. (27) is obtained between the spin and velocity of

$$\omega_z^1 = \frac{5}{2} \left(\frac{h - a}{a^2} \right) V_x^1 \qquad (27)$$

the ball immediately after the impulse. The velocity of the ball at the point of contact with the table is given by Eq. (28). If the ball is to roll without slipping

$$V_c = V_x^1 - a\omega_z^1 = V_x^1 \left(\frac{7a - 5h}{2a} \right) \qquad (28)$$

($V_c = 0$), the height at which the ball must be hit is given by Eq. (29). Only if this condition is satisfied

$$h = \frac{7}{5} a \qquad (29)$$

does pure rolling take place from the very start.

General rotations. A rotation is a motion such that a point i located at $\vec{r}_i$ has a constant distance $|\vec{r}_i - \vec{r}_O|$ to each point O on a line, the axis of rotation (**Fig. 3**). The path traced out by the point i is a circle whose symmetry axis is the axis of rotation. If the z axis is chosen to be parallel to the rotation axis, the circular path lies in the x, y plane and the vector $\vec{r} - \vec{r}_O$ is given by Eq. (30). Here a is the radius of the circle,

$$\vec{r} - \vec{r}_O = a \cos \phi \hat{x} + a \sin \phi \hat{y} + b \hat{z} \qquad (30)$$

ϕ is the rotation angle, and b is the distance from O to the center of the circle. The time derivative of Eq. (30) gives the velocity of point i in Eq. (31), which

$$\vec{v} = \dot{\phi}(-a \sin \phi \hat{x} + a \cos \phi \hat{y}) \qquad (31)$$

is perpendicular to $\vec{r} - \vec{r}_O$. Defining an angular velocity vector $\vec{\omega} \equiv \dot{\phi}\,\hat{z}$, the velocity of rotation can be written as in Eq. (32). Equation (32) is generally valid

$$\vec{v} = \vec{\omega} \times (\vec{r} - \vec{r}_O) \qquad (32)$$

for any coordinate system if $\vec{\omega}$ is defined as $\vec{\omega} \equiv \omega\hat{n}$ with the unit vector $\hat{n}$ directed along the rotation axis. If the point O is moving with a translational velocity $\vec{v}_O$ relative to a fixed reference frame, the velocity of the point i in the fixed frame is the vector sum of the translational velocity $\vec{v}_O$ and $\vec{\omega}$ may change with time. The expression, Eq. (33), is the

$$\vec{v}(r) = \vec{v}_O + \vec{\omega} \times (\vec{r} - \vec{r}_O) \qquad (33)$$

most general for the velocity of the points of a rigid body. The vectors $\vec{v}_O$ and $\vec{\omega}$ amount to six velocity components, agreeing with the number of coordinates of a rigid body.

As a single-axis example, consider a wheel of radius R which rolls without slipping on a level surface (**Fig. 4**). For no slipping, $dx = R\,d\theta$, where dx and $d\theta$ are infinitesimal horizontal and angular displacements, respectively. Dividing both sides by the time interval dt, $\vec{v}_O = R\omega\hat{x}$, where $\vec{v}_O$ is the velocity of the center of mass. The velocity of a point on the wheel
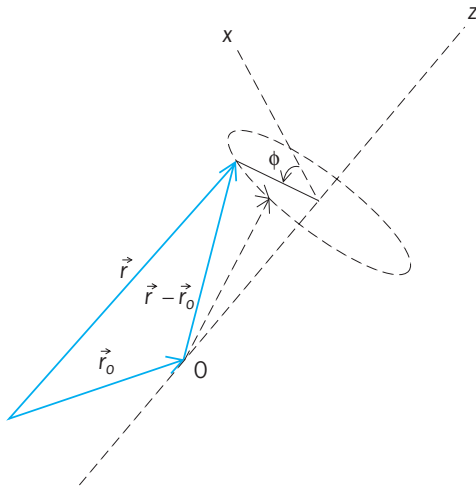
Fig. 3. Rotation of a point located at $\vec{r}$ about the *z* axis.
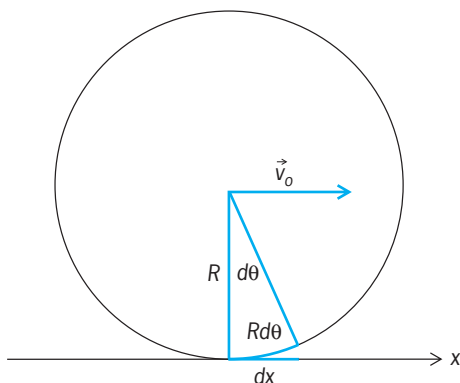
Fig. 4. Wheel rolling without slipping on a level surface. (*After V. Barger and M. Olsson, Classical Mechanics: A Modern Perspective, McGraw-Hill, 1973*)
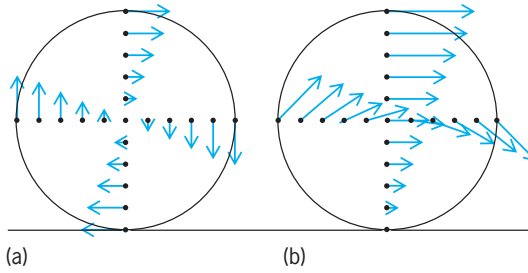
Fig. 5. Velocity of points on a wheel (*a*) relative to the center of mass and (*b*) relative to the ground. (*After V. Barger and M. Olsson, Classical Mechanics: A Modern Perspective, McGraw-Hill, 1973*)

relative to the center of mass is given in Eq. (34),

$$\vec{v}_{\text{rot}} = \vec{\omega} \times \vec{r} = v_O \hat{\omega} \times \left(\frac{\vec{r}}{R}\right) \qquad (34)$$

where $\hat{\omega} = \vec{\omega}/\omega$. The rotational velocities of points on the wheel are illustrated in **Fig. 5**. With respect to a reference frame at rest on the level surface, the velocity of a point on the wheel is given by Eq. (35).

$$\vec{v} = \vec{v}_O + \vec{v}_{\text{rot}} = \vec{v}_O + v_O \hat{\omega} \times \left(\frac{\vec{r}}{R}\right) \qquad (35)$$

The velocity $\vec{v} = 0$ at the point of contact, since no slipping means that there is no relative motion of the wheel and surface at the contact point.

Moments and products of inertia. The dynamics of rigid-body rotations are contained in Eq. (10), which relates the time rate of change of the total angular momentum to the external torque. The angular momentum $\vec{L}$ about a point O can be computed in terms of the angular velocity $\vec{\omega}$ in Eq. (21). The location of a point mass m_i in the rigid body relative to O is hereafter denoted by $\vec{r}_i$. The velocity of the mass m_i relative to O is then $\vec{v}_i = \vec{\omega} \times \vec{r}_i$, and the angular momentum about O is expressed by Eq. (36). The triple

$$\vec{L} = \sum_i m_i (\vec{r}_i \times \vec{v}_i) = \sum_i m_i \vec{r}_i \times (\vec{\omega} \times \vec{r}_i) \qquad (36)$$

cross-product can be expanded by using the identity $\vec{a} \times (\vec{B} \times \vec{C}) = \vec{B}(\vec{a} \cdot \vec{C}) - \vec{C}(\vec{a} \cdot \vec{B})$, leading to the result in Eq. (37). It can be seen that the angular

$$\vec{L} = \vec{\omega} \left(\sum_i m_i |\vec{r}_i|^2 \right) - \sum_i m_i \vec{r}_i (\vec{r}_i \cdot \vec{\omega}) \qquad (37)$$

momentum vector $\vec{L}$ will not necessarily be parallel to the angular velocity vector $\vec{\omega}$.

In cartesian components of $\vec{\omega} = (\omega_x, \omega_y, \omega_z)$ and $\vec{r} = (x, y, z)$, the components of $\vec{L} = (L_x, L_y, L_z)$ have the form of Eqs. (38). The coefficients of the angu-

$$L_x = I_{xx}\omega_x + I_{xy}\omega_y + I_{xz}\omega_z$$

$$L_y = I_{yx}\omega_x + I_{yy}\omega_y + I_{yz}\omega_z \qquad (38)$$

$$L_z = I_{zx}\omega_x + I_{zy}\omega_y + I_{zz}\omega_z$$

lar velocity components are written in Eqs. (39). The three quantities I_x, I_y, and I_z are known as moments of inertia and the other six quantities are called products of inertia. For a continuous mass distribution

$$I_x \equiv I_{xx} = \sum_i m_i(y_i^2 + z_i^2)$$
$$I_y \equiv I_{yy} = \sum_i m_i(x_i^2 + z_i^2)$$
$$I_z \equiv I_{zz} = \sum_i m_i(x_i^2 + y_i^2)$$
$$I_{xy} = I_{yx} = -\sum_i m_i x_i y_i \tag{39}$$
$$I_{xz} = I_{zx} = -\sum_i m_i x_i z_i$$
$$I_{yz} = I_{zy} = -\sum_i m_i y_i z_i$$

three quantities I_x, I_y, and I_z are known as moments of inertia and the other six quantities are called products of inertia. For a continuous mass distribution

$$\sum_i m_i$$

is replaced by

$$\int dm$$

where $dm = \rho dV$ as in Eq. (3). The moments and products of inertia will be denoted henceforth by I_{jk}, where the subscripts j, k take on the values x, y, z. The products of inertia are symmetric: $I_{jk} = I_{kj}$. *See* MOMENT OF INERTIA; PRODUCT OF INERTIA.

The nine quantities I_{jk} form a symmetric tensor which can be written as the 3×3 matrix in Eq. (40).

$$\mathbf{I} \equiv \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{pmatrix} \tag{40}$$

In matrix notation, $\vec{L}$, $\vec{\omega}$, and $\vec{N}$ are written as column vectors, with their x, y, and z components forming the elements, as in Eq. (41). Then the an-

$$\vec{L} = \begin{pmatrix} L_x \\ L_y \\ L_z \end{pmatrix} \quad \vec{\omega} = \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} \quad \vec{N} = \begin{pmatrix} N_x \\ N_y \\ N_z \end{pmatrix} \tag{41}$$

gular momentum and the equation for rigid-body rotational motion can be written in compact form as in Eqs. (42).

$$\vec{L} = \mathbf{I} \cdot \vec{\omega} \tag{42a}$$

$$\vec{N} = \frac{d\vec{L}}{dt} = \frac{d}{dt}(\mathbf{I} \cdot \vec{\omega}) \tag{42b}$$

See MATRIX THEORY.

The kinetic energy of a rigid body can likewise be expressed in terms of the moments and products of inertia. The steps leading to this form for T are given in Eq. (43). In tensor notation the

kinetic energy is then given by Eq. (44).

$$T = \frac{1}{2} \sum_i m_i \vec{v}_i \cdot \vec{v}_i$$
$$= \frac{1}{2} \sum_i m_i (\vec{\omega} \times \vec{r}_i) \cdot (\vec{\omega} \times \vec{r}_i)$$
$$= \frac{1}{2}\vec{\omega} \cdot \left[\sum_i m_i \vec{r}_i \times (\vec{\omega} \times \vec{r}_i) \right] = \frac{1}{2}\vec{\omega} \cdot \vec{L} \tag{43}$$

$$T = \frac{1}{2}\vec{\omega} \cdot \mathbf{I} \cdot \vec{\omega} = \frac{1}{2} \sum_{j,k} \omega_j I_{jk} \omega_k \tag{44}$$

Euler's equations. The moments and products of inertia I_{jk} relative to the fixed coordinate system change as a function of time as the rigid body rotates, making the description of the motion directly from Eq. (40) difficult. Great simplification can often be achieved by choosing instead a coordinate system that rotates with the rigid body. In the rotating reference frame the moments and products of inertia are time-independent. The equation of motion with respect to the moving "body" axes is given by Eq. (45).

$$\vec{N} = \frac{\delta\vec{L}}{\delta t} + \vec{\omega} \times \vec{L} \tag{45}$$

The time derivative $d\vec{L}/dt$ in a fixed reference frame consists of a part $\delta\vec{L}/\delta t$ from the time rate of change of $\vec{L}$ relative to the axes of the moving frame, and part $\vec{\omega} \times \vec{L}$ from the rotation of these axes relative to the fixed axes.

It is always possible to make a choice of axes in the body so that all the products of inertia vanish (that is, $I_{jk} = 0$ for $j \neq k$), thus making the moment of inertia matrix diagonal. These axes are called the principal axes of the rigid body. The angular momentum components along these axes (which are labeled 1, 2, 3) are given in terms of the angular velocity components along these axes and the principal moments of inertia I_1, I_2, I_3 by Eqs. (46). The resulting Euler's

$$L_1 = I_1\omega_1$$
$$L_2 = I_2\omega_2 \tag{46}$$
$$L_3 = I_3\omega_3$$

equations of motion for a rigid body in terms of principal-axes components of $\vec{N}$ and $\vec{\omega}$ are given in Eqs. (47). These equations are a convenient starting

$$N_1 = I_1\dot{\omega}_1 + (I_3 - I_2)\omega_3\omega_2$$
$$N_2 = I_2\dot{\omega}_2 + (I_1 - I_3)\omega_1\omega_3 \tag{47}$$
$$N_3 = I_3\dot{\omega}_3 + (I_2 - I_1)\omega_2\omega_1$$

point for many discussions of rigid-body rotations. *See* EULER'S EQUATIONS OF MOTION.

Poinsot's method. The motion of a freely rotating rigid body can be described by a geometrical approach known as Poinsot's construction, which is based on conservation of kinetic energy and angular momentum. In the coordinate system defined by the principal axes, the kinetic energy T is given by

Eq. (48). A vector $\vec{\rho}$ is defined as in Eq. (49). Rewrit-

$$2T = I_1\omega_1^2 + I_2\omega_2^2 + I_3\omega_3^2 \qquad (48)$$

$$\vec{\rho} \equiv \frac{\vec{\omega}}{\sqrt{2T}} \qquad (49)$$

ing Eq. (48) in terms of the components of $\vec{\rho}$, Eq. (50) is obtained. In the space specified by the

$$I_1\rho_1^2 + I_2\rho_2^2 + I_3\rho_3^2 = 1 \qquad (50)$$

coordinates (ρ_1,ρ_2,ρ_3), this describes the surface of an ellipsoid, known as the inertial ellipsoid. The inertial ellipsoid is imagined to be rigidly attached to the body and rotating with it. Taking the gradient of the left-hand side of Eq. (50) with respect to $\vec{\rho}$ gives the direction of the normal to the ellipsoid surface. According to Eq. (51), the normal direction is

$$\nabla_{\vec{\rho}}(I_1\rho_1^2 + I_2\rho_2^2 + I_3\rho_3^2) = 2\frac{\vec{L}}{\sqrt{2T}} \qquad (51)$$

along the constant vector $\vec{L}$. Thus the tangent plane to the ellipsoid at $\vec{\rho}$ gives the direction of the normal to the ellipsoid surface. Furthermore, the dot product $\vec{\rho} \cdot \vec{L}/L$ is also constant, as demonstrated in Eq. (52).

$$\frac{\vec{\rho} \cdot \vec{L}}{L} = \frac{2T}{L} \qquad (52)$$

Since $\vec{\rho} \cdot \vec{L}/L$ is the distance from the center of the ellipsoid to the tangent plane, the center of the ellipsoid remains at a constant height above the plane. Consequently the force-free motion of the rigid body is represented by the rolling motion of the inertial ellipsoid on the tangent plane, with the center of mass at a fixed height above the tangent plane (**Fig. 6**). The point of contact with the plane is specified by $\vec{\rho}$, which is along the direction of the rotation axis. Thus the point of contact is at rest and the ellipsoid rolls without slipping. The tangent plane is known as the invariable plane, the curve traced out by $\vec{\rho}$ on the inertial ellipsoid is called the polhode, and the corresponding curve on the invariable plane is called the herpolhode.

Tennis racket theorem. The solution of Euler's equation's for a free rigid body with unequal principal moments of inertia can be simply illustrated with
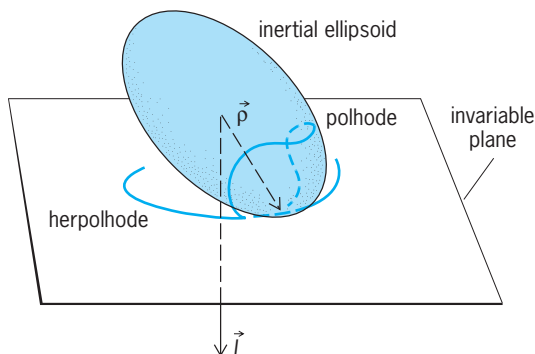
Fig. 6. Rolling motion of inertial ellipsoid on the tangent plane describing the force-free motion of a rigid body.
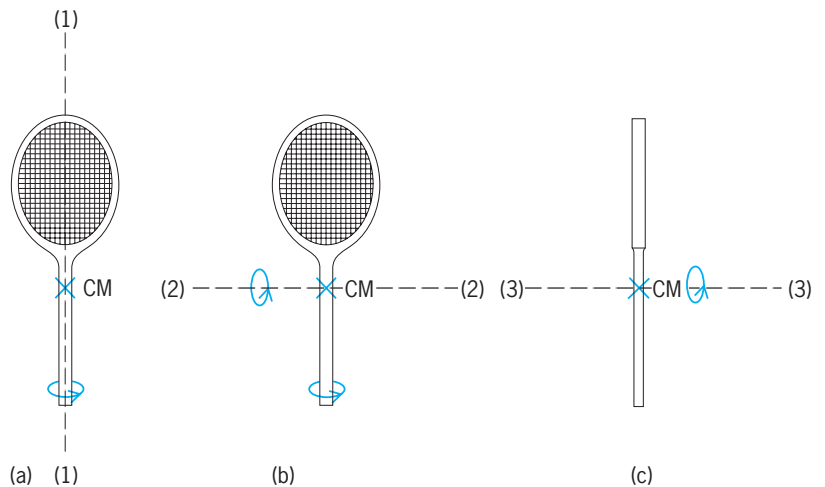

Fig. 7. Principal axes of a tennis racket (*a*) along the handle, (*b*) perpendicular to the handle in the plane of the strings, and (*c*) perpendicular to the handle and strings. (*After V. Barger and M. Olsson, Classical Mechanics: A Modern Perspective, McGraw-Hill, 1973*)

a tennis racket. A tennis racket has three principal axes (**Fig. 7**); the moments of inertia around these three principal axes are ordered as in Eq. (53). In a

$$I_1 < I_2 < I_3 \qquad (53)$$

coordinate system with the racket in the x, y plane and x and y axes along the (1) and (2) principal axes, the principal moments of inertia are given by Eqs. (54), where σ is the mass per unit area. Thus the

$$I_1 = \int y^2\sigma \, dx \, dy$$

$$I_2 = \int x^2\sigma \, dx \, dy \qquad (54)$$

$$I_3 = \int (x^2+y^2)\sigma \, dx \, dy$$

three moments of inertia for this plane body satisfy the relation of Eq. (55).

$$I_3 = I_1 + I_2 \qquad (55)$$

Suppose that a racket is tossed into the air with a spin about one of the principal axes. In the approximation that the Earth's gravity is a uniform force, there are no gravitational torques about the center of mass of the racket, and Euler's equations with $\vec{N} = 0$ apply. They can be written as in Eqs. (56),

$$\dot{\omega}_1 + \omega_3\omega_2 = 0$$
$$\dot{\omega}_2 - \omega_1\omega_3 = 0 \qquad (56)$$
$$\dot{\omega}_3 + K\omega_2\omega_1 = 0$$

where $K = (I_2 - I_1)/(I_2 + I_1)$; for a tennis racket $I_2 > I_1$ and thus K is positive.

If the initial spin is nearly around the axis (1), the product $\omega_3\omega_2$ is small and ω_1 is nearly constant. Then by time differentiation and substitution, equations for ω_2 and ω_3 separately can be found; they are of simple harmonic form, Eqs. (57). These equations

have the solution given in Eqs. (58). By the initial

$$\ddot{\omega}_2 + K\omega_1^2\omega_2 = 0$$
$$\ddot{\omega}_3 + K\omega_1^2\omega_3 = 0 \tag{57}$$

$$\omega_2(t) = a \sin\left(\sqrt{K}\omega_1 t + \alpha\right)$$
$$\omega_3(t) = a \cos\left(\sqrt{K}\omega_1 t + \alpha\right) \tag{58}$$

conditions the amplitude a is small and the ω_2 and ω_3 components of the angular velocity remain small. Thus when the tennis racket is tossed into the air with a spin nearly about the (1) axis, the racket continues to spin nearly about this axis and can be easily recaught. A similar stability of the motion applies for axis (3).

For initial motion nearly about axis (2), the second-order differential equations for ω_1 and ω_3 are given in Eqs. (59). Here the solutions for $\omega_1(t)$ and $\omega_3(t)$ grow

$$\ddot{\omega}_1 - K\omega_2^2\omega_1 = 0$$
$$\ddot{\omega}_3 - K\omega_2^2\omega_3 = 0 \tag{59}$$

exponentially with time, and the motion quickly develops spin around all three principal axes, which makes it difficult to recatch the racket. The instability of rotations about the principal axis with the intermediate moment of inertia (I_2) and the stability of rotations about the other two principal axes is true for any free rigid body. *See* DIFFERENTIAL EQUATION; HARMONIC MOTION.   Vernon D. Barger

Bibliography. V. Barger and M. Olsson, *Classical Mechanics: A Modern Perspective*, 2d ed., McGraw-Hill, 1995; H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, 3d ed., Addison-Wesley, 2002; D. Halliday, R. Resnick, and K. S. Krane, *Physics*, 5th ed., Wiley, 2002; K. R. Symon, *Mechanics*, 3d ed., Addison-Wesley, 1971.

Rinderpest

An acute or subacute, contagious viral disease of ruminants and swine, manifested by high fever, lachrymal discharge, profuse diarrhea, erosion of the epithelium of the mouth and of the digestive tract, and high mortality.

Infectious agent. Rinderpest (also known as cattle plague) is caused by a ribonucleic acid (RNA) virus classified in the genus *Morbillivirus* within the Paramyxoviridae family. This virus is closely associated with the viruses of human measles, peste des petits ruminants of sheep and goats, canine distemper, phocine distemper, and the dolphin morbillivirus. Although only one serotype of rinderpest virus is known, the various field strains, grouped by molecular characterization into three lineages known as the Asian lineage and African lineages 1 and 2, vary significantly in virulence. The rinderpest virus is easily inactivated by heat and survives outside the host for a short time. Therefore, transmission of rinderpest is by direct contact between animals. *See* ANIMAL VIRUS; MEASLES; PARAMYXOVIRUS.

Pathogenesis. Rinderpest is one of the oldest diseases known to affect cattle. It is characterized by the development of high fever that lasts for several days until just prior to death. Animals become depressed and have profuse clear serous nasal and lachrymal secretions. The mucosal membranes of the eyes and mouth become congested, and as the disease progresses, there is development of severe erosions on the inside of lips, gums, tongue, and hard and soft palate areas. Cattle suffering from rinderpest develop a very severe, profuse diarrhea and a drop in the number of white blood cells. The morbidity in susceptible cattle and buffalo is greater than 90%, with death of almost all clinically affected animals. *See* DIARRHEA.

The pathological characteristics of rinderpest are related to the ability of the virus to grow in lymphoid and epithelial tissues. In addition to the oral erosions, there is a destruction of epithelium over aggregated lymphoid follicles (Peyer's patches) in the submucosa of the small intestine, and mucosal areas (rich in lymphoid cells) in the large intestine and colon. *See* LYMPHATIC SYSTEM.

Epidemiology. Rinderpest affects virtually all cloven-hoofed animals, both domesticated and wild. Cattle, Asian buffaloes, yaks, African buffaloes, lesser kudus, elands, warthogs, and giraffes are particularly susceptible. Wildlife does not appear to act as a long-term reservoir of infection independently of cattle. All swine are susceptible and may suffer severe disease; they have been important in disease transmission in the Far East. Sheep and goats undergo infection but rarely exhibit clinical signs. Earlier confusing reports of rinderpest in Indian sheep are now mainly considered to have referred to the related virus disease "peste des petits ruminants." Strains of virus causing only mild clinical reactions have been reported, as have others causing peracute (severe and brief) disease. The clinical signs of disease appear in as little as 3 days after exposure to strains causing acute, severe disease, however, the incubation period can extend to 15 days for milder strains. In most fatal cases the animal collapses and dies within 6 to 12 days after the onset of fever; otherwise a protracted convalescence ensues. If the animal survives, it will develop a life-long sterile, solid immunity; there is no carrier state.

Immunity and prevention. Modified (attenuated), live rinderpest vaccines generate a strong, protective immune response that lasts for life. One such vaccine developed in East Africa has been widely used in Africa and Asia and has proved safe and efficacious in large-scale vaccination campaigns and outbreak control. Another vaccine used for decades in Russia might have reverted to virulence on several occasions. A vaccinia recombinant vaccine has been described but has not been licensed for general use. There is now growing confidence that rinderpest has been eliminated from Asia. The last reservoirs of infection persisted in Yemen until 1997 and in the Indus River buffalo tract of Pakistan until late 2000. Thus, the Asian lineage of rinderpest virus is now almost certainly extinct.

Similarly African lineage 1 persisted in southern Sudan until 2001 but was eliminated by vaccination of all cattle belonging to the last tribe to be affected.

The rest of Africa is also now considered to be free from rinderpest except, possibly, for a relatively small area of the Somali pastoral ecosystem spanning the border of Somalia and Kenya, where a reservoir of African lineage 2 could still be present in mild form. However, even there the last confirmation of virus presence was in 2001 in wild African buffaloes in a Kenyan national park. Investigations continue to determine whether or not rinderpest is still present.

In the rest of the world, progress is being made in accreditation of rinderpest freedom by the Office International des Epizooties. The Global Rinderpest Eradication Programme hosted by the United Nations Food and Agriculture Organization has provided international coordination since 1993. *See* VACCINATION.   Alfonso Torres; Peter Roeder

Bibliography. E. P. J. Gibbs (ed.), *Virus Diseases of Food Animals*, vols. 1 and 2, 1981; W. Plowright, Rinderpest, *Ann. N.Y. Acad. Sci.*, 101:327–382, 1962; P. L. Roeder and W. P. Taylor, Rinderpest in Veterinary Clinics of North America, Food Animal Practice, *Animal Health Emergency Diseases*, 18:515–547, 2002; P. B. Rossiter, Rinderpest in J. A. W. Coetzer et al. (eds.), *Infectious Diseases of Livestock with Special Reference to Southern Africa*, Chap. 74, Oxford University Press, pp. 735–757, 1994; G. R. Scott Rinderpest, *Adv. Vet. Sci.*, 9:113–124, 1964.

Ring

A tie member or chain link. Tension or compression applied through the center of a ring produces bending moment, shear, and normal force on radial sections. Because shear stress is zero at the boundaries of the section where bending stress is maximum, it is usually neglected. A quadrant of the ring is a curved bar with moments M_A and M_B at the ends (see **illus.**).

An approximate solution for moments at the ends of the quadrant neglects curvature and employs the procedures used for statically indeterminate straight beams. The numerically larger moment occurs at the load section A where $M_A = 0.318PR$ and $M_B = 0.181PR$, where P is load and R is radius. The variation of moment is found by statics in terms of angle θ defining the section. For tensile load, $M_x = PR(0.318 - \frac{1}{2}\sin\theta)$. Moment at A tends to increase the curvature; the ring becomes flatter at B. The end moments have opposite signs and $M = 0$ near $\theta = 39°$.

Important stresses occur at the inside and outside of sections where the normal and bending stresses reach maximum values. At any section, the relationship is as shown in Eq. (1), where S_b is found by

$$S = S_b + \frac{P\sin\theta}{2A} \qquad (1)$$

curved bar theory and A is sectional area. Or stresses can be calculated from the corrected straight beam formula, shown in Eq. (2), where K depends on end

$$S = K\frac{M_c}{I} + \frac{P\sin\theta}{2A} \qquad (2)$$
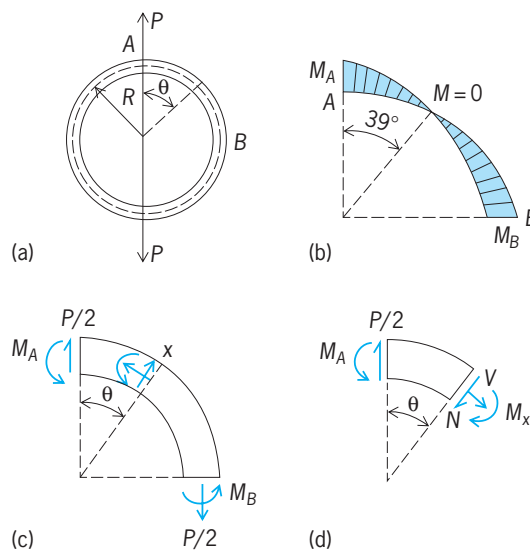
Stresses in rings. (a) Entire ring, showing relation of quadrant **AB** and angle θ to load **P** and radius **R**.
(b) Stresses on quadrant **AB**. **(c)** Variation of moment with angle θ. **(d)** Stresses on arbitrary section.

restraints, I is moment of inertia, and c is distance from centroids to extreme element. The greatest stress (compressive for tensile load) occurs at the inside of the section where load is applied. Where the greatest tensile stress occurs depends on the ring dimensions. For a thick ring, whose radius of curvature is small compared to the dimensions, maximum tension may occur at top and bottom sections, whereas for thin rings with relatively large radius of curvature, greatest tension is at the side sections perpendicular to the load line. Thin rings subjected to external hydrostatic pressure may fail by compressive buckling. Deflection of the ring is most conveniently found by energy methods, which include the moments, shearing, and normal forces acting in the ring. Bending has greater influence in thin rings.
                                    W. J. Krefeld; W. G. Browman

Ring-opening polymerization

The formation of macromolecules from cyclic monomers such as cyclic hydrocarbons, ethers, esters, amides, siloxanes, and sulfur (eight-membered ring). Thus, ring-opening polymerization is of particular interest, since macromolecules of almost any chemical structure can be prepared.
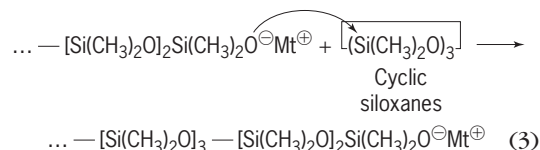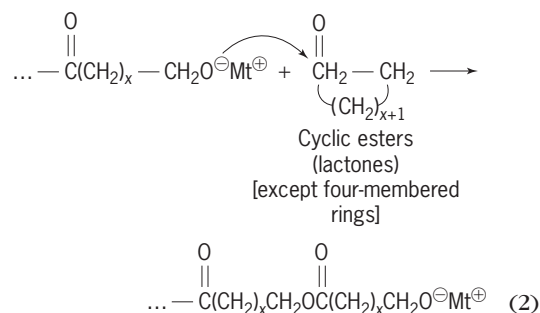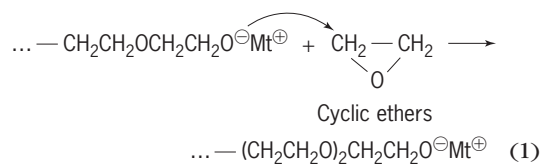
Polymerization is initiated with the breaking of a single (sigma) bond in the cyclic monomer. The driving force of ring-opening polymerization comes from the strain of the rings. The major source of the ring strain is the angular strain, with the release of the strain being energetically favorable.

As the ring size increases, ring strain decreases, although this change is not absolute. Some cyclic monomers, such as six-membered cyclic ethers, are virtually strain-free, and their polymerization to a high-molecular-weight polymer is not possible. By contrast, six-membered cyclic esters and siloxanes

(for example, hexamethylcyclotrisiloxane) polymerize easily, due to their particular conformation. Conformational strain (related to the opposition of hydrogen atoms across the rings) is responsible for the strain for some of the larger cyclic monomers.

The majority of heterocyclic monomers polymerize by ionic mechanisms. That is, either anions or cations located at the end of a growing macromolecule attack the monomer molecule, breaking a bond between the heteroatom (for example, nitrogen or oxygen) and the adjacent atom so that both bonding electrons remain with one of the atoms (heterolytic cleavage). *See* HETEROCYCLIC COMPOUNDS; POLYMER; POLYMERIZATION.

Anionic polymerization. In the anionic polymerization of cyclic ethers or esters, mostly alkoxide anions are involved. Examples of monomers and propagating species are shown in reactions (1)–(3), where

$$\ldots - CH_2CH_2OCH_2CH_2O^{\ominus}Mt^{\oplus} + \underset{O}{CH_2 - CH_2} \longrightarrow$$

Cyclic ethers

$$\ldots - (CH_2CH_2O)_2CH_2CH_2O^{\ominus}Mt^{\oplus} \quad (1)$$

$$\ldots - \underset{O}{\overset{O}{\underset{\|}{C}}}(CH_2)_x - CH_2O^{\ominus}Mt^{\oplus} + \underset{(CH_2)_{x+1}}{\overset{O}{\underset{\|}{CH_2 - CH_2}}} \longrightarrow$$

Cyclic esters
(lactones)
[except four-membered rings]

$$\ldots - \underset{O}{\overset{O}{\underset{\|}{C}}}(CH_2)_x CH_2 O \underset{O}{\overset{O}{\underset{\|}{C}}}(CH_2)_x CH_2 O^{\ominus}Mt^{\oplus} \quad (2)$$

$$\ldots - [Si(CH_3)_2O]_2Si(CH_3)_2O^{\ominus}Mt^{\oplus} + [(Si(CH_3)_2O)_3] \longrightarrow$$

Cyclic siloxanes

$$\ldots - [Si(CH_3)_2O]_3 - [Si(CH_3)_2O]_2Si(CH_3)_2O^{\ominus}Mt^{\oplus} \quad (3)$$

Mt^+ denotes metal cations, mostly Na^+ or K^+, and less often divalent cations.

The polymerization of some of these monomers (for example, ethylene oxide or some lactones) are called "living polymerizations" because there is no termination to stop chain growth.

Polymers of ethylene oxide find numerous applications. Very high molecular weight polymer, up to several million, is used to increase the fluidity of water. Water containing a fraction of a percent of poly(ethylene oxide) can be pumped (for example, by firefighters) with much less energy and therefore for much longer (higher) distances. Lower-molecular-weight products are used as water-soluble surface-active agents. In addition, poly(ethylene oxide) is biocompatible, and it is finding applications in biomedicine, such as drug delivery systems and gels (highly swelling in water) used for healing wounds. *See* POLYETHER RESINS; POLYETHYLENE GLYCOL.

Polysiloxanes are very versatile polymers used in engineering applications due to their outstanding heat resistance, as well as in biomedicine because of their biocompatibility. Silicone rubber (sealant) is another popular product. It can be made as a liquid which solidifies on contact with air. *See* SILICONE RESINS.

Another technically important family of polymers made by anionic ring-opening polymerization is based on ε-caprolactam. Polyamides derived from this monomer are used as fibers for textiles known as nylon (for example, Nylon 6). *See* MANUFACTURED FIBER; POLYAMIDE RESINS.

Coordinated polymerization. The most recently developed living polymerizations of cyclic ethers and cyclic esters are based on multicenter concerted mechanisms of propagation. This mechanism of coordinated polymerization was elaborated in Lodz, Poland, and is described in reaction (4) for the poly-

(Ground state of the deaggregated, actually growing species)   (ε-Caprolactone)

(Transition state)

(Ground state)   (4)
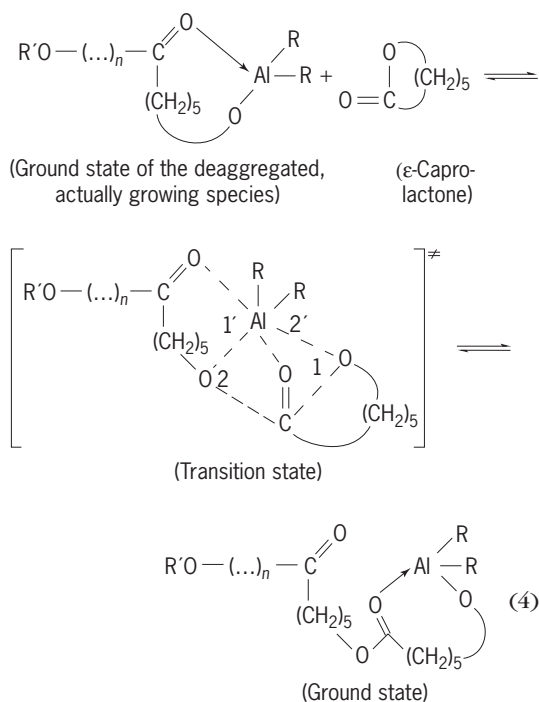
merization of ε-caprolactone initiated with dialkylalkoxyaluminum (R and $R' = C_2H_5$), where in a concerted manner two bonds (1 and 1′) are broken and simultaneously two new bonds (2 and 2′) are formed. The coordinated propagating species are less reactive than ions and therefore are much more selective. Alkoxides of iron (Fe), tin (Sn), titanium (Ti), and several others polymerize cyclic esters in the same way. Only with these initiators is it possible to avoid formation of the unwanted cyclic oligomers, which notoriously contaminate the desired linear macromolecules.

The most comprehensively studied cyclic ester is L,L-dilactide, giving the novel industrial polymer poly(L-lactide) [reaction (5)]. It meets several criteria of future polymers; namely, it is based on renewable starting materials, and its biodegradation products

L-Lactic acid   L,L-Dilactide

Poly(L-lactide)   (5)
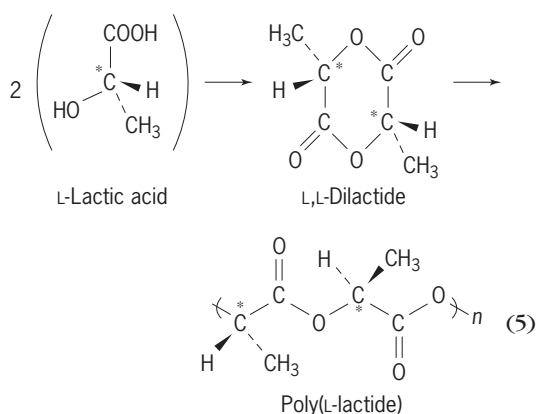
are water and carbon dioxide. Municipal waste is used as the starting material and converted in a high yield into L,L-dilactide. L,L-dilactide is a six-membered ring with two chiral centers in the molecule, and is prepared from L-lactic acid, a product of bacterial fermentation of mono- or polysaccharides.

The most common initiator is tin octoate [$Sn(Oct)_2$ or $Sn(OC(O)CH(CH_2CH_3)(CH_2)_3CH_3)$]. It has recently been shown that $Sn(Oct)_2$ is not an initiator by itself, and like other carboxylates reacts with impurities present in monomer (for example, with lactic acid or water) and gives tin alkoxide, which initiates and then propagates in the same way, as described above for $ROAlR_2$ (that is, on the $-SnOR$ bond).

Although L,L-dilactide is a six-membered cyclic compound that has relatively low ring strain, its thermodynamics of polymerization are favorable enough to ensure high monomer conversion at the polymer-monomer equilibrium. The polymerization proceeds with almost complete retention of configuration (at the stereocenter), in contrast to the anionic polymerization, in which substantial loss of optical activity (racemization) occurs.

The mechanical properties of poly(L-lactide) [containing approximately 5% of D,L-dilactide units] resemble those of polystyrene. Poly(L-lactide) is colorless and transparent, and is mostly used as a commodity plastic for films and fibers. When prepared with special precautions and at the high purity conditions, poly(L-lactide) can be used as a biocompatible material (for example, surgical sutures) whenever bioresorption is required.

In the recent years, the number of published works on, and related to, poly(L-lactide) has been fast increasing, reaching 500 papers in the year 2002.

Cationic ring-opening polymerization. A majority of the monomers that polymerize by anionic mechanisms can also be polymerized cationically. However, several monomers that undergo cationic ring-opening polymerization do not polymerize anionically, such as cyclic acetals, some cyclic ethers (oxetanes, tetrahydrofurans and larger rings), cyclic 1,3-oxaza monomers (for example, oxazolines), and cyclic imines (for example, aziridines). Cationic ring-opening polymerization can be initiated by protonic acids, carbocations, onium ions, or Friedel-Crafts catalysts, most often initiating in the form of complex acids with water. Polymerization proceeds with

onium ions as active species, shown in notations (6) and (7). Polyoxazolines upon hydrolysis can be converted into the linear polyamines.

Polytetrahydrofuranium ion   Tetrahydrofuran

(6)

Polyoxazolinium ion   Oxazoline

(7)
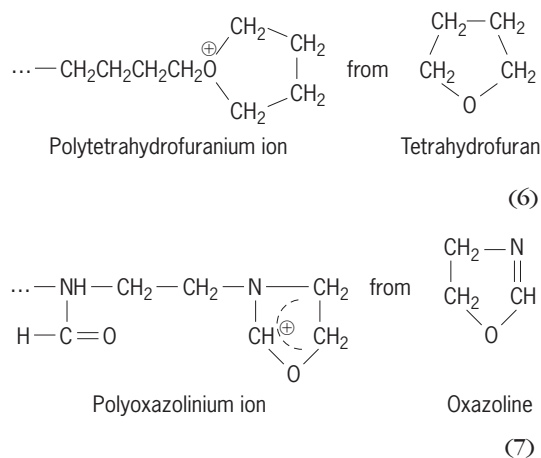
See POLYACETAL.

Cationic ring-opening polymerization is used in industry mostly for polymerization of 1,3,5-trioxane and its copolymerization with ethylene oxide or 1,3-dioxolane. This is the largest single polymeric material of technical importance made by cationic ring-opening polymerization. It has a number of applications as a high-melting polymer with excellent mechanical properties and high dimensional stability, such as electronic components and automotive parts. *See* COPOLYMER.

The poly(1,3,5-trioxane) [that is, polyformaldehyde] hemiacetal end groups resulting directly from polymerization are unstable, and on heating produce formaldehyde by an unzipping reaction (8).

$$\ldots-CH_2OCH_2OCH_2OCH_2OH \longrightarrow$$
$$\ldots-CH_2OCH_2OCH_2OH + CH_2O \quad (8)$$
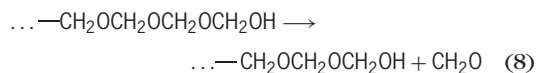
The homopolymer is stabilized by blocking the end groups, thus increasing the energy of activation for the first step of depolymerization. Another way to increase the polyacetal stability is to copolymerize 1,3,5-trioxane with ethylene oxide (or 1,3-dioxolane). At the end groups or at their immediate vicinity are units, $\ldots CH_2CH_2O-$, that are stable toward depropagation, since removal of the monomer molecule from the chain's end would require much more energy—in this instance a highly strained three-membered ring would have to be formed.

Another industrially important product is polytetrahydrofuran (polyTHF), mostly used as an oligomer with a molecular weight in the range of a few thousand. The α,ω-dihydroxypolyTHF are used as soft, elastic blocks in elastoplastic multiblock copolymers, namely with aromatic polyesters or polyamide rigid blocks, as well as in elastic fibers. Elastic fibers (such as spandex), which are additives to many modern textiles, are polyurethanes which contain soft and elastic poly(THF) blocks.

Activated monomer mechanism. Cationic ring-opening polymerization of heterocyclic compounds is plagued by inevitable side reactions. Recently a

novel mechanism of cationic ring-opening polymerization was elaborated. It was based on the observation that formation of cyclic side products by backbiting is suppressed in the presence of alcohols or water. Backbiting is the only way (except end-to-end biting) of macrocyclics formation when onium ions' active centers are located at the chain ends [reaction (9)]. In the traditional process, a macromolecule

Growing chain with
an active center*
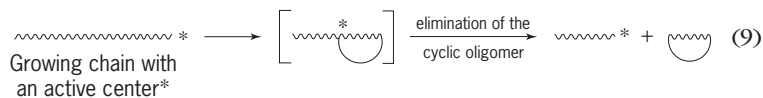
having active cationic species at its end ($\sim*$) attacks the nucleophilic unit in the chain (for example, an ether or an ester bond) and expels an oligocyclic compound in the next step.

The observed absence of macrocyclics when polymerization is conducted in the presence of alcohols led to the formulation of a mechanism in which cationic species are no longer at the end of the macromolecules, but are located on the monomer molecules. This is called the activated monomer mechanism (AMM). The simplest system, namely AMM in the cationic ring-opening polymerization of ethylene oxide, is as follows.

1. Protonation of the monomer [reaction (10)].

(Ethylene oxide)
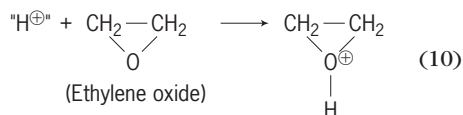
("$H^{\oplus}$" is a proton, linked ionically to a counterion; counterions are omitted, however.)

2. Initiation: addition (S_N2 substitution) of the initiator (for example, an alcohol molecule ROH) to the activated (that is, protonated) monomer molecule [reaction (11)].
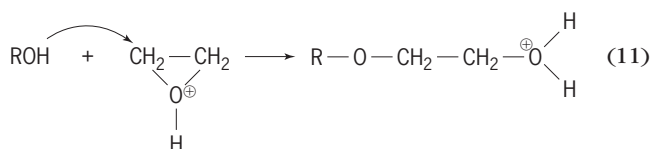
3. Transfer of a proton from the protonated end group to the next monomer molecule, followed by addition of thus activated monomer molecule to the . . .—OH end group in the growing macromolecule.

Propagation proceeds by repeating this sequence. The tertiary oxonium ions at the end of macromolecules are not formed, eliminating formation of macrocyclics.   S. Penczek

Bibliography. G. Allen and J. C. Bevington (eds.), *Comprehensive Polymer Science*, vol. III, Chap. 31–37 (Anionic ROP) and 46–52 (Cationic ROP), Pergamon Press, Oxford, 1984; A. Duda and S. Penczek, Mechanisms of aliphatic polyester formation, Chap. 12 in Biopolymers, vol. 3b: *Polyesters II—Properties and Chemical Synthesis*, edited by A. Steinbüchel and Y. Doi, Wiley-VCH, Weinheim, 2002; K. J. Ivin and T. Saegusa (eds.), *Ring-Opening Polymerization*, vols. I–III, Elsevier Applied Science, London,

1984; S. Penczek, P. Kubisa, and K. Matyjaszewski, *Cationic Ring-Opening Polymerization*, Springer, Berlin, vol. I, 1980, vol. II, 1985; M. Szwarc and M. van Beylen, *Ionic Polymerization and Living Polymers*, Chapman & Hall, New York, 1993.

Ring theory

The mathematical term ring is used to designate a type of algebraic system with two compositions satisfying most but not all the properties of addition and multiplication in the system of integers, 0, ± 1, ± 2, ... In precise terms a ring is a set R with two binary compositions called addition and multiplication whose results on an ordered pair (a,b), a,b in R, are denoted by $a + b$ and ab, respectively. These compositions must satisfy the following conditions:

C. $a + b$ and ab belong to R (closure).

A1. $a + b = b + a$ (commutative law).

A2. $(a + b) + c = a + (b + c)$ [associative law].

A3. There exist an element 0 (called zero) in R satisfying $a + 0 = a$ for every a in R.

A4. For each a in R there exist an element $-a$ (called the negative of a) in R such that $a + (-a) = 0$.

M1. $(ab)c = a(bc)$.

D. $a(b + c) = ab + ac$; $(b + c)a = ab + ca$ (distributive laws).

In the ring I of integers (addition and multiplication as usual) there are further conditions, for example, the commutative law of multiplication ($ab = ba$) and the cancellation law that if $a \neq 0$ and $ab = ac$, then $b = c$. *See* SET THEORY.

Types of rings. The importance of the concept of a ring stems from the fact that it embraces many special cases which are fundamental in all branches of mathematics. Thus it includes the ring I of integers, the ring R_0 of rational numbers, the ring $R^{\#}$ of real numbers, the ring C of complex numbers, various rings of functions, rings of matrices, and so on. An example of matrix rings is the following: Let $R_n^{\#}$ denote the collection of all the n by n arrays or matrices

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

whose entries a_{ij} are taken in the ring $R^{\#}$ of real numbers. Two such matrices $A = (a_{ij})$ and $B = (b_{ij})$ are regarded as equal if and only if $a_{ij} = b_{ij}$ for every $i,j = 1, 2, \ldots, n$. If $A = (a_{ij})$ and $B = (b_{ij})$, then $A + B$ is defined to be $S = (s_{ij})$ where $s_{ij} = a_{ij} + b_{ij}$, and AB is defined to be $P = (p_{ij})$ where $p_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$. The conditions C, A1–A4, M, and D are fulfilled, so that $R_n^{\#}$ is a ring. If $n = 1$, this is essentially the same as the ring $R_n^{\#}$, but if $n > 1$, then $R_n^{\#}$ differs from $R^{\#}$ because the commutative law of multiplication does not hold; that is, there are instances in which $AB \neq BA$. The example $R_n^{\#}$ has a geometric counterpart, namely, the ring of linear transformations of an n-dimensional vector

space. In fact these rings are isomorphic in the sense defined below. The first example of a ring in which the cummutative law of multiplication does not hold is the system Q of quaternions introduced by W. R. Hamilton as an appropriate tool for studying rotations in three-dimensional space. This system is the set of all the expressions of the form $a1 + bi + cj + dk$ where a, b, c, d are in $R^\#$. Addition is defined by adding corresponding coefficients as in ordinary algebra, and multiplication is defined to satisfy the associative and distributive laws and the following rules for the quaternion units 1, i, j, k: $1u = u = u1$ for $u = 1, i, j, k$; $i^2 = k^2 = -1$: $ij = -ji = k$, $jk = -kj = i, ki = -ik = j$. *See* QUATERNIONS.

The conditions A1–A4 on the addition composition are exactly equivalent to the statement that any ring is a commutative group relative to its addition composition. This group is called the additive group of the ring. The algebraic system consisting of the set of elements of a ring together with its multiplication composition is called the multiplicative semigroup of the ring. *See* GRAPH THEORY.

Various classes of rings are singled out by imposing conditions on the multiplicative semigroup. Thus integral domains are rings in which the product of nonzero elements is nonzero. Division rings are groups relative to the multiplication composition, and fields are division rings satisfying the commutative law of multiplication. The system of quaternions Q is a noncommutative division ring, $R^\#$ and C are fields. Important instances of integral domains are the rings $F[x_1, x_2, \ldots x_r]$ of all formal polynomials in the letters x_i with coefficients taken out of some field F.

If R is a ring, then a subring of R is a subset S of R, which is a ring with respect to the addition and multiplication defined in R. The conditions for this are that if s_1 and s_2 belong to S, then so do $s_1 - s_2 [= s_1 + (-s_2)]$ and $s_1 s_2$. In a similar manner the concept of a subfield of a field is defined.

Ideals, difference rings, homomorphism. In elementary number theory it is often important to separate the ring I of integers into subsets defined by a divisibility condition. For example, there are the sets of even and odd integers. More generally, if m is a positive integer, then the set of integers decomposes into m subsets $I_0, I_1, \ldots, I_{m-1}$ where I_j is the set of integers which gives the remainder j on division by m. Two integers a and b are in the same I_j if and only if $a - b$ is divisible by m. If this condition holds, one writes, following Gauss, $a \equiv b \pmod{m}$ which is read, "a is congruent to b modulo m." Congruences can be added and multiplied, leading to a new ring whose elements are the m sets $I_0, I_1, \ldots, I_{m-1}$. The study of this ring gives a natural setting for important results of number theory.

The construction just indicated can be carried over to any ring R. One begins with an ideal M in R which is defined to be a subset of R having the following closure properties:

1. If m_1, m_2 are in M, then $m_1 - m_2$ is in M.
2. If m is in M and a is any element of R, then am and ma are in M. If $R = I$ the ring of integers, then

the set M of multiples of a fixed positive integer m is an ideal.

As in this special case, two elements a,b of any ring R are said to be congruent modulo an ideal M if $a - b$ is in M. The ring M can be decomposed into nonoverlapping congruence classes where such a class $[a]$ is the complete set of elements x of R which are congruent to a fixed a. The congruence classes can be added and multiplied and are the elements of a new ring R/M called the difference (or factor or quotient) ring. This is the ring analog of the quotient group defined in the theory of groups. Also, as in group theory, the mapping which associates with every a of R the congruence class $[a]$ of R/M is the prime instance of a homomorphism. If R and R' are any two rings, a mapping $a \rightarrow f(a)$ of R into R' is a homomorphism if $f(a + b) = f(a) + f(b)$ and $f(ab) = f(a)f(b)$. If distinct elements have distinct images, then f is called an isomorphism, and R and its image are said to be isomorphic. The image under any homomorphism is isomorphic to the difference ring R/M, where M is the ideal of elements mapped into 0. This basic result is the fundamental theorem of homomorphism for rings. *See* GROUP THEORY.

Advanced aspects. Ideals play an important role in higher arithmetic, which deals with the arithmetic aspects of number fields. Ideals are also basic in algebraic geometry. In fact, the theory of algebraic curves makes considerable use of a type of ideal theory, called Dedekind ideal theory, which includes the ideal theory of number fields. On the other hand, higher-dimensional algebraic geometry can be founded on the ideal theory of so-called noetherian rings.

The structure theory of rings is essentially an analysis of the anatomy of rings. The general idea is that of reducing the study of classes of rings to that of simpler classes. Intertwined with this is the theory of representations of rings. A representation of a ring is defined to be a homomorphism of a ring into a ring of endomorphisms of a commutative group. *See* BOOLEAN ALGEBRA; LATTICE (MATHEMATICS); NUMBER THEORY.   Nathan Jacobson

Bibliography. N. Jacobson, *Lectures in Abstract Algebra*, vols. 1 and 2, 1951 and 1953, reprint 1981 and 1984; N. Jacobson, *Structure of Rings*, 2d ed., 1964, reprint 1990; I. Kaplansky, *Commutative Rings*, 1970, reprint 1994; S. MacLane and G. Birkhoff, *Algebra*, 3d ed., 1987; D. Passman, *A Course in Ring Theory*, 1991; L. H. Rowen, *Ring Theory*, student edition, 1991; O. Zariski and P. Samuel, *Commutative Algebra*, 2 vols., 1958 and 1960, reprint 1986 and 1991.

Ripple voltage

The time-varying part of a voltage that is ideally time-invariant. Most electronic systems require a direct-current voltage for at least part of their operation. An ideal direct-current voltage is available from a battery, but batteries are impractical for many applications. To obtain a direct-current voltage from
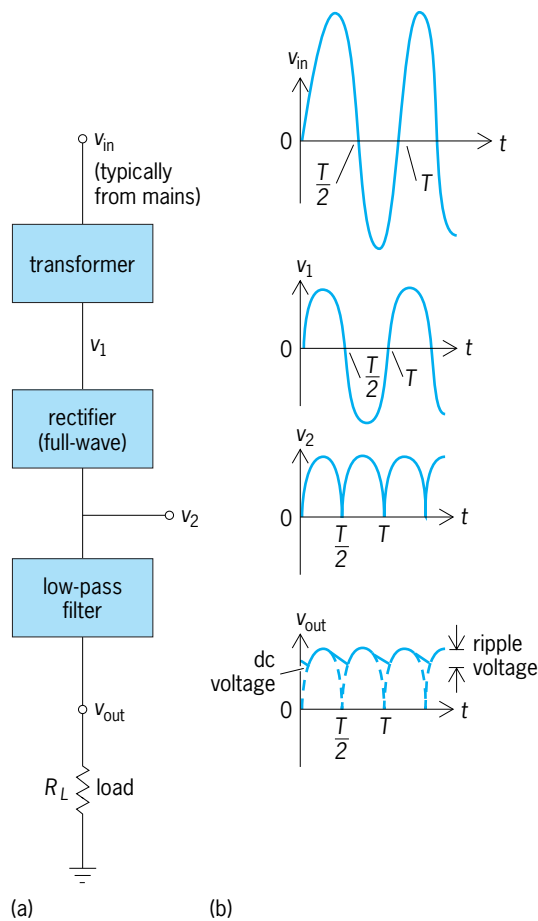
Power supply system with ripple voltage defined.
(*a*) **System configuration.** (*b*) **Voltage waveforms. *T* = period
of input voltage.**

the alternating-current power mains requires using some type of power supply.

A typical linear power supply system configuration (see **illus.**) consists of a transformer to change the voltage at the mains to the desired level, a rectifier to convert the alternating-current input voltage v_1 to a pulsating direct-current voltage v_2, followed by a low-pass filter. The output voltage v_{out} of the filter consists of a large direct-current voltage with a superimposed alternating-current voltage. This remaining superimposed alternating-current voltage is called the ripple voltage. The percent ripple (in contrast to the ripple voltage) is defined by the equation below.

$$\% \text{ ripple} = \frac{\text{root-mean-square (rms) value of the ac component}}{\text{dc or average value of } v_{out}} \times 100\%$$

The root-mean-square value of the alternating-current component (that is, the ripple voltage) is typically what is measured by an alternating-current voltmeter, while the direct-current value of v_{out} is measured by a direct-current voltmeter. From these two measurements, this formula can be used to obtain a quantitative value for the percent ripple. The ripple voltage has a fundamental frequency of twice the input voltage frequency for a single-phase, full-wave rectifier, and a fundamental frequency equal to that of the input voltage for a single-phase, half-wave rectifier. The ideal ripple would be 0%, which would mean that v_{out} consists of only a direct-current component. Ripple is often less than 1% in a well-designed system. The ripple voltage can be reduced at the expense of increasing the size of some of the components of the low-pass filter or by reducing the load-current drain on the circuit by increasing the value of R_L. The low-pass filter is more effective in reducing ripple voltage if the frequency of the input voltage is higher. A higher input frequency is not practical for operation from 60-Hz alternating-current mains but is used in switched-mode power supplies.

Practical linear power supplies often include a voltage regulator between the low-pass filter and the load. The voltage regulator is usually an electronic circuit that is specifically designed to provide a very stable dc output voltage even if large variations occur in the input. Nonlinear power supplies, which are often termed switching power supplies or switched-mode power supplies, are a practical alternative for producing a low-ripple dc output. *See* ELECTRIC FILTER; ELECTRONIC POWER SUPPLY; RECTIFIER; VOLTAGE REGULATOR.   Stanley G. Burns

Risk assessment and management

The scientific study of risk, the potential realization of undesirable consequences from hazards arising from a possible event, the assessment of the acceptability of the risks, and the management of unacceptable risks. For example, the probability of contracting lung cancer (unwanted consequence) is a risk caused by carcinogens (hazards) contained in second-hand tobacco smoke (event). The risk is estimated using scientific methods and then the acceptability of that risk is assessed by public health officials. Risk management is the term for the systematic analysis and control of risk, such as prohibiting smoking in public places. Risks are caused by exposure to hazards. Sudden hazards are referred to as acute (for example, a flash flood caused by heavy rains); prolonged hazards are referred to as chronic (for example, carcinogens in second-hand tobacco smoke and polluted air).

The definition of risk contains two components: the probability of an undesirable consequence of an event and the seriousness of that consequence. In the example of a flash flood, risk can be defined as the probability of having a flood of a given magnitude. Sometimes the probability is expressed as a return period, which means, for instance, that a flood of a specified magnitude is expected to occur once every 100 years. The scope of a flood can be expressed as the level or stage of a river, or the dollar amount of property damage.

Most human activities involve risk. The risk of driving, for example, can be subdivided according to property damage, human injuries, fatalities, and harm to the environment. Even the stress and lack of exercise due to driving create health risks. Although

risk pervades modern society and is widely acknowledged, it continues to cause unending controversy and debate. In the example of driving a car, the issues include which risks should be studied and how they should be measured and managed; which safety features should be incorporated into cars in order to reduce certain risks; whether governments should require stronger automobile frames to reduce the severity of accidents; and whether the use of safety belts should be required in the front and rear seats. People perceive risks in different ways; decision makers can exercise more effective risk management by improved public risk communication.

Risk analysis and assessment are applied to new technologies, chemicals, air-quality standards, the adequacy of emergency response forces, and other risk problems. However, risk estimates are seldom accurate to even two orders of magnitude, and widely varying perceptions of risk by different interest groups can add confusion and conflict to the risk management process. As a result, risk analysis and assessment are often carried out conservatively with a safety factor as high as 1000. This, in turn, leads to nonoptimal decision making, because the risks are exaggerated.

Environmental risk assessment is laden with uncertainty, particularly with respect to the quantification of chemical emissions; the nature of contaminant transport (such as the region over which a chemical may spread and the velocity of movement) in the water, air, and soil; the type of exposure pathway (such as inhalation, ingestion, and dermal contact); the effects on people based on dose-response studies (which are extrapolated from animals); ecological impacts; and so forth. Moreover, the public perception of risk is shaped by powerful psychological and social processes such as the dread of cancer, the involuntary nature of the exposure, and the risks to future generations. This outlook may lead to biases when judging risks (such as overestimating the frequency of very uncommon events), public demands for zero-risk criteria (such as the no-threshold criterion), and improved compliance with environmental regulations; and also may raise a myriad of socioeconomic issues, from social equity and competiveness to who bears the costs for improved environmental stewardship. To this end, the use of effective risk communication and risk principles such as ALARA (as low as reasonably achievable) and de minimus (the risk is so small that most people are uninterested in giving up the risk-producing activity, or the risk does not warrant regulatory enforcement) may help to promote more effective environmental risk management, particularly for the most noxious environmental risks.

Risk management. Thousands of natural and other hazards are subjected to the statistical analysis of mortality and morbidity data. Society selects a small number of risks to manage, but often some high risks (such as radon in houses) may not be managed, while some low risks (such as movement of dangerous goods) may be selected for management. Management alternatives include banning of the haz-
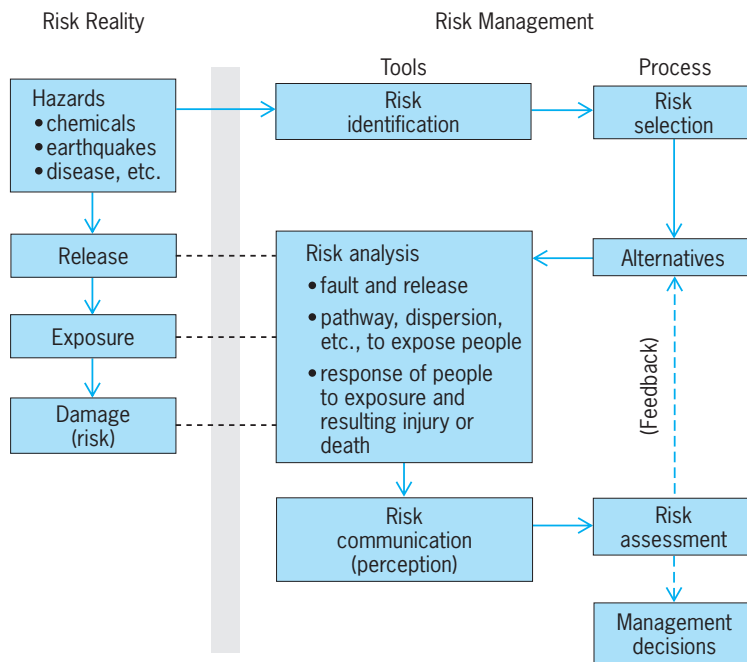

Fig. 1. Risk analysis, assessment, and management.

ard (drugs), regulating the hazard (drivers' tests and licensing), controlling the release and exposure of hazardous materials, treatment after exposure, and penalties for damages. Each management alternative may be analyzed to estimate the impact on risk, which may be known only within a few orders of magnitude (**Fig. 1**).

Risk estimates are uncertain, are described in technical language, and are outside the general understanding or experience of most people. Perception plays a crucial role, tending to exaggerate the significance, for example, of risks that are involuntary, catastrophic, or newsworthy. Effective risk management therefore requires effective risk communication.

Risk assessment is the evaluation of the relative importance of an estimated risk with respect to other risks faced by the population, the benefits of the activity source of the risk, and the costs of managing the risk. For risks due to long-term exposure to chemicals, the risk assessment activity generally incorporates the estimation of the response of people to the exposure (that is, risk analysis is a part of risk assessment). The methods used include studies on animals, exposure of tissues, and epidemiology.

Examples of applications. The general approach to risk management outlined above is needed for solving risk problems arising in areas ranging from the disposal of nuclear wastes to the design of safe toys. Below, examples of risk analysis and risk assessment are used to illustrate how risk management is carried out in practice.

Risk analysis. The nuclear and chemical industries have developed methods of probabilistic risk assessment, using techniques such as event trees and fault trees. The top event in a fault tree is the fault, and the tree indicates how it can occur and with what frequency. In a fault tree for an overflowing gasoline
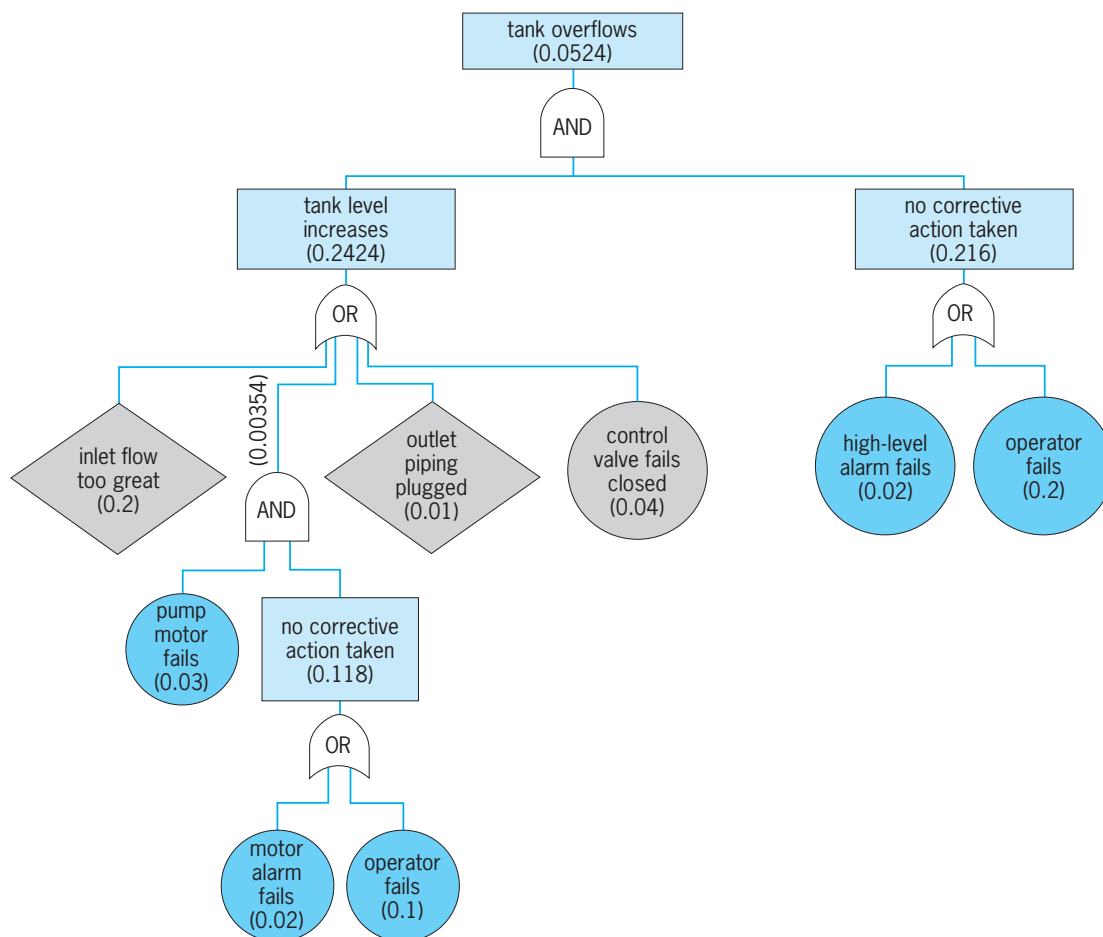
Fig. 2. Fault tree for an overflowing tank. The numbers in parentheses represent probabilities. (*After Chemical Manufacturers Association, Process Safety Management (Control of Acute Hazards), Washington, D.C., May 1985*)

tank, one of the most common causes of industrial accidents (**Fig. 2**), the tank overflows if it is being filled too fast AND the filling is not stopped. The probability associated with the AND logical operator can be calculated by the product of the relevant probabilities. An OR gate indicates that any of the input events is sufficient, and the probabilities are found by the complement of the probability that none of the input events would occur. For example, "no corrective action" can occur if the high-level alarm fails OR the operator fails to respond. Basic events for which data exist on failure rates are shown in circles. For example, data on thousands of motors can be used to estimate the mean failure rate per year and the variation in failure rates. Events that are specific to the situation, and whose probabilities must be estimated, are shown in diamond-shaped boxes. Events whose probabilities are calculated internal to the fault tree are shown in squares. Fault trees permit probability calculations of rare events through the application of logic and the knowledge of failure rates for relatively common events.

If a basic event such as "electrical power supply fails" is used as input to a number of roots of the failure tree, the probability of the fault or the top event is calculated by reducing the tree to a number of independent cut sets using boolean logic. Fault trees can also be used to evaluate alternative risk management strategies such as using an automatic shut-off in place of operator response to the high-level alarm.

An event tree generally starts from a fault such as the tank overflow problem. The event tree then continues as a typical decision tree to list the occurrence or nonoccurrence of subsequent events that can lead to the risk of loss of life. *See* DECISION THEORY.

Risk assessment. Health risks in the home or office are caused by many kinds of hazardous air pollutants, of which cigarette smoke is one of the best known. Another pollutant is radon, a radioactive gas occurring naturally by the nuclear decay of radium in the ground, which can enter houses through the basements. Less familiar indoor air pollutants include the great variety of complex organic chemicals that are given off by building materials, paints, furniture, cleaning fluids, and other manufactured products. Because buildings are often designed to be relatively airtight, air pollutants can build up to fairly high levels. Many indoor air pollutants not only are carcinogenic but can also lead to respiratory problems. Unfortunately, causal relationships are usually difficult to establish and consequently risk analyses are subject to great uncertainty.
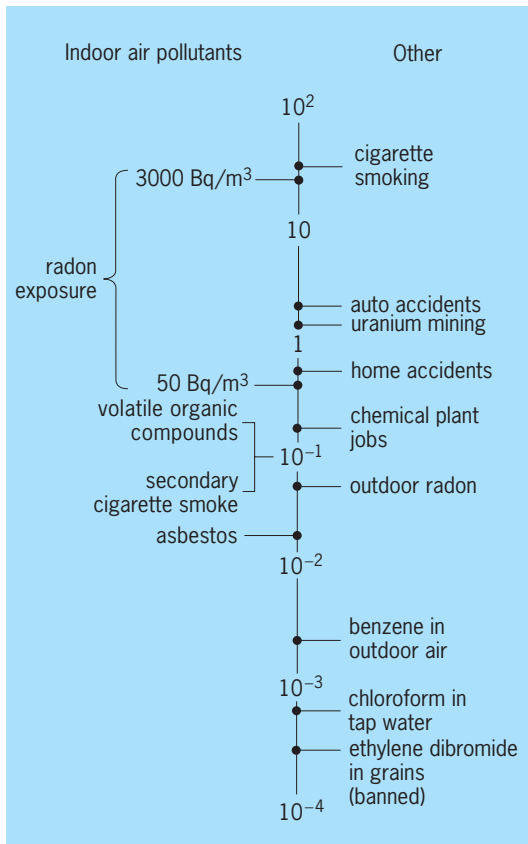
Fig. 3. Risk assessment scale, showing percentage lifetime risk of premature death from indoor radon exposure compared to other hazards; 1 becquerel per cubic meter = 0. 027 picocurie per liter. (*After A. V. Nero, Jr., Controlling indoor air pollution, Sci. Amer., 258(5):42–48, May 1988*)

A risk comparison (**Fig. 3**) shows that indoor radon exposure varies widely but is more dangerous than many familiar hazards. Such a comparison also shows that the estimated probability of suffering a fatal disease is significantly higher for exposure to indoor air pollutants than for exposure to pollutants in outdoor air, drinking water, and food; the average indoor concentration of radon in the United States translates to about 10,000 lung cancer deaths per year. The outdoor air pollution is kept low by regulation, and the resulting risk of cancer is actually two orders of magnitude below the risk due to chemicals in the indoor environment. *See* AIR POLLUTION; ENVIRONMENTAL ENGINEERING; ENVIRONMENTAL TOXICOLOGY; HAZARDOUS WASTE; NUCLEAR POWER; ONCOLOGY; RADIOACTIVE WASTE MANAGEMENT; RADON.

Keith W. Hipel; John Shortreed

Bibliography. P. L. Bernstein, *Against the Gods: The Remarkable Story of Risk*, 1996; Canadian Standards Association, *Risk Management: Guideline for Decision-Makers*, 1997; Y. Y. Haines, *Risk Modeling, Assessment and Management*, 1998; K. W. Hipel and L. Fang (eds.), *Effective Environmental Management for Sustainable Development*, 1994; G. Holmes, B. R. Singh, and L. Theodore, *Handbook of Environmental Management and Technology*, 1993; S. Kaplan, The words of risk analysis, *Risk Analysis*, 17(4): 407-417, 1997; M. G. Morgan, Risk analysis and management, *Sci. Amer.*, 269(1):32–41, July 1993; National Research Council, *Improving Risk Communication*, 1989; National Research Council, *Science and Judgment in Risk Assessment*, 1994; Royal Society of London, *Risk: Analysis, Perception and Management*, 1992.

Ritz's combination principle

The empirical rule, formulated by W. Ritz in 1905, that sums and differences of the frequencies of spectral lines often equal other observed frequencies. The rule is an immediate consequence of the quantum-mechanical formula $hf = E_i - E_f$ relating the energy hf of an emitted photon to the initial energy E_i and final energy E_f of the radiating system; h is Planck's constant and f is the frequency of the emitted light. For example, the **illustration** shows the photon
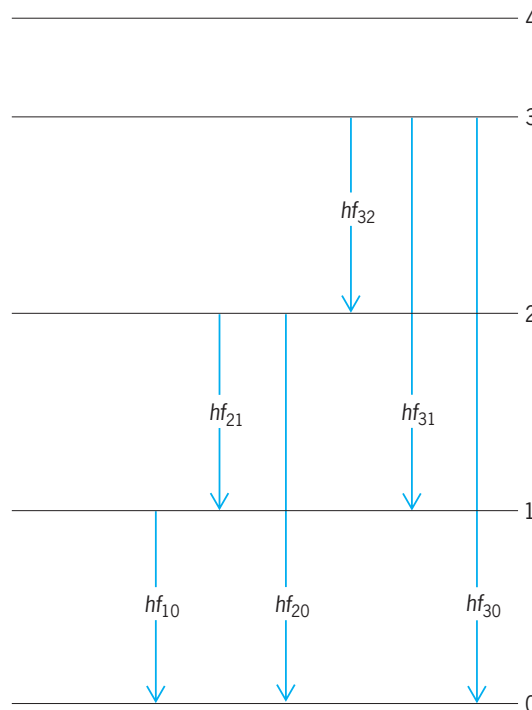
Energy levels and emitted frequencies.

energies hf_{32}, hf_{31}, hf_{30} associated with transitions from level 3 to lower-lying levels, and so on. Level 3 may radiate directly to the ground state 0, emitting f_{30}, or it may first make a transition to level 2, which subsequently radiates to the ground state, and so on. Since the total energy emitted in these two alternative means of making transitions from 3 to 0 is exactly the same, namely $E_3 - E_0$, it follows that $hf_{30} = hf_{32} + hf_{20}$. Similarly, $hf_{30} = hf_{32} + hf_{21} + hf_{10}$, and so forth. *See* ATOMIC STRUCTURE AND SPECTRA; ENERGY LEVEL (QUANTUM MECHANICS); QUANTUM MECHANICS.

Edward Gerjuoy

River

A natural, fresh-water surface stream that has considerable volume compared with its smaller tributaries. The tributaries are known as brooks, creeks, branches, or forks. Rivers are usually the main stems and larger tributaries of the drainage systems that convey surface runoff from the land. Rivers flow from headwater areas of small tributaries to their mouths, where they may discharge into the ocean, a major lake, or a desert basin.

Rivers flowing to the ocean drain about 68% of the Earth's land surface. The remainder of the land either is covered by ice or drains to closed basins (common in desert regions). Regions draining to the sea are termed exoreic, while those draining to interior closed basins are endoreic. Areic regions are those which lack surface streams because of low rainfall or lithoogic conditions.

Sixteen of the largest rivers (see **table**) account for nearly half of the total world river flow of water. The Amazon River alone carries nearly 20% of all the water annually discharged by the world's rivers. Rivers also carry large loads of sediment. The total sediment load for all the world's rivers averages about 22×10^9 tons (20×10^9 metric tons) brought to the sea each year. Sediment loads for individual rivers vary considerably. The Yellow River of northern China is the most prolific transporter of sediment. Draining an agricultural region of easily eroded loess, this river averages about 2×10^9 tons (1.8×10^9 metric tons) of sediment per year, one-tenth of the world average. *See* DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; LOESS.

River floods. River discharge varies over a broad range, depending on many climatic and geologic factors. The low flows of the river influence water supply and navigation. The high flows are a concern as threats to life and property. However, floods are also beneficial. Indeed, the ancient Egyptian civilization was dependent upon the Nile River floods to provide new soil and moisture for crops.

Floods are a natural consequence of the spectrum of discharges exhibited by a river. Rivers in humid temperate regions often exhibit less variable flood behavior than rivers in semiarid regions of rugged terrain. Floods in the humid regions tend to occur on average once each year. Rarer, larger floods are often no more than one or two times the magnitude of more common annual floods. The semiarid floods, however, are usually very small for common events, such as the average annual flood. However, rare, high-magnitude floods may be catastrophic.

The greatest floods in the geologic record occurred during the Pleistocene about 13,000 years ago in the northwestern United States. Lake Missoula, an ice-dammed lake in western Montana, released several catastrophic floods across the Channeled Scabland region of eastern Washington. The largest of these floods discharged as much as 20×10^6 ft^3 (570,000 m^3) of water per second. Flow velocities in the Scabland channelways ranged from 33 to 100 ft/s (10 to 30 m/s) for water 100–330 ft (30–100 m) deep. These phenomenal floods created a bizarre landscape of anastomosing channels, abandoned cataracts, streamlined hills, and immense gravel bars. So much water entered the preflood river valleys that they filled to overflowing, and floodwater scoured the divide areas between the valleys.

The Missoula floods were certainly among the most spectacular fluvial phenomena of all time. However, it should be remembered that most rivers do their work very slowly. Rivers are mainly transport agents, removing debris produced by the prolonged action of rainsplash, frost action, and mass movement. The many smaller floods probably accomplish much more of this work of transport than the rare large flood.

Floods are but one attribute of rivers that affect human society. Means of counteracting the vagaries of river flow have concerned engineers for centuries. In modern times many of the world's rivers are managed to conserve the natural flow for release at times required by human activity, to confine flood flows

---

Characteristics of some of the world's major rivers

| River	Average discharge, ft^3/s (m^3/s)	Drainage area, 10^3 mi^2 (10^3 km^2)	Average annual sediment load, 10^3 tons (10^3 metric tons)	Length, mi (km)
Amazon	6,390,000 (181,000)	2770 (7180)	990,000 (900,000)	3899 (6275)
Congo	1,400,000 (39,620)	1420 (3690)	71,300 (64,680)	2901 (4670)
Orinoco	800,000 (22,640)	571 (1480)	93,130 (86,490)	1600 (2570)
Yangtze	770,000 (21,790)	749 (1940)	610,000 (550,000)	3100 (4990)
Brahmaputra	706,000 (19,980)	361 (935)	880,000 (800,000)	1700 (2700)
Mississippi-Missouri	630,000 (17,830)	1240 (3220)	379,000 (344,000)	3890 (6260)
Yenisei	614,000 (17,380)	1000 (2590)	11,600 (10,520)	3550 (5710)
Lena	546,671 (15,480)	1170 (3030)	—	2900 (4600)
Mekong	530,000 (15,000)	350 (910)	206,850 (187,650)	2600 (4180)
Parana	526,000 (14,890)	1200 (3100)	90,000 (81,650)	2450 (3940)
St. Lawrence	500,000 (14,150)	564 (1460)	4,000 (3,630)	2150 (3460)
Ganges	497,600 (14,090)	451 (1170)	1,800,000 (1,600,000)	1640 (2640)
Irrawaddy	478,900 (13,560)	140 (370)	364,070 (330,280)	1400 (2300)
Ob	440,700 (12,480)	1000 (2590)	15,700 (14,240)	2800 (4500)
Volga	350,000 (9,900)	591 (1530)	20,780 (18,840)	2320 (3740)
Amur	338,000 (9,570)	788 (2040)	—	2900 (4670)
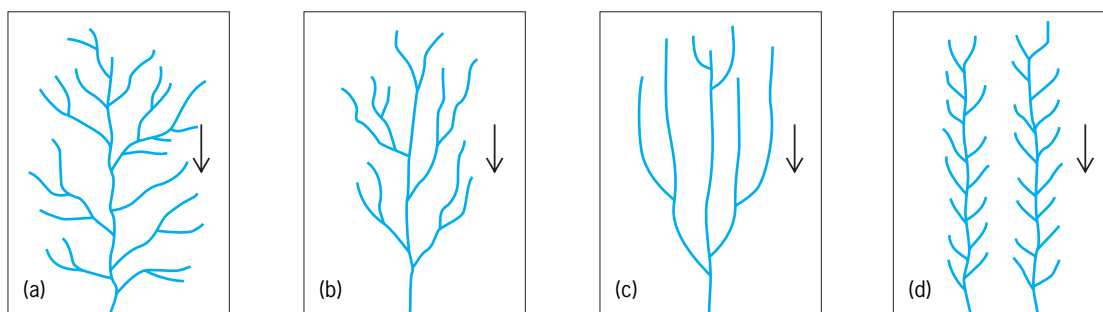
Fig. 1. Typical slope angles and associated drainage patterns forming on relatively homogenous planar beds: (*a*) surface gradient 1%, dendritic; (*b*) surface gradient 3%, subparallel; (*c*) surface gradient 5%, parallel; and (*d*) surface gradient 5%, pinnate. Arrows indicate direction of flow. (*After L. F. Phillips and S. A. Schumm, Effects of regional slope on drainage networks. Geology, 15:813–816, 1987*)

to the channel and to planned areas of floodwater storage, and to maintain water quality at optimum levels. *See* FLOODPLAIN; RIVER ENGINEERING.

Geologic history. Some rivers possess a long heritage related to their location along relatively stable continental cratons or their correspondence to structural lows, such as rift valleys. More commonly, rivers have been disrupted through geologic time by the Earth's active tectonic and erosional processes. The most recent disruptions were caused by the glaciations of the Pleistocene. Many rivers, like the Mississippi, became heavily loaded with coarse glacial debris delivered to their headwaters by glaciers. Since the last glacial maximum about 18,000 years ago, most of the world's rivers have adjusted their channel sizes, patterns, and gradients to the new environmental conditions of postglacial time.

In tropical regions the effects of glaciation were indirect. During full-glacial episodes of the Pleistocene, tropical areas of the Amazon and Congo river basins experienced relative aridity. Forests were replaced by savanna and grassland. The greater erosion rates on the land contributed large amounts of coarse sediment to the rivers. Today these regions have returned to their high-rainfall condition. The rivers receive relatively little sediment from interfluves that are stabilized by a dense forest cover. However, human exploitation of the tropical forests seems to be effectively returning the landscape to its glacial condition. This will undoubtedly induce profound changes in many tropical rivers.   Victor R. Baker

Stream pattern analysis. The patterns of intersecting stream channels provide valuable information about the geology, topography, climate, and hydrology of the region, and about some of the ways in which humans have altered the land surface. The alignment of the streams (for example, branched, rectangular, or parallel) is controlled by bedrock type, topography, and the locations of faults, folds, and joint patterns. The density of the network (that is, how closely spaced the channels are) is controlled by the climate, vegetation, age, permeability, and slope of the surface. Stream patterns are thus a product of the physical geography of a region and provide valuable clues to that geography.

Geologists and geographers have long used stream patterns for landscape analysis. An interesting application of pattern analysis is in extraterrestrial studies; scientists have used stream patterns to identify surficial processes on planets where on-site collection of data is impossible.

Form, topography, and geology. The general form of stream patterns reflects the underlying topography and geology of a region. Dendritic, or branchlike, drainage (**Fig. 1***a*) is the most commonly occurring pattern; it indicates a lack of strong structural controls (that is, faults, joints, or folds) or complex topography. Dendritic patterns thus commonly occur on relatively horizontal sediments or beveled plains of any rock type. Large drainage networks such as the Mississippi River almost always display dendritic patterns, with other patterns occurring at smaller area scales.

If there are no structural controls or significant variations in rock type and the surface gradient is increased to 3% (that is, a rise of 3 m in 100 m of horizontal distance), the dendritic pattern typically assumes a subparallel pattern (Fig. 1*b*). At slopes of 5% and higher on planar surfaces, the subparallel pattern usually goes through a transition to a parallel pattern (Fig. 1*c*) on most rock surfaces, or a pinnate pattern on silts and clays (Fig. 1*d*).

The shape of the topography as well as the gradient also plays a role in controlling the form of drainage patterns. Volcanoes create radial drainage patterns (**Fig. 2***a*). Topographic lows, such as occur in the closed basins of Nevada, create the centripetal pattern shown in Fig. 2*b*. The hummocky terrain of permafrost areas, regions of limestone solution, and recently deposited materials from glaciers, volcanic explosions, and slides creates a deranged pattern, which consists of many ponds and unconnected tributaries (Fig. 2*c*). Deranged patterns may evolve to become dendritic as streams erode the divides between ponds and eventually form an interconnected network of tributaries (Fig. 2*d*).

Structural controls can also override the natural tendency for drainages to assume a dendritic form, or a parallel form on slopes with gradients greater than 5%. For example, parallel drainages can occur on slopes of less than 5% in areas such as the south shore of Lake Superior, where ancient glacial grooves restrict flow to parallel channels. Trellis patterns (Fig. 2*e*) generally indicate parallel fractures or
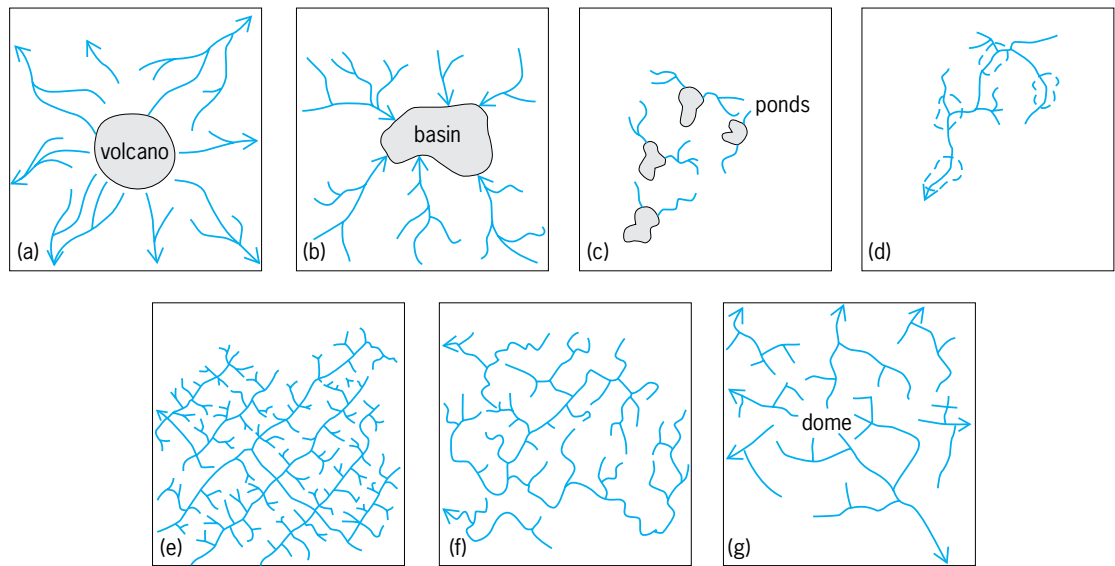
Fig. 2. Diagrams of broad categories of structurally and topographically controlled drainage pattern forms. (*a*) Radial. (*b*) Centripetal. (*c*) Deranged. (*d*) Dendritic. (*e*) Trellis. (*f*) Rectangular. (*g*) Annular. The dendritic drainage (*d*) shows the pattern that evolves at the deranged pattern (*c*) location after sufficient time has passed for erosion of the divides between ponds and formation of a connected channel network. Broken outlines in *d* show old pond locations in the deranged pattern (*c*). (*After A. D. Howard, Drainage analysis in geologic interpretation: A summation, Amer. Ass. Petrol. Geol. Bull., 51:2246–2259, 1967*)

parallel dipping rock beds (for example, old beaches that form resistant sandstone), which create long reaches of parallel streams with smaller tributaries to either side. The rectangular pattern shown in Fig. 2*f* reflects an underlying structure or joints or faults at right angles. Annular patterns occur in areas where the rocks have been bent to form domes and basins (Fig. 2*g*).

Pattern density. The various stream patterns can occur at different densities; in other words, channels may range from being closely to widely spaced. Density can be controlled by climate, vegetation, bedrock and soil permeability and strength, slope and shape of surface, and the age of a surface and stream. In turn, human disruption of the landscape can affect all these factors at a variety of scales, from a garbage dump to a river basin.

The evaluation of controls on channel density is important because channel density provides key clues to the physical geography of drainage basins. Furthermore, erosion and flash floods are controlled by drainage density, because more channels mean more runoff and excavation of sediment. Drainage densities therefore have been used by hydrologists to estimate mean annual flows, annual flood sizes, low flows due to ground-water discharge into streams, and sediment yields. Finally, by understanding the controls on drainage density, engineers can design mine tailing piles, landfills, and other artificial landscapes that help maintain drainage densities that reduce erosion and flash floods. *See* GROUND-WATER HYDROLOGY; HYDROLOGY.

Climate and vegetation. Drainage density is most strongly controlled by the intensity of precipitation (the amount of precipitation falling per unit time) at regional scales; other factors generally play a role at local scales. The widest variation in drainage densities occurs in semiarid climates, where local factors often override climatic controls.

Despite the impact of local factors, climate-driven factors generally cause densities to be greatest in semiarid areas such as the western Great Plains the United States, where occasional, intense thunderstorm activity coupled with sparse vegetation cover
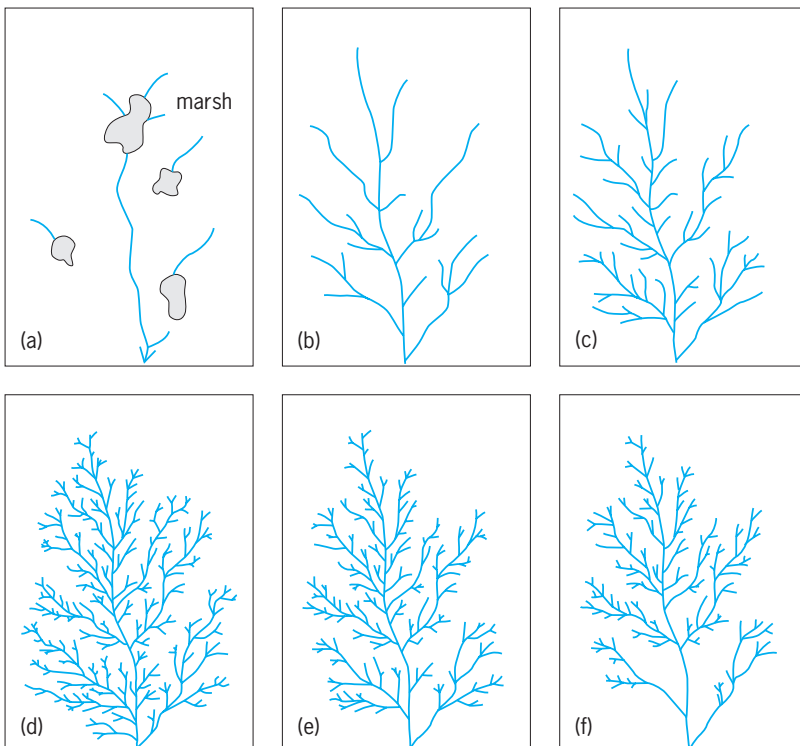

Fig. 3. Drainage pattern evolution from (*a*) Initiation as a deranged pattern through (*b*)–(*d*) maximum channel development in a dendritic pattern to (*e, f*) erosional lowering of the basin and loss of channels from the dendritic pattern. (*After W. S. Glock, The development of drainage systems: A synoptic view, Geog. Rev., 21:475–482, 1931*)

promotes erosion and the formation of many channels. Densities typically decrease in more arid climates because of the absence of rainfall and runoff. Densities also typically decrease in more humid climates, probably because of (1) increased vegetation, which holds soil in place and reduces erosion, and (2) increased soil cover, which acts in a spongelike manner to absorb and store water, thus reducing the surface runoff that erodes channels.

Rock type and soil permeability. Surfaces possessing high permeability, such as sands or loose volcanic debris, generally produce low stream densities, because precipitation can soak into the surface rather than generating runoff and erosion. Similarly, surfaces such as clays or silts that have low permeability produce high drainage densities, because water cannot sink into the surface, and so runs off and erodes channels. Badlands, which are barren, heavily eroded and dissected landscapes, almost invariably consist of shales or silts that cannot absorb precipitation. On occasion, however, this relationship can be confusing. Granites, for example, are impermeable to water, but are also sufficiently strong to resist extensive erosion and channel formation. In addition, the soils that form from granites are very porous and can absorb significant quantities of rainfall.

Age of the surface. The older the surface, the more time has been available for erosion and network formation. Thus, if there are identical surfaces that have experienced identical climates, the surface will progress through stages of little erosion, tributary development and high drainage density, and finally reduction in density as erosion reduces lower portions of the basin to a relatively flat plane (**Fig. 3**). This process may be repeated if the basin is uplifted or the base level (for example, sea level or the level of a reservoir) drops, both events starting a new cycle of erosion.

At the scale of small areas (for example, one hillside) in soft bedrock or soils, the drainage density may also change on a seasonal basis. During the cold or dry season, channels may be destroyed as frost action loosens dry sediment that fills in the channel, while during the wet season, channels are reestablished by runoff.

Slope and shape of the surface. If all other factors are equal, drainage densities increase with gradient up to a material-dependent threshold value. Densities increase because steeper slopes have less soil cover to absorb rain, and runoff flows more quickly and has more erosive power on steeper gradients. At lower gradients, upwardly convex slopes tend to produce slightly lower drainage densities than planar or concave slopes, but this difference in densities disappears as gradients increase to the threshold value.   W. Andrew Marcus

Bibliography. E. Derbyshire (ed.), *Geomorphology and Climate*, 1976; W. L. Graf, *Fluvial Processes in Dryland Rivers*, 1988; L. Leopold, M. G. Wolman, and J. P. Miller, *Fluvial Processes in Geomorphology*, 1995; L. F. Phillips and S. A. Schumm, Effects of regional slope on drainage networks, *Geology*, 15:813–816, 1987; S. A. Schumm, *The Fluvial System*, 1977; S. A. Schumm, *River Morphology*, 1982; S. A. Schumm, M. P. Mosley, and W. E. Weaver, *Experimental Fluvial Geomorphology*, 1987; M. M. Smart et al. (eds.), *Ecological Perspectives of the Upper Mississippi River*, 1986; B. A. Whitton (ed.), *Ecology of European Rivers*, 1989.

River engineering

River engineering involves the control and utilization of rivers for the benefit of humankind. Its scope, in the broad sense, may include river training, channel design, flood control, water supply, navigation improvement, hydraulic structure design, hazard mitigation, and environmental enhancement. River engineering is also necessary to provide protection against floods and other river disasters.

Throughout recorded history, rivers have been a subject of study by engineers and scientists who have been fascinated by the geometric shapes that rivers form and their responses to natural and human interference. River engineering is somewhat different from other aspects of civil engineering because the emphasis is often on river responses, long-term and short-term, to changes in nature, and stabilization and utilization, such as damming, channelization, diversion, bridge construction, and sand or gravel mining. Evaluation of river responses is essential at the conceptual, planning, and design phases of a project. The use of fundamental principles of river and sedimentation engineering is required in the analysis of special conditions for each river project. *See* CANAL; DAM; RIVER.

River training. This, in a broad sense, covers all the engineering works to regulate river flow and sediment transport for the sake of flood control, navigation, irrigation, and channel stabilization. River training by embankments can be traced to early human history, exemplified by the works on the Nile, Yellow, Euphrates, Tigris, and Ganges rivers. Such practice continues in modern times. Rivers are often trained because of urbanization; land reclamation from flood plains has resulted in the channelization of many rivers in different parts of the world. *See* FLOODPLAIN; STREAM TRANSPORT AND DEPOSITION.

The principle of river training is twofold: (1) The training works must be designed to be strong enough to withstand the design flow, and (2) the impacts on the river should be understood and evaluated whenever feasible. The commonly used types of training works can be broadly classified as bank protection, dikes, and grade-control structures.

Bank protection. Different types of bank protection are in use, including rock riprap, mattress, gabion, soil cement, concrete blocks, cribs, Armorflex (**Fig. 1**), and so on. Because of environmental concerns, materials that also allow vegetation growth are preferred (**Fig. 2**). Soil cement blocks are used as bank protection as a suitable and inexpensive alternative to rock riprap. They are used mainly in places where riprap is not readily available at a reasonable cost. The adequate cement content was
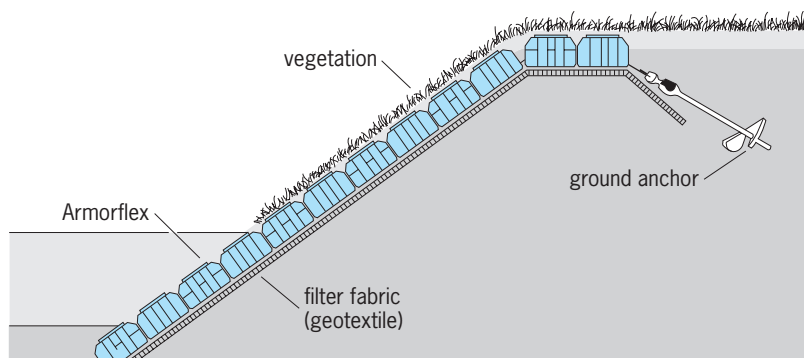
Fig. 1. Armorflex mats made of concrete blocks connected by nylon cables are used for bank protection. (*Armortec Concrete Erosion Control Systems*)

determined to range 9–10% by weight of the soil. The soil-cement bank protection is popular in the semiarid southwestern United States, where the construction is facilitated by the usually dry river bed. Such bank protection is also desirable from the environmental viewpoint because it preserves the soil appearance. *See* GEOSYNTHETIC.

Dikes. These are training structures that extend from the bank to the river at an angle, or perpendicular, to the flow. Most dikes are aligned in a slightly downstream direction. They are also known as groins, spurs, spur dikes, or transverse dikes and represent a very widely used type of training works. Dikes are often used to form a system covering a certain river reach. Dikes serve one or more of the following functions: (1) training a river along a desired course, (2) creating a region of low velocity to induce siltation, (3) protecting the bank by keeping the flow away, and (4) contracting a wide river channel usually for the improvement of depth for navigation. There are two principal types of dikes: permeable and impermeable. Permeable dikes, such as timber pile dikes and jetty fields, permit flow through the dikes at reduced velocities, thereby preventing bank erosion and causing deposition of suspended sediment from the flow. Experience has shown that per-

meable dikes are more effective than solid ones as a bank protection, especially in silt and sand rivers. Jetties, or steel jacks, consist of angles fastened together and strung with wire. The jetty field is typically used in shallow streams to train the stream into a single narrower channel; it reduces the velocity near the bank and hence protects the bank from erosion. The jetty field is more effective in streams that have a considerable amount of debris and a high concentration of suspended load.

Grade-control structures. These are also called drop structures, stabilizers, weirs, barrages, or check dams, and are generally constructed normal to the channel flow and traverse the channel bed (Fig. 2). Grade-control structures are used in river channels to maintain a slope flatter than the slope of the terrain. Stabilizers refer to sediment control structures that are used primarily to stabilize the upstream channel bed where scour may endanger certain structures such as bridge foundations. The crest of a structure usually extends across the channel, and the side walls should extend into the bank and have adequate bank protection to prevent flanking at high flows. Each structure should also have adequate upstream and downstream protection. Dumped stones should be placed on the downstream side to the anticipated scour depth. While a grade-control structure stabilizes the upstream channel bed, it usually induces downstream changes. When major gradation change is expected, a series of grade-control structures may be used to limit the extent of change at each structure.

River responses. Alluvial rivers are self-regulatory in that they adjust their characteristics in response to any change in the environment. These environmental changes may occur naturally, as in the case of climatic variation or changes in vegetative cover, or may be a result of such human activities as river training, damming, diversion, sand and gravel mining, channelization, bank protection, and bridge and highway construction. Such responses usually occur at places where the bed is mobile (**Fig. 3**) or where the bank is

Fig. 2. Channelization of the Santa Gertrudis Creek in California showing bank protection and grade control structure traversing the channel bed.

unprotected. If a river reach is straightened and channelized, its freedom for meandering is constrained. In response to this change in slope due to straightening, changes in the adjacent natural reaches may occur in order to restore the equilibrium of the river system. In that process, the river adjusts to the new conditions by changing its slope, roughness, bed-material size, cross-sectional shape, or meandering pattern. Within the existing constraints, any one or a combination of these characteristics may adjust as the river seeks to maintain the balance between its ability to transport sediment and the load provided.

Any plan for river training needs to be evaluated with regard to channel responses. In fact, channel responses should be the criteria for determining the adequacy of a design scheme. Modifications in design are often necessary in arriving at the final plan. In modern river engineering, a very challenging task is the prediction of river channel responses. Studies of river hydraulics, sediment transport, and river channel changes may be conducted through physical modeling or mathematical modeling, or both. Traditionally, physical modeling has been relied upon to obtain the essential design information. It nevertheless often involves large expenditures and is time-consuming in model construction and experimentation. What limits the accuracy of physical modeling is the scale distortion, which is almost unavoidable whenever it involves sedimentation. Mathematical modeling of erodible channels has been advanced with the progress in the physics of fluvial processes and computer techniques. Since the actual size of a river is employed in mathematical modeling, there is no scale distortion. The applicability and accuracy of a model depends on the physical foundation and numerical techniques employed. There exist several mathematical models that simulate river flow, sediment transport, and river channel changes, such as the HEC-6 model by the U.S. Army Corps of Engineers and the FLUVIAL-12 model by Chang. The computer-aided approach is illustrated by case studies for two types of problems: general scour at bridge crossings, and fluvial design of river bank protection. *See* COMPUTER-AIDED ENGINEERING; MATHEMATICAL GEOGRAPHY.

Bridge scour. Since a bridge may be damaged by river bed scour (Fig. 3), the scour potential should be determined in bridge design or restoration. River bed scour at bridge crossings consists of local and general scour. Local scour that occurs around bridge piers and abutments is caused by the local obstruction to flow. General scour develops when more sediment is removed from an area than is supplied from upstream. For example, the proposed Rancho Santa Fe Road Bridge on San Marcos Creek in California is located a few miles below a dam. The deficit in sediment supply caused by the dam will induce general scour at the bridge crossing. General scour at bridge crossings is related to the flow and sediment-transport processes of the adjacent river as a system; therefore, evaluation of such scour requires modeling of a river reach.

Fig. 3. River bed scour at a bridge crossing induced by sand mining in the river channel.

Channelization. The major trend at present is to stay away from totally lined channels for the sake of preserving ground-water recharge, wildlife habitat, and so on. Channelization is therefore accomplished by constructing bank protection while leaving an alluvial bed (Fig. 2). The bank protection must be designed to withstand the design flood, and it must be entrenched beyond the potential channel-bed scour. Therefore, an important task in channel design is to arrive at a hydraulic geometry which will minimize potential channel-bed changes. The water-surface profile and bed topography during the passage of the design flood must also be determined for the design of bank height and toe-down (entrenchment) depth. The design for channelization is exemplified by the Santa Cruz River. The river reach in Tucson, Arizona, underwent dramatic changes during the October 1983 flood, which reached the 100-year magnitude at many locations. River channel changes resulted in severe damages. For the purposes of flood control and hazard mitigation, Pima County took the responsibility of channelizing several reaches of the river. Channelization of the Santa Cruz River was designed on the basis of the pattern of river channel responses simulated using the FLUVIAL-12 model. Scour and fill of the channel bed related to the longitudinal imbalance in sediment load were simulated; they were also tied in with the effects of the spiral motion that induces an uneven bed topography in channel bends, characterized by greater scour near the concave bank. Variable top and toe elevations for hydraulic structures were determined and employed to make more efficient use of construction materials while providing effective protection against erosion.   Howard H. Chang

Bibliography. H. H. Chang, *Fluvial Processes in River Engineering*, Wiley, 1988; H. H. Chang and Z. Osmolski, Computer-aided design for channelization, *J. Hydraulic Eng.*, 114(11):1377–1389, 1988;

M. S. Petersen, *River Engineering*, Prentice Hall, Englewood Cliffs, NJ, 1986; C. R. Thorne, R. D. Hey, and M. D. Newson (eds.), *Applied Fluvial Geomorphology for River Engineering and Management*, 1998.

River tides

Tides that occur in rivers with an open connection to a tidal estuary or ocean. These tides are highly modified from coastal ocean tides as they propagate landward into shallow water, often through a brackish estuary and then into a fresh-water river. Ocean tides form in response to the gravitational forces exerted by the Sun and Moon acting on the water masses on the surface of the rotating Earth. Within estuaries or rivers, water masses are small enough that tidal motions are not generated directly from lunar and solar forces; rather, the motions occur because of tide-generating forces at the estuary entrance or river mouth. However, variations in the tide-generating forces in a coastal ocean produce corresponding variations in tides within the upland estuary/river system. In addition to the well-known diurnal or semidiurnal occurrences of high and low waters, numerous other important cycles occur that are necessary to describe tidal oscillations in estuaries and tidal rivers. Tides have their greatest amplitude when the Sun, Moon, and Earth are in alignment during full and new moon and the tidal forces of the Moon and Sun reinforce each other. These tides are referred to as spring tides. Alternatively, neap tides occur when the moon is in quadrature (the solar and lunar forces oppose rather than reinforce each other) and tidal amplitudes are much smaller; that is, higher low tides and lower high tides occur during neap periods. In addition to the spring/neap cycle that has an approximately fifteen-day period, many other astronomical cycles occur. Two of the more important cycles are the ones responsible for diurnal inequalities between the heights of the two high and two low waters experienced daily in some locations and the effect on tidal amplitude that results from the moon's varying distance from earth. *See* ESTUARINE OCEANOGRAPHY; RIVER; TIDE.

Tidal range. Tide-generated waves are known as "long" or "shallow water" waves because of their long wavelengths and typically small-wave amplitudes relative to water depths. Their wave speed, given approximately by the time of travel of a long progressive wave, is defined as $\sqrt{gD}$, where $g = 9.81$ m/s^2 (32.2 ft/s^2) is gravitational acceleration and D is water depth. Thus, wave speed is affected by local bottom topography, which can significantly modify incoming waves in shallow bays, estuaries, and rivers. Shoaling within estuaries and rivers slows progress of the tidal wave; the speed/depth relation creates a condition where the tide crest moves faster than the trough causing the wave front to steepen. This results in a situation where the duration of the rising tide is generally less than that of the falling tide and the slope of the tide curve becomes more asymmetrical as it propagates further inland. Under certain conditions usually associated with large tide range, shallow depth, and sharp increase of slope of riverbed, the distortion of the tidal curve between low water and high water becomes sufficiently large that a discontinuity of water level occurs. The condition is referred to as a tidal bore. Most tidal bores are small, possibly only a few centimeters (0.1 ft), but extreme bores may reach heights of 3 m (10 ft) or more and travel at speeds of 25 km/h (15.5 mi/h) during spring tides. *See* TIDAL BORE.

Whereas tidal bores are often spectacular, they are generally rare. Typically, tidal range increases as the width of the estuary or river narrows. Both high- and low-water levels rise with bed level, and tidal range begins to decrease. River tides propagate upstream until friction dissipates the tidal energy or some other physical impediment is encountered. Each estuary and river system has different characteristics because of its unique topography and geometry. As an example, 36 h of predicted tide heights starting January 1, 2005, at four locations within the San Francisco Estuary and Sacramento River system in California are shown (see **illustration**). Delay in arrival time of the tide and general damping of tide amplitude are observed as the tide propagates from the bay entrance at Golden Gate near San Francisco to an upstream location at Walnut Grove on the Sacramento River south of Sacramento. Tide tables indicate that by the time the tide reaches Rio Vista at 100 km (62 mi) up river, the time of high water occurs about 4 h later and the time of low water is delayed more than 5 h. In addition, the range of tide is reduced; high waters are reduced by about 20% and low waters are only about half as low as the corresponding values of high and low waters at Golden Gate.

While the effects of astronomical tides on estuary/river systems are periodic and predictable, other nonperiodic factors can greatly modify the astronomical tides in these upland systems. Meteorological influences such as variations in air pressure and wind and hydrologic influences such as river inflows, including the secondary influence that river flows have on salinity gradients within the estuary, can affect river tides. Air pressure changes, often referred to as the inverse barometer effect, create a rise in water level of about 1 cm (0.4 in.) for every 1 millibar (0.1 kPa) drop in pressure. Sustained winds may cause water to "pile up" along a coastline or at the downwind end of open water bodies, such as estuaries, creating changes in water level. Finally, runoff from rain or snowmelt may cause river flows to rise and lead to an increase in water level heights. River flows can retard the progress of rising tides, while enhancing falling tides.

Tidal currents. Tidal currents in rivers are the horizontal motions of water produced by the river tides. In the lower reaches of an estuary or river, tidal currents are usually bidirectional but become unidirectional (but still oscillatory) further upstream. In a river, tidal current reversals may occur far upstream under conditions of large tidal amplitudes and low
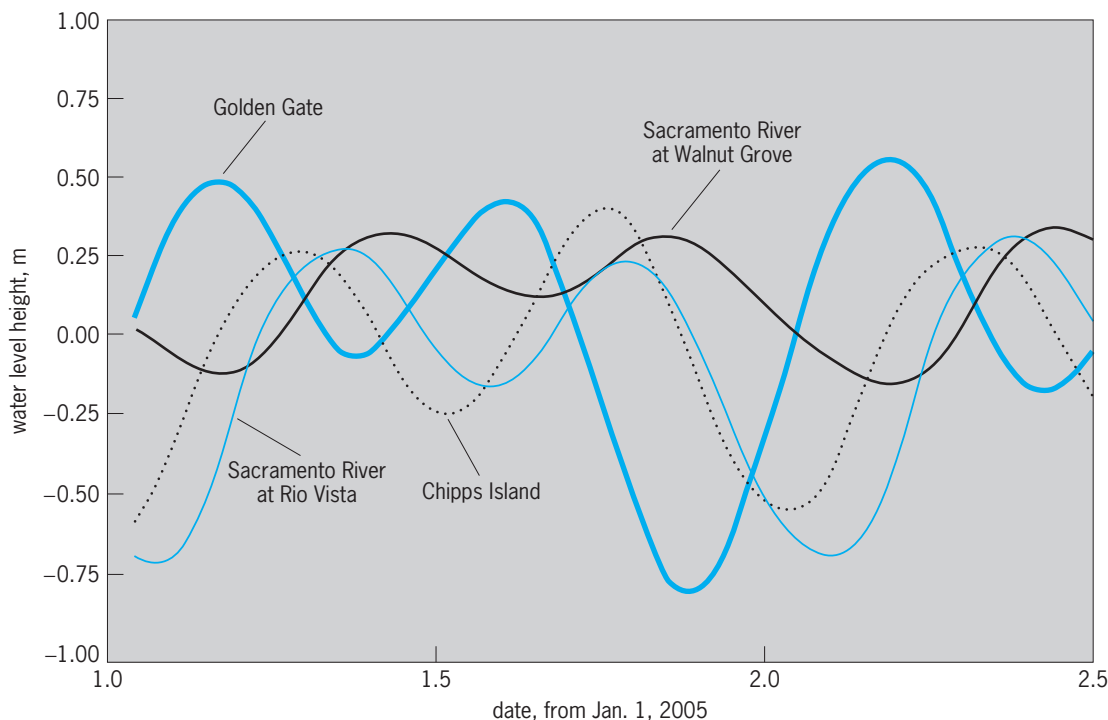
Time series of predicted water level heights referenced to mean water level at four locations in the San Francisco Bay and Sacramento River system. Golden Gate is at the entrance to bay (km = 0); Chipps Island is 72 km (45 mi) up-estuary; Sacramento River at Rio Vista is 100 km (62 mi) from Golden Gate; and Sacramento River at Walnut Grove is 130 km (81 mi) from Golden Gate.

river flows. Tidal currents are highly modified by geometry and bathymetry, with stronger currents occurring in deep, narrow channels and weaker currents occurring in wide, shallow areas. Upstream-directed currents that occur on rising tides are referred to as flood currents and downstream-directed currents that occur on falling tides are referred to as ebb currents. Slack water occurs at times when the flow changes direction and there is little or no water motion.

Times of maximum and minimum currents and their relation to high and low tides are related to how the tide wave progresses within the estuary/river system. The tide is a progressive wave that advances with respect to a fixed point of observation. Tides and tidal currents are in phase in a pure progressive wave, which means that maximum flood currents occur at high waters and maximum ebb currents occur at low waters. Slack waters occur at midtide levels. In contrast to a progressive wave, a standing wave appears stationary with respect to a fixed point of observation. Standing waves occur when two progressive waves of equal amplitude and period traveling in opposite directions are superimposed. Standing waves can occur in estuaries with closed ends or in tidal rivers that are blocked by a barrier such as a dam or falls from which the incident tide wave is reflected. A pure standing wave is characterized by tides and tidal currents that are 90° out of phase; maximum currents occur at midtide level and slack waters occur at times of high and low water. Reflections of the incident tide wave from channel sidewalls can result in a wave progressing down the estuary or

river. This ordinarily leads to tides that propagate in some combination of progressive and standing waves. Under this type of combined motion, the time delay as the tide progresses landward is generally less than for a pure progressive wave. Where rivers flow into estuaries, tides are generally more progressive than standing and are subject to decreases in amplitude from attenuation by bottom friction and channel constrictions, as well as attenuation from reflection from sides of channels.   Jeffrey W. Gartner

Bibliography. A. Defant, *Ebb and Flow: the Tides of Earth, Air, and Water*, 1958, University of Michigan Press, 1958; C. B. Officer, *Physical Oceanography of Estuaries (and Associated Coastal Waters)*, Wiley, 1976; D. T. Pugh, *Tides, Surges and Mean Sea-level, a Handbook for Engineers and Scientists*, Wiley, 1987.

Rivet

A short rod with a head formed on one end. A rivet is inserted through aligned holes in two or more parts to be joined; then by pressing the protruding end, a second head is formed to hold the parts together permanently. The first head is called the manufactured head and the second one the point. In forming the point, a hold-on or dolly bar is used to back up the manufactured head and the rivet is driven, preferably by a machine riveter. For high-grade work such as boiler-joint riveting, the rivet holes are drilled and reamed to size, and the rivet is driven to fill the hole completely. Structural riveting uses punched holes.
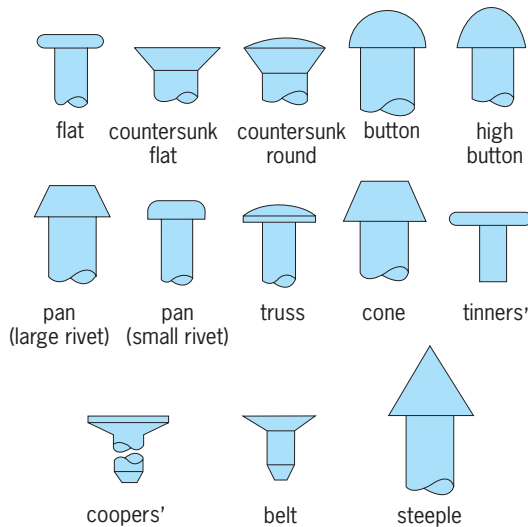
Fig. 1. Standard rivet heads.

Small rivets 7/16 in. or 11 mm and under) are used for general-purpose work with head forms as follows: flat, countersunk, button, pan, and truss (**Fig. 1**). These rivets are commonly made of rivet steel, although aluminum and copper are used for some applications. The fillet under the head may be up to 1/32 in. or 0.8 mm in radius.

Large rivets (1/2 in. or 13 mm and over) are used for structural work and in boiler and ship construction with heads as follows: roundtop countersunk, button (most common), high button or acorn, pan, cone (truncated), and flattop countersunk.

Boiler rivets have heads similar to large rivets with steeple (conical) added but have different proportions from large rivet heads in some cases.

Special-purpose rivets are tinners' rivets, which have flatheads for use in sheet-metal work; coopers' rivets, that are used for riveting hoops for barrels, casks, and kegs; and belt rivets, used for joining belt ends.

Blind rivets are special rivets that can be set without access to the point. They are available in many designs but are of three general types: screw, mandrel, and explosive (**Fig. 2**). In the mandrel type the rivet is set as the mandrel is pulled through. In the explosive type an explosive charge in the point is set off by a special hot iron; the explosion expands the point and sets the rivet.

Standard material for rivets is open-hearth steel (containing Mn, P, S) with tensile strength 45,000–55,000 lb/in.^{2} or 310–380 megapascals. Standards include acceptance tests for cold and hot ductility and hardness. Materials for some special-purpose rivets are aluminum and copper.   Paul H. Black

Bibliography. American National Standards Institute (ANSI), *Semitubular Rivets, Full Tubular Rivets, Split Rivets and Rivet Caps*, B18.7–1972; ANSI, *Small Solid Rivets 7/16 Inch Nominal Diameter and Smaller*, B18.1.1–1972; E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; Institute for Power System, *Handbook of Industrial Fasteners*, 3d ed., 1997.

Fig. 2. Two types of blind rivet.

Robotics

A field of engineering concerned with the development and application of robots, and computer systems for their control, sensory feedback, and information processing. There are many types of robotic systems, including robotic manipulators, robotic hands, mobile robots, walking robots, aids for disabled persons, telerobots, and microelectromechanical systems.

The term "robotics" has been broadly interpreted. It includes research and engineering activities involving the design and development of robotic systems. Planning for the use of industrial robots in manufacturing or evaluation of the economic impact of robotic automation can also be viewed as robotics. This breadth of usage arises from the interdisciplinary nature of robotics, a field involving mechanisms, computers, control systems, actuators, and software. *See* BIOMECHANICS; COMPUTER; CONTROL SYSTEMS; CYBERNETICS; ELECTRICAL ENGINEERING; INDUSTRIAL ENGINEERING; MECHANICAL ENGINEERING; SOFTWARE ENGINEERING.

Mechanisms. Robots produce mechanical motion that, in most cases, results in manipulation or locomotion. For example, industrial robots (**Fig. 1**) manipulate parts or tools to perform manufacturing tasks such as material handling, welding, spray painting, or assembly; automated guided vehicles are used for transporting materials in factories and warehouses. Telerobotic mechanisms provide astronauts with large manipulators for remotely performing spacecraft maintenance. Walking machines explore active volcanoes. Mechanical characteristics for robotic mechanisms include degrees of freedom of movement, size and shape of the operating space, stiffness and strength of the structure, lifting capacity, velocity, and acceleration under load. Performance measures include repeatability and accuracy of positioning, speed, and freedom from vibration.

Manipulator geometries. An important design for robotic manipulators is the articulated arm, a configuration with rotary joints that resembles a human arm (Fig. 1*a*). An extension of this idea is the selective compliance assembly robot arm (SCARA) configuration (Fig. 1*b*), which has been extensively applied to assembly in manufacturing. SCARA manipulators have high stiffness during vertical motions, but they

Fig. 1. Industrial robots. (*a*) Articulated, with rotary joints (*GM Fanuc Robotics Corp.*). (*b*) Selective compliance assembly robotn arm (SCARA) configuration (*Adept Technology, Inc.*) (*c*) Overhead gantry.

may be fitted with mechanical and sensory control devices that allow freedom in lateral motions during assembly operations.

Cylindrical configurations having a rotary joint at the base and a prismatic joint at the shoulder are useful for simple material transfer and assembly. One of the earliest manipulators, marketed in the late 1950s, was based on a configuration that had two rotary axes at the shoulder; this configuration has disappeared from commercial application. Overhead gantry (cartesian geometry) robots are used for high-payload applications and those requiring linear positioning over large work spaces (Fig. 1*c*).

Degrees of freedom. Some manipulators have simple mechanical configurations involving only two or three degrees of freedom of movement. Most robotic manipulators have six degrees of freedom so that they can position a part or approach a part with any desired position and orientation. The wrist is positioned at a desired x,y,z position in the work space. Then, the end effector is rotated to a desired orientation (roll, pitch, yaw). In effect, the wrist represents the origin of a three-axis coordinate system fixed to the gripper. Moving the first three joints of the arm translates this origin to any point in a three-axis coordinate system fixed to the work space; motion of the final three joints (in the wrist) orients the gripper coordinate system in rotation about an origin at the wrist point. While not all robots are constructed this way, the mathematics of mechanical motion (kinematics) is considerably simplified when these axes of the wrist do not intersect. *See* DEGREE OF FREEDOM (MECHANICS); KINEMATICS.

Types of joints. Robotic mechanisms can have joints that are prismatic (linear motion), articulated (rotary motion), or a combination of both types. While many robots use only articulated joints to imitate the human arm, limited actions can be produced by using prismatic joints alone. Robot locomotion can also be obtained by the use of wheels and treads.

Walking robots have been developed that use articulated legs. *See* MECHANISM.

Joint actuators. Actuators for moving joints of a robotic mechanism are usually electric motors. In the past, larger robots were built with hydraulic actuation due to their high payload capacity and ability to work in explosion-prone environments. In the 1990s there was a trend by the major manufacturers to use electric actuators in most of their industrial robots.

Actuators can be placed directly at the joints; however, the weight and bulk of these motors and associated gear transmissions limit the performance of the robot, particularly at the wrist of industrial robots. Another design involves placement of the actuators in the base of the robot and transmission of motion to the joints through mechanical linkages such as shafts, belts, cables, or gears. This approach overcomes many of the problems associated with locating actuators at the joints, but requires the design of backlash-free mechanical linkages that can transmit power effectively through the arm in all of its positions and orientations. An innovation in the design of assembly robots is the use of direct-drive motors, which eliminate the need for gear transmissions. This type of robot has become popular for high-speed assembly tasks. *See* BELT DRIVE; CHAIN DRIVE; GEAR; LINKAGE (MECHANISM).

End effectors. A two-jaw parallel gripper actuated by a pneumatic cylinder is a common type of end effector for industrial robots. End effectors are specialized for particular applications, such as parts handling, welding, and drilling. Using a quick-disconnect mechanism, robots can change their own end effectors to suit the task from a selection of special-purpose attachments. Since position errors of robots, fixtures, and grippers may occur during high-precision assembly tasks, compliance is needed to compensate for misalignment between parts during operations such as inserting a piston into an
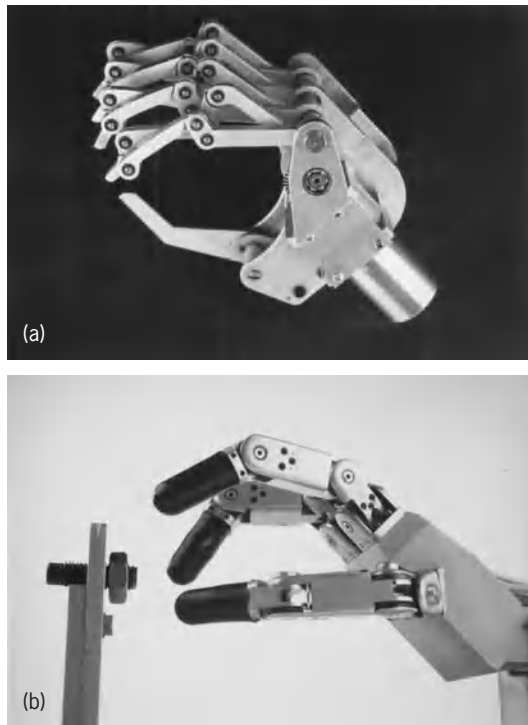
Fig. 2. Robotic hands. (*a*) Five-finger mechanical hand actuated by a single direct-current motor. (*b*) Three-finger, nine-degree-of-freedom dextrous mechanical hand. (*Intelligent Robotics Corp.*)

engine head or fitting a gland on a valve assembly. A spring-actuated mechanism, called a remote center compliance (RCC) wrist, may be attached between the robot arm and the end effector for this purpose.

Robotic hands. Dexterity is as basic to manufacturing assembly as it is to prosthetic and human hands. Multifingered robot hands can accomplish fine manipulation, achieve robotic hand cooperation for fixtureless assembly, operate in hazardous environments, explore remote locations, and perform other important functions. The usefulness of multifingered robot hands to perform fine manipulation and grasping tasks is widely acknowledged, but it is costly to implement. Mechanisms for robotic hands are commercially available with two to four fingers and a thumb, each with one to four degrees of freedom and actuated by servomotors (**Fig. 2**).

Mobile robots. Mobile robots can have wheels, tracks, legs, or a combination of these parts. They also can crawl, for example, for inspecting pipes. Tracked mobile robots developed for military and surveillance applications are among the most stable and rugged. Automated guided vehicles (AGV) are wheeled mobile robots that deliver parts and tools in factories by following a guide path embedded in the floor. The payload capacity of these vehicles varies from 220 lb (100 kg) to several tons. An AGV for light-duty work such as mail delivery in office buildings can follow a stripe attached to the floor. An AGV can be programmed like an industrial robot to follow a preprogrammed trajectory and to make decisions along the way depending on inputs from external signals. The path of an AGV can be guided by a wire underneath the floor or a stripe painted on the surface. Free-ranging AGVs are also available for industrial and commercial use to avoid the need for installing a fixed path.

Walking machines. Walking robots have been built with two, four, and six legs. One company marketed a four-legged walking machine that could walk over obstacles, lift up to five times its weight while stationary, and lift two times its weight while moving (**Fig. 3**). The control of gait stability for legged locomotion systems is a difficult implementation issue, particularly for two-legged (biped) machines. Research in locomotion has provided results on kinematic and dynamic modeling, control algorithms, and other forms of stability. Solutions to these research issues have resulted in a better understanding of high-level planning, control requirements, and their implementation.

Robot control and planning. A robot control system directs the motion and sensory processing of a robot or system of cooperating robots. The controller may consist of only a sequencing device for simple robots, although most multiaxis industrial robots today employ servo-controlled positioning of their joints by a microprocessor-based system.

Servo systems. A measure of the actual joint position is obtained from a transducer, such as an optical shaft encoder, and this signal is compared with the position specified for the joint. If the desired position (called the set point) and the measured position differ, the circuitry applies a correcting drive signal through the digital-to-analog converter (DAC) to the power amplifier of the joint motor. Such servo systems employ analog and digital circuitry; in addition
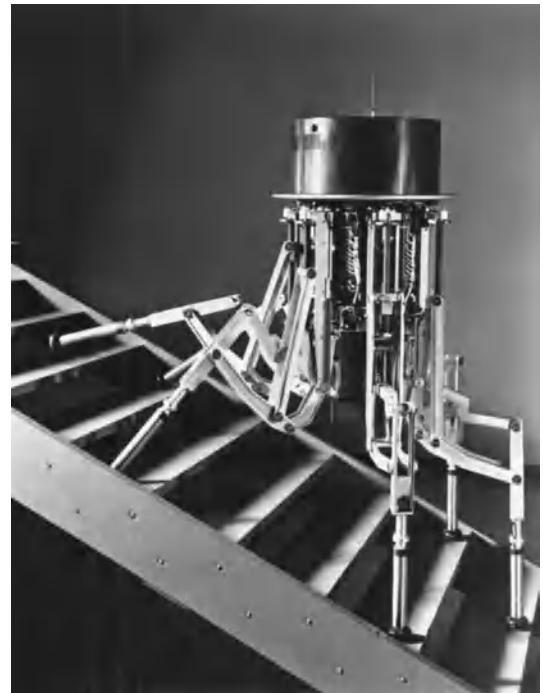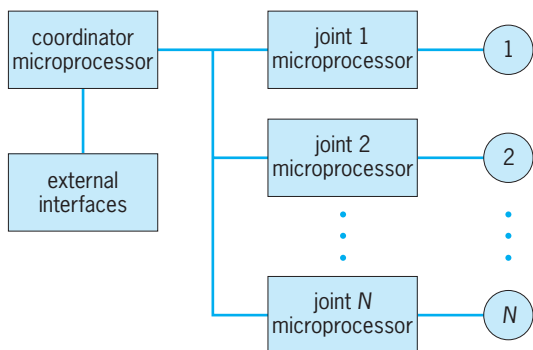

Fig. 3. Six-legged walking robot. (*Odetics*)

Fig. 4. **Joint-parallel controller.**

to position, they usually control joint velocity. *See* DIGITAL-TO-ANALOG CONVERTER.

Servo systems allow the robot to be moved through any selected sequence of positions without the need to preset mechanical stops. The sequence of positions defining the robot's trajectory may be preprogrammed by numerically specifying the positions (and orientations) in the robot's coordinate frame, or they may be taught to the system by moving the robot to the desired point and recording the outputs of the joint position transducers. Such teaching methods allow the user to direct a complex action once, and to have the robot repeat it under preprogrammed computer control as long as the motion is required. *See* SERVOMECHANISM; TRANSDUCER.

Robot controllers. Because of the computing power and low cost of microprocessors, most robot controllers are computer-based systems, often with individual microprocessors for each joint (**Fig. 4**). The controller generates the robot's trajectory based on mathematical representation of the work space, work-space objects, and tasks stored in the computer's database. The computer may also generate trajectories for the robot that are not fixed but vary with the state of the external world as monitored by the robot's sensors. This type of sensory-interactive control permits the robot to act appropriately in relation to conditions, rather than relying on assumptions about the world. For example, without custom fixtures and timing of the workflow, the actual location of a part may vary from one instance to the next; without sensory-interactive control in such cases, the robot could proceed blindly through a set of actions at preprogrammed but incorrect positions. *See* COMPUTER PROGRAMMING; DIGITAL CONTROL; MULTIPROCESSING.

When separate microprocessors are used for control of each joint of the robot, coordination of these axes can be obtained by an upper-level processor that also generates the path for each joint and converts a description of the end-effector position to joint angles (**Fig. 5**). The robot controller receives inputs from a teach pendent that records joint positions of the desired path. An input/output (I/O) module monitors on-off signals from sensors, such as proximity switches, indicating part presence prior to gripping. Programs are stored on hard or floppy disk drives. Another function of the upper-level con-

trol is to interpolate the path so that the robot moves smoothly. This requirement, called continuous-path control, is needed for applications such as arc welding and spray painting. *See* CONTROL SYSTEMS; MICROPROCESSOR.

Mobile robot navigation. Navigation of mobile robots requires constructing a map of where a robot has been, a plan to tell it where to go, and a controller to drive it there. As in the case of manipulator-motion planning, there is a need for the automatic creation of plans or programs for guiding a mobile robot to its desired state. Path planners for mobile robots find a path through a mapped environment so that the robot can travel along it without colliding with another object. Methods designed for expert systems have been used to combine the vast amount of poorly structured knowledge obtained for control. *See* EXPERT SYSTEMS.

Robot path planning. Automatic path generation algorithms have been developed to enable robots to avoid obstacles, such as those encountered in a cluttered environment, or to coordinate the actions of multiple robots to lift a heavy object. One category of path-planning algorithms is based on an explicit representation of free space, subsets of robot configurations that are free of collisions. Such an approach will find a path if it exists, within the known subset of free space. Unfortunately, it is difficult to find a representation for a multidimensional path-planning problem that can be solved. In cartesian space, a manipulator is easy to visualize, but its description is very complex. Regions of free space are easy to find and model, but planning through them can be very difficult. A configuration of a moving object is a set of parameters that completely specifies every point on the object. Configuration space (C-space) is a graphical representation that helps the systems designer to visualize the motion of robots. The C-space for a robot manipulator is a rectangular coordinate space in which each axis represents one joint angle. In C-space, path planning is reduced to planning the
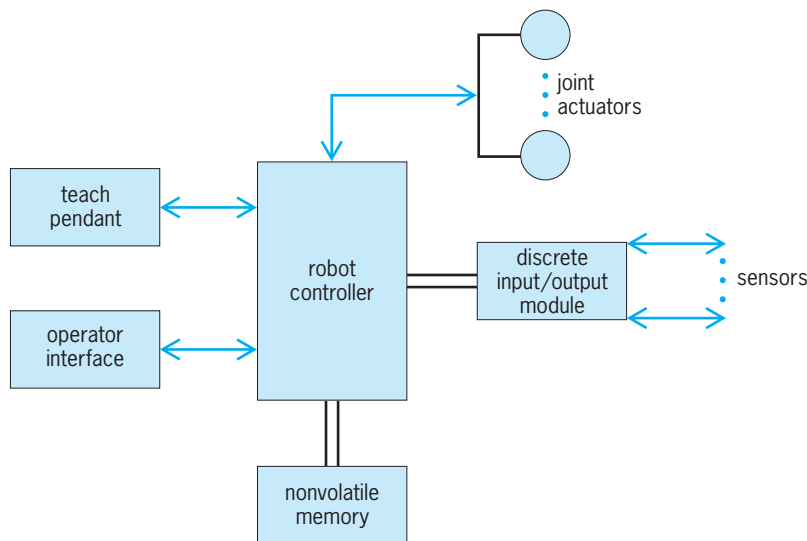

Fig. 5. **Robot controller architecture.**

path of a point through intersecting configuration space volumes. Computational algorithms for solving C-space plans have been based on "A* -Search," a graph search algorithm for selecting the best path from many possible paths. Other techniques based on nonlinear programming and heuristics have also been developed. *See* ALGORITHM; NONLINEAR PROGRAMMING.

Sensory systems. The robot sensory system gathers specific information needed by the control system and, in more advanced systems, maintains an internal model of the environment to enable prediction and decision making. The joint position transducers on industrial robots provide a minimal sensory system for many industrial applications, but other sensors are needed to gather data about the external environment. Sensors may detect position, velocity, acceleration, visual, proximity, acoustic, force-torque, tactile, thermal, and radiation data.

Tactile sensors. Tactile sensors may be mounted on the robot gripper to detect contact with objects in the work space. These sensors may be simple mechanical switches or transducers indicating the degree and direction of force. Arrays of such transducers may be used to discriminate types and orientations of objects. Force-torque sensors, mounted in the robot's wrist or end effector, are often used to sense the degree and direction of resistance due to the object being handled. These resistive forces may be due to the weight of the object being manipulated, or contact with other objects or surfaces. Such sensors are used to adjust gripping pressure, avoid applying destructive forces, and guide proper mating of surfaces and parts. In combination, these senses allow a robot to feel the proper fit of parts as does a human worker. *See* PRESSURE TRANSDUCER.

Robot vision. An inexpensive means of noncontact sensing for robots is to use a photoelectric diode that provides on-off signals such as "gripper open" or "part missing." These devices can also determine proximity. In robot vision, solid-state video cameras are used to capture an image such as a part to be grasped or a distance to be measured. The image can be processed by computer analysis using pattern recognition techniques, and can be used to evaluate grasp positions and automatically determine flaws based on comparison with stored images in the computer. Computer vision systems may use ambient light or structured light sources. Systems with ambient light sources rely on normal sources of illumination, while structured light sources provide special patterns of illumination whose shape and orientation are known to the sensory system. The advantage of structured light is that special patterns of illumination may be chosen to simplify and speed up the processing needed to interpret the image. Procedures for determining depth in images rely fundamentally on geometric calculations associated with triangulation procedures, but in structured light systems the triangulation is between the camera and the light source. In ambient light systems, corresponding points in two images taken from different viewing positions must be triangulated, and this is much more difficult to achieve. Speed is important in robot vision, because visual information can be used in real time by the control system to correct the robot's movements, a process called visual servoing. Sensor fusion allows the combining of independent and qualitatively different sensor data (for example, from vision and force sensing) to reinforce consistent interpretations, and disregard inconsistent interpretations. Research in this area has led to methods for extracting three-dimensional shapes from two-dimensional images; such methods require dealing with uncertain and imprecise data. *See* COMPUTER VISION.

Multisensor systems. The robot's world model can be continually updated on the basis of new sensory inputs, and can also generate expectations about how the view will transform under movements in progress. The world model's hypotheses in turn assist the interpretation of incoming sensory data. The most advanced robot sensory systems process information from many different sensor modalities, and are also structured to generate successive levels of description, at increasing degrees of complexity and decreasing rates, suited to use by the successive levels of the control system (**Fig. 6**). At the lowest level are the motor controllers and processes that describe and use feedback from sensor signals. These processes occur with cycle times on the order of 10–100 milliseconds. At the next level, perceptual information is brought together and overall control is computed to direct the robot. These processes occur at 100 milliseconds to 1 second. The third level in the architecture consolidates perception and locomotion. Cycle times in this level operate at 100 milliseconds for road following to 10 seconds for local path planning. Approaches for planning these actions have been based on preprogrammed algorithmic methods, while other methods have employed artificial intelligence rule-based decision making. At the highest level, the supervisory computer operates at control cycles of 1–100 seconds. *See* ARTIFICIAL INTELLIGENCE.

Intelligent systems. As information moves up from the device, the amount of information increases and the speed of data acquisition decreases. These control architectures form the basis for computer integrated manufacturing (CIM), a hierarchical approach to organizing automated factories. Recently, a new paradigm has emerged, based on the interconnection of intelligent system elements that can learn, reason, and modify their configuration to satisfy overall system requirements. One of the most important of these approaches is based on holonic systems. *See* AUTOMATION; COMPUTER-INTEGRATED MANUFACTURING; INTELLIGENT MACHINE.

Holonic systems. A holon is an intelligent, autonomous, cooperative unit of a system for transforming, transporting, storing, or validating information and physical objects. Holons are capable of processing information and performing physical tasks. A holon is a building block for holonic systems (**Fig. 7**). It can be either a fully self-functioning entity or an integral part of another holon. From an
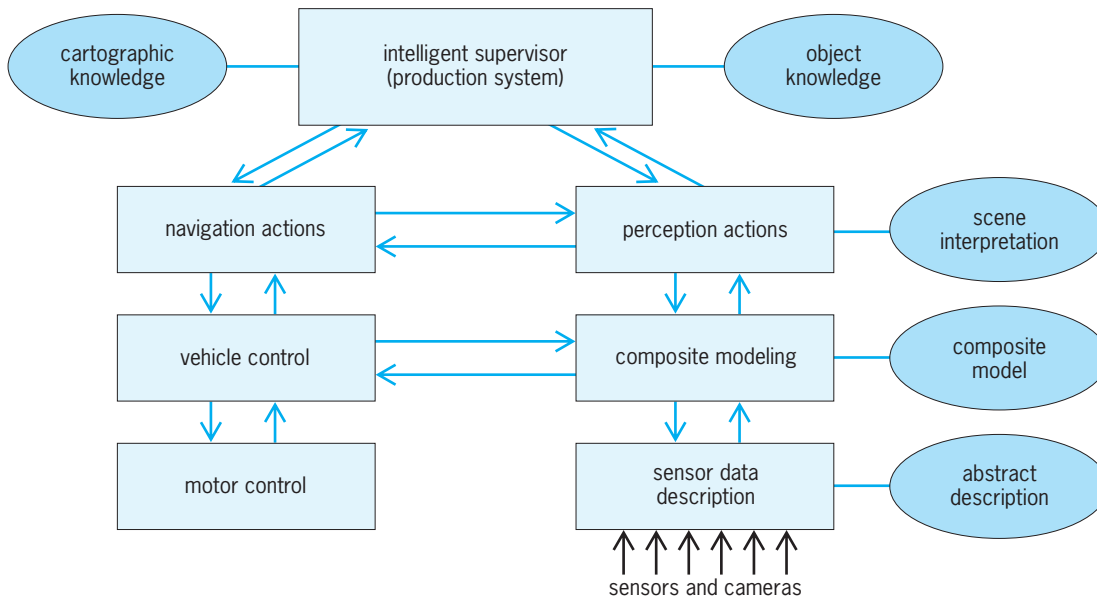
Fig. 6. Typical framework or architecture for mobile robot control. (*After R. C. Dorf, ed., International Encyclopedia of Robotics: Applications and Automation, vol. 3, Wiley, 1988*)

implementation viewpoint, holons are simply system hardware and software components that are modular, reusable, and reconfigurable. Holons may be robots, software packages, machine tools, part feeders, automated guided vehicles, or human operators. Object-oriented programming languages, such as C++, and computer-aided software engineering (CASE) tools provide a development environment to implement holonic systems. *See* OBJECT-ORIENTED PROGRAMMING.

A holarcy is a structure composed of holons to achieve a given objective or to execute a task optimally—for example, in terms of the amount of time taken or resources consumed. It can be distinguished from a hierarchical top-down structure and a group of loosely organized free agents in that actions take place in a cooperative and autonomous manner. Cooperation is obtained through the holons' interaction with the environment, the communication between holons themselves, and the communication from an upper-level controller. Autonomy is achieved through decentralized and distributed intelligence across the overall system, thereby enabling holons to interpret commands as well as sensory feedback to maintain a robust operation and handle unforeseeable disturbances. *See* DISTRIBUTED SYSTEMS (CONTROL SYSTEMS).

With traditional hierarchical architectures, high investments may be needed for system implementation to accommodate the required variety in production lot sizes and product mixes. Although related to multiagent systems, holonic systems exhibit not only autonomy but also a high degree of cooperation. Holonic systems provide a means for designing and integrating robotic and manufacturing systems. They offer the user quick installation and start-up, as well as easy development and operation with flexibility to meet changes in the environment. A holonic manufacturing system exhibits a high degree of strategic flexibility while maintaining a fixed or relatively stable set of rules for its internal organization and functioning. Human experiences and interaction may also be incorporated into holonic systems.

Neural systems. A neural network is an information-processing system that has performance characteristics in common with biological neural systems. This class of systems is useful for implementing difficult control and sensing tasks in robotics without requiring an accurate representation of the system. A neural net consists of a large number of simple processing elements called neurons. Each neuron is connected to other neurons by means of directed communication links, each with an associated weight. The weights represent information being used by the net to solve a problem. Neural nets can be used to classify patterns obtained from a digital camera image and perform general mappings from input patterns to output patterns, such as in fusing information from multiple sensors in map building. They
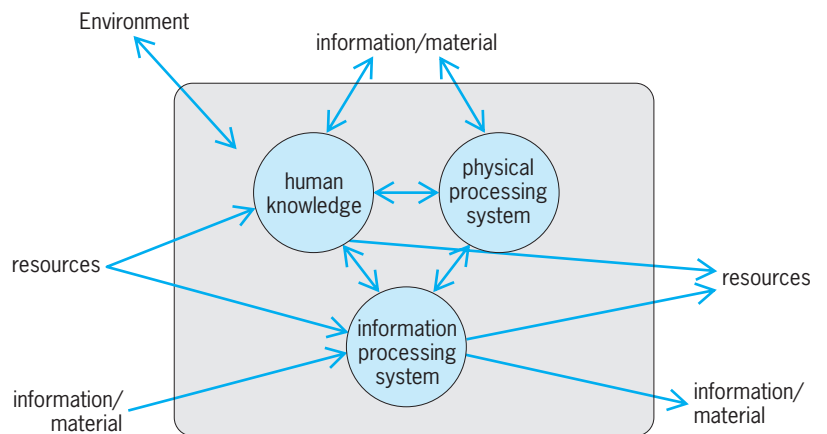

Fig. 7. Holonic system element. (*Holonic Manufacturing Systems Consortium*)

can also be used to find solutions to constrained optimization problems such as those required for robot path planning. *See* NEURAL NETWORK.

Fuzzy systems. Fuzzy sets, fuzzy logic, and fuzzy control of complex systems are characterized by high degree of nonlinearity, strong coupling, and time variance. Fuzzy logic is a superset of boolean logic that has been extended to handle the concept of partial truth that falls between "completely true" and "completely false." A fuzzy system uses knowledge that experts have accumulated over time on topics for which it is very hard to establish precise mathematical models, and expresses this knowledge in explicit "if-then" linguistic rules. The design of a fuzzy system involves the following steps. Fuzzification maps the actual input variables to their fuzzy subsets according to membership functions. Fuzzy inference and reasoning evaluate each fuzzy subset and integrate all the variables to obtain the output fuzzy subset based on fuzzy operators and control rules. Defuzzification is the process of converting fuzzified values into crisp variables to command the physical plants. Fuzzy systems have been applied in many engineering fields, including process control, artificial intelligence, system diagnosis, and pattern recognition, as well as in economics, sociology, management, and medical research. Fuzzy systems provide an important tool for implementing the control of sensory interactive robotic systems. *See* FUZZY SETS AND SYSTEMS.

Human-machine interfaces. A telerobotic system augments humans by allowing them to extend their ability to perform complex tasks in remote locations. It is a technology that couples the human operator's visual, tactile, and other sensory perception functions with a remote manipulator or mobile robot. These systems are useful for performing tasks in environments that are dangerous or not easily accessible for humans. Telerobotic systems are used in nuclear handling, maintenance in space, undersea exploration, and servicing electric transmission lines. Perhaps the most important sensory data needed for telepresence are feedback of visual information, robot position, body motion and forces, as well as tactile information. Master-slave systems have been developed in which, for example, a hand controller provides control inputs to an articulated robotic manipulator. These systems are capable of feeding back forces felt by the robot to actuators on the exoskeletal master controller so that the operator can "feel" the remote environment. *See* HUMAN-MACHINE SYSTEMS; REMOTE MANIPULATORS.

Programming languages. Software development is a significant part of the cost of engineering a robotic system. System architectures and programming environments are generally not able to handle the demands of flexibility, performance, and complexity that are required by modern production processes. For this reason, high-level, textual programming languages are provided for specifying a sequence of actions that allow the operator to program external sensors, such as vision and force, and to use computer-aided design (CAD) systems to program the robot off-line. Programming languages have been developed for use in factory automation as well as robots in unstructured environments. Some of these languages have attained several generations of development, and they now include concepts from knowledge-based systems, model-based systems, discrete-event systems and Petri nets (a graphical method for analyzing and modeling discrete-event dynamic systems), graphical languages, and object-oriented programming languages, as well as related methodologies and tools from database engineering and communication networks. Building on textual languages developed for industrial robots, these programming environments provide a common interface to devices and sensors that would otherwise require specialized programming by people with different levels of skills. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING; COMPUTER PROGRAMMING; PROGRAMMING LANGUAGES.

Telepresence. This technology is a combination of interface technologies that enables users to intuitively interact with immersive and dynamic computer-generated environments. It frequently involves the use of computer-generated visuals to enhance the perception of the physical environment. Such systems, called virtual reality (VR), have cameras at the worksite that are slaved to the operator's head movement. The operator views the scene at the worksite through a binocular display mounted on a helmet. While virtual reality has been used by flight simulators and military displays for many years, the low cost of computers and peripherals, such as data gloves, is helping to make this technology available to new applications of robotics. *See* AIRCRAFT TESTING; COMPUTER GRAPHICS; VIRTUAL REALITY.

Robotic system applications. An early justification for industrial robots was to replace humans in dangerous and repetitive labor, and to replace fixed-function automation with more flexible equipment that could be reprogrammed for new situations and adapted through the use of sensory inputs from the external world. Because of their initial limited sensory abilities, robots were best suited to those tasks with clearly prescribed paths of motion. For these reasons, most applications of robots were found in tasks such as material transfer, machine loading and unloading, spot welding, and spray painting.

The robots employed for these purposes range from small models with lifting capacities of several pounds to very large types that can lift many hundreds of pounds. Overhead gantry robots with three or four axes have been developed for heavy lifting tasks involving payloads up to 1000 lb (450 kg). Designs also vary according to the size and shape of the work envelope. For example, many robots are designed to reach sideways and backwards so that they can achieve higher productivity by tending several machines at the same time. Graphical simulation is used to design and evaluate a workcell layout before it is built. The robot motion can be programmed on the simulation and downloaded to the robot controller. Companies market software systems that include libraries of commercially available robots and

postprocessors for off-line robot programming. *See* SIMULATION.

Automated manufacturing. Advances in robotics stimulated the development of industrial applications that exploited the robot's speed and accuracy in such tasks as precision drilling and other light machining, gluing and fastening, simple assembly, packing and sealing, and testing and measurement. Some of the machines developed for these tasks have repeatability of movement better than 0.001 in. (25 micrometers), yet are able to move at high rates throughout the work envelope. Usually they are not capable of heavy payloads. These robots are often fitted with highly specialized end effectors that are automated tools themselves. Their control systems are more elaborate and frequently include external sensory capabilities. Devices of this type may, for example, place odd-sized components on electronic circuit boards, with very small clearances, after visually locating positioning marks on the circuit board.

Flexible manufacturing requires high levels of sensory interaction and processing. Robots must be capable of locating and positioning three-dimensional parts, mating them to high tolerances, and performing complex joining operations, such as screwing bolts and riveting skins on aircraft wings. Robots with redundant degrees of freedom (more than six for a manipulator) can reach a specific position and orientation in more than one way, for example, as needed for inspection inside the body of a car that is being assembled. Robots may be equipped with sensors that enable them to feel and adaptively respond to forces from sticking and jamming of parts during positioning and mating. Since assembly requires the presence of many parts in the work space, intelligent robots have been developed with visual sensing that can identify and locate parts without the need for complex, inflexible part-presentation equipment such as feeders, positioners, and custom end effectors. Vision capabilities required for assembly range from simple to complex. In many cases, parts have only a few stable positions in which they rest, so that recognition and location of two-dimensional outlines is sufficient to acquire and orient them. *See* FLEXIBLE MANUFACTURING SYSTEM.

Autonomous mobile robots. Simple mobile robots have long been employed for material movement in factories, mail delivery in office buildings, and similar tasks requiring intelligent choice of path and a limited ability to avoid obstacles. Mobile robots for service in the home and office are becoming available. Developments in robot control and sensing, combined with advances in mechanisms for robot locomotion, are expanding the applications for mobile robots. Among these applications are robots for guard and sentry duty, hospital care, and commercial cleaning.

Research and development has resulted in robots for agriculture, house cleaning, construction, mining, fire fighting, rescue, surveillance, handling of hazardous materials encountered in waste cleanup, military applications on land, and undersea exploration. In some cases, this involves the provision of sensory, world-modeling, and control devices for existing machines which already possess mobility and actuators, such as earth-moving equipment. In other cases, entirely new devices and system architectures must be developed. The holonic system paradigm described previously is well suited for these applications because of its ability to adapt to changing environments.

One of the difficulties in using autonomous mobile robots is the complex sensory requirement imposed by navigation in unstructured environments. Control requirements for mobile robots are considerably more difficult than those for industrial robots. Since the control and perception processes in a mobile robot decompose naturally across a set of layers, and may be implemented on different processors, a holonic systems architecture can be used for control. Because the control of autonomous mobile robots requires decision making with large amounts of poorly structured knowledge, expert systems and fuzzy logic are useful tools.

Prosthetic devices and aids for disabled persons. Prosthetic devices have been developed to provide replacements for lost or damaged human functions. Prostheses for upper limbs must be reliable and versatile so that they can partially replace the functionalities of a human arm and hand. The earliest upper-limb prostheses offered simple motional functionalities without providing anthropomorphic features. Command signals for electrical prostheses may be obtained from electromyographic (EMG) signals on the surface of the skin resulting from muscle action or directly from nerve signals with implanted sensors. Commercially available prostheses, however, have not exploited modern robotic technologies. Most of these devices lack fingertip tactile sensing, a feature that humans routinely use in most daily tasks. The largest number of disabilities involves the loss or impairment of lower limbs. An active exoskeletal walking mechanism can significantly improve a person's capability for movement in daily living. Prototype mechanisms have been built to assist a disabled person in walking. With the aid of such mechanisms, a person can walk smoothly using a rolling support or crutches. The device also can be used as a training device for patients with functional disabilities. Exoskeletel mechanisms can also provide therapeutic exercise of a patient's arm and shoulder during stroke rehabilitation. *See* BIOMEDICAL ENGINEERING; MEDICAL CONTROL SYSTEMS; PROSTHESIS.

Micromachines. Miniature mechanical and sensor systems are being developed using the technology of integrated circuit fabrication to etch micro-electromechanical systems (MEMS) directly onto silicon chips. Through the use of multiple structural polysilicon layers, it is possible to generate structures on the scale of a few hundred micrometers, such as a 16-joint anthropomorphic skeleton of a hand or a cantilever for measuring gas flow. By integrating thin-film transistors into the same process, interface electronics may be fabricated simultaneously with the mechanical structure. Pressure sensors,

accelerometers, resonators, and micromotors have been built. Another process generates a photoresist pattern using deep x-ray lithography to create a design on a conductive substrate. This process is useful for making small gears, linkages, or other parts. *See* MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS).   William A. Gruver

Bibliography. J. Craig, *Introduction to Robotics: Mechanics and Control*, 2d ed., 1989; R. Dorf (ed.), *International Encyclopedia of Robotics: Applications and Automation*, 3 vols., 1988; J. Engelberger, *Robotics in Practice*, AMACOM, New York, 1980; J. Engelberger, *Robotics in Service*, MIT Press, Cambridge, MA, 1989; L. Fausett, *Fundamentals of Neural Networks*, Prentice Hall, 1994; J. W. Gardner, *Microsensors: Principles and Applications*, Wiley, 1994; W. A. Gruver and J. Boudreaux, *Intelligent Manufacturing: Programming Environments for CIM*, Springer-Verlag, 1993; P. McKerrow, *Introduction to Robotics*, Addison Wesley, 1991; Micromachines on the march, *IEEE Spectrum*, pp. 20–31, May 1993; T. Sheridan, Telerobotics, *Automatica*, 23(4):487–507, 1988; Virtual reality is for real, *IEEE Spectrum*, pp. 22–39, October 1993; P. H. Winston, *Artificial Intelligence*, Addison Wesley, 1984; E. H. van Leeuwen and D. Norrie, Inteligent manufacturing: Holons and holarchies, *Manuf. Eng.*, 76(2):86–88, April 1997.

Roche limit

The closest distance which a satellite, revolving around a parent body, can approach the parent without being pulled apart tidally. The simplest formal definition is that the Roche limit is the minimum distance at which a satellite can be in equilibrium under the influence of its own gravitation and that of the central mass about which it is describing a circular orbit. If the satellite is in a circular orbit and has negligible mass, the same density as the primary, and zero tensile strength, the Roche limit is 2.46 times the radius of the primary.

The concept of an equilibrium state was first given by Édouard Roche in 1849. The radius r of the Roche limit appropriate to satellite formation is given by the equation below, where B is the radius and ρ the

$$r = 2.46 \left(\frac{\rho}{\rho_s} \right)^{1/3} B$$

mean density of the parent planet, and ρ_s is the mean density of the satellite. This is the radius at which the surface of a homogeneous, fluid satellite can remain in stable hydrostatic equilibrium.

The popular definition of Roche's limit, that is, the distance from the planet at which a satellite would suffer tidal loss of particles, depends on properties of the material and the shape and density of the satellite. If, for instance, the material can flow to adopt the hydrostatic equilibrium shape, then the body will be elongated in the direction of the radius vector. For a binary system, the Roche limit is

defined by the value of the Jacobi constant such that the zero velocity surfaces around the two bodies intersect in the lagrangian libration point L_2 between the finite masses. This implies an upper limit to the size of the components. While magnetohydrodynamic forces act on the outer layers of stars and the theory holds only for the circular restricted problem with the masses acting as point masses, the Roche limit seems to correspond to the observational data. When a star has exhausted the supply of hydrogen in its core, its radius will increase by a factor of 10 to 100. A star in a binary system may then exceed its Roche limit, material will thus escape from that star, and its companion will receive the excess material. *See* BINARY STAR; CELESTIAL MECHANICS.   P. K. Seidelmann

Rock

A relatively common aggregate of mineral grains. Some rocks consist essentially of but one mineral species (monomineralic, such as quartzite, composed of quartz); others consist of two or more minerals (polymineralic, such as granite, composed of quartz, feldspar, and biotite). Rock names are not given for those rare combinations of minerals that constitute ore deposits, such as quartz, pyrite, and gold. In the popular sense rock is considered also to denote a compact substance, one with some coherence; but geologically, friable volcanic ash also is a rock. A genetic classification of rocks follows:

Igneous
   Intrusive
      Plutonic (deep)
      Hypabyssal (shallow)
   Extrusive
   Flow
      Pyroclastic (explosive)
Sedimentary
   Clastic (mechanical or detrital)
   Chemical (crystalline or precipitated)
   Organic (biogenic)
Metamorphic
   Cataclastic
   Contact metamorphic and pyrometasomatic
   Regional metamorphic (dynamothermal)
Hybrid
   Metasomatic
   Migmatitic

Exceptions to the requirement that rocks consist of minerals are obsidian, a volcanic rock consisting of glass; and coal, a sedimentary rock which is a mixture of organic compounds. *See* COAL; VOLCANIC GLASS.

Igneous rocks. Igneous rocks are those that have solidified from a molten condition. The parent material is magma—a natural, hot, mutual solution of silicates with minor amounts of water and other volatiles. Igneous rocks are divided into those which crystallized before magma reached the Earth's

TABLE 1. Simplified classification of major igneous rocks on the basis of composition and texture

	SiO$_2$-rich (acidic) ←			→ SiO$_2$-poor (basic)	
	Light colored ←	——— Gray ———	—— Dark colored ——	→ Black	
Mineral composition:	Quartz, potash feldspar, biotite	Potash feldspar, biotite, or amphibole	Sodic plagioclase, hornblende, or augite	Augite, olivine, hypersthene, calcic plagioclase	Olivine, enstatite, augite
Intrusive Medium-grained	Granite*	Syenite	Diorite	Gabbro	Peridotite
Extrusive Fine-grained to aphanitic	Rhyolite ←	Trachyte ——— Felsite ———	Andesite →	Basalt	
Porphyritic	Rhyolite porphyry	Tracyte porphyry	Andesite porphyry	Basalt porphyry	
Glassy	Obsidian				
Vesicular	Pumice			Scoria	
Fragmental		Tuff and agglomerate of each type			

*Exceptionally coarse-grained rock of general granitic composition is pregmatite.

surface (intrusive rocks) and those that solidified at the surface, some as layers of lava (the extrusive flow rocks) and others as pyroclastic debris in explosive eruption (**Table 1**). *See* MAGMA; PYROCLASTIC ROCKS.

Chief elements in igneous rocks are oxygen, O; silicon, Si; aluminum, Al; iron, Fe; magnesium, Mg; calcium, Ca; sodium, Na; and potassium, K. Compositions range from about 40% SiO$_2$ (peridotites) to as much as 70% SiO$_2$ (granites). Silica-poor rocks contain relatively large amounts of Ca, Mg, and Fe^2 (basic rocks), whereas silica-rich types contain larger amounts of Na and K (acidic rocks). *See* IGNEOUS ROCKS.

Sedimentary rocks. Clastic sedimentary rocks (consisting of mechanically transported particles) are subdivided on the basis of particle size (**Table 2**). Those having intermediate and fine-grain sizes are further subdivided on the basis of composition (see **illus.**). Other significant clastic rocks are those consisting of detrital calcite, the calcarenites and calcilutites.

Textures of clastic rocks derive from grain size, sorting, form, and arrangement. Form includes sphericity (shape), the degree to which a particle approximates a sphere; and roundness, the measurement of the sharpness of edges and corners. Only glacial sedimentary rocks (tillite) do not show layering or stratification.

Chemical sedimentary rocks are those precipitated from ocean, lake, and ground water. The most important ones are shown in the following list:

Rock	Chief Mineral
Chert	Chalcedony, SiO$_2$
Limestone	Calcite, CaCO$_3$
Travertine (spring deposit)	Calcite, CaCO$_3$
Dolomite	Dolomite, CaMg(CO$_3$)$_2$
Phosphorite	Apatite, Ca$_{10}$(PO$_4$)$_5$(CO$_3$)F$_3$
Salines (evaporites)	
Rock salt	Halite, NaCl
Rock anhydrite	Anhydrite, CaSO$_4$
Rock gypsum	Gypsum, CaSO$_4$ · 2H$_2$O

Organic sedimentary rocks include (1) siliceous types made up of opaline tests of diatoms, diatomite, or radiolaria, radiolarite; (2) calcareous types, consisting of calcite shell, namely, shell limestone and coquina; and (3) carbonaceous types, namely, coal and other accumulations of altered plant debris. *See* SEDIMENTARY ROCKS; SEDIMENTOLOGY.

Metamorphic rocks. Metamorphic rocks owe their complexity of composition and texture not only to the existence of several types of metamorphism but also to the application of these types under different intensities to a variety of parent rocks; thus, both sedimentary and igneous rocks may be metamorphosed (**Table 3**). Regional metamorphic rocks are distinguished by foliation, a parallel orientation of platy or

TABLE 2. Size classification of clastic sedimentary particles and aggregates

| Size, in. (mm)	Particle	Aggregate
>10 (>256)	Boulder	Gravel, conglomerate (psephite, rudite)
10–2.5 (256–64)	Cobble	Breccia (angular)
2.5–0.16 (64–4)	Pebble	
0.16–0.079 (4–2)	Coarse sand	Sandstone (psammite, arenite)
0.079–0.0025 (2–0.63)	Sand	
0.0025–0.00015 (0.63–0.0039)	Silt	Siltstone (pelite, lutite)
0.00015 (<0.0039)	Clay	Clay
		Shale

TABLE 3. Simplied classification of metamorphic rocks, with selected examples

Parent rocks	Contact metamorphism	Regional metamorphism		
		Low grade	Intermediate grade	High grade
Sandstone, arkose	Quartzite (quartz)	Quartzite and quartz-feldspar gneiss		
Shale	Hornfels (andalusite, cordierite)	Slate, phyllite, (chlorite, muscovite)	Mica schist (biotite, garnet, kyanite, staurolite)	Sillimanite gneiss (sillimanite, almandite)
Limestone	Marble (calcite)	Marble and calc-silicate gneiss		
		(Calcite, tremolite)	(Calcite, wollastonite)	(Calcite, diopside, anorthite)
Basalt	Hornfels (plagioclase, hypersthene)	Greenschist (chlorite, albite, epidote)	Amphibole schist, (plagioclase, actinolite)	Amphibolite (andesine, hornblende, garnet)

TABLE 4. Physical properties of some common rocks

Rock	Specific gravity	Porosity, %	Compressive strength, lb/in.^{2}*	Tensile strength, lb/in.^{2}*
Igneous				
Granite	2.67	1	30,000–50,000	500–1000
Basalt	2.75	1	25,000–30,000	
Sedimentary				
Sandstone	2.1–2.5	5–30	5,000–15,000	100–200
Shale	1.9–2.4	7–25	5,000–10,000	
Limestone	2.2–2.5	2–20	2,000–20,000	400–850
Metamorphic				
Marble	2.5–2.8	0.5–2	10,000–30,000	700–1000
Quartzite	2.5–2.6	1–2	15,000–40,000	
Slate	2.6–2.8	0.5–5	15,000–30,000	

*1 lb/in.^{2} = 6.9 kPa.
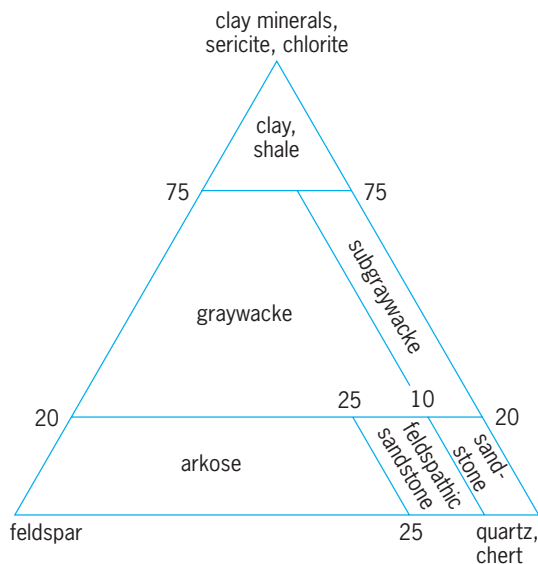
Classification of psammitic and pelitic rocks based on proportions of the three most common clastic minerals: quartz, feldspars, and clays—excluding calcilutites and calcarenites. (*After E. W. Heinrich, Microscopic Petrography, McGraw-Hill, 1956*)

prismatic minerals. *See* METAMORPHIC ROCKS; METAMORPHISM.

Physical properties and behavior. When stresses are applied to rocks, either natural (those active in mountain building) or human-made, resulting from loading with structures (such as dams), deformation (strain) may result. Rocks subjected to stress normally undergo three deformation stages: elastic—the rock returns to its original size or shape upon withdrawal of stress; plastic—beyond a limiting stress (elastic limit) there is only partial restoration upon stress removal; and fracture—breakage with further increase in stress. With increases in confining pressure (load of overlying rocks) and temperature, the interval between the elastic limit and fracture increases. Thus rocks that behave as brittle substances near the Earth's surface, failing by fracturing, will at depth be deformed plastically by solid flow. The effects of stress depend on physical properties (**Table 4**). *See* ENGINEERING GEOLOGY; HIGH-PRESSURE MINERAL SYNTHESIS; ROCK MECHANICS; STRUCTURAL GEOLOGY.

Rock cycle. Igneous rocks exposed at the Earth's surface are subject to weathering, which alters them chemically and physically. Such material when transported, deposited, and consolidated becomes

sedimentary rock, which, through heat and pressure, may be converted to metamorphic rock. Both sedimentary and metamorphic rocks also may be weathered and transformed into younger sediments. Deeply buried metamorphic rocks may be remelted to yield new igneous material. *See* LITHOSPHERE; PETROGRAPHY; PETROLOGY; STONE AND STONE PRODUCTS.

E. William Heinrich

Rock, electrical properties of

The effect of changes in pressure and temperature on electrical properties of rocks. There has been increasing interest in the electrical properties of rocks at depth within the Earth and the Moon. The reason for this interest has been consideration of the use of electrical properties in studying the interior of the Earth and its satellite, particularly to depths of tens or hundreds of kilometers. At such depths pressures and temperatures are very great, and laboratory studies in which these pressures and temperatures are duplicated have been used to predict what the electrical properties at depth actually are. More direct measurements of the electrical properties deep within the Earth have been made by using surface-based electrical surveys of various sorts. An important side aspect of the study of electrical properties has been the observation that, when pressures near the crushing strength are applied to a rock, marked changes in electrical properties occur, probably caused by the development of incipient fractures. Such changes in resistivity might be used in predicting earthquakes, if they can be measured in the ground. *See* GEOELECTRICITY; GEOPHYSICAL EXPLORATION.

Electrical zones in the Earth. Attempts to measure the electrical properties of rocks to depths of tens or hundreds of kilometers in the Earth indicate that the Earth's crust is zoned electrically. The surface zone, with which scientists are most familiar, consists of a sequence of sedimentary rocks, along with fractured crystalline and metamorphic rocks, all of which are moderately good conductors of electricity because they contain relatively large amounts of water in pore spaces and other voids. This zone, which may range in thickness from a kilometer to several tens of kilometers, has conductivities varying from about $1/2$ ohm-m in recent sediments to 1000 ohm-m or more in weathered crystalline rock.

The basement rocks beneath this surface zone are crystalline, igneous, or metamorphic rocks which are much more dense, having little pore space in which water may collect. Since most rock-forming minerals are good insulators at normal temperatures, conduction of electricity in such rocks is determined almost entirely by the water in them. As a result, this part of the Earth's crust is electrically resistant. There are numerous experimental difficulties involved in trying to measure the electrical properties of this part of the Earth's crust, but it appears that the resistivity lies in the range 10,000–1,000,000 ohm-m.
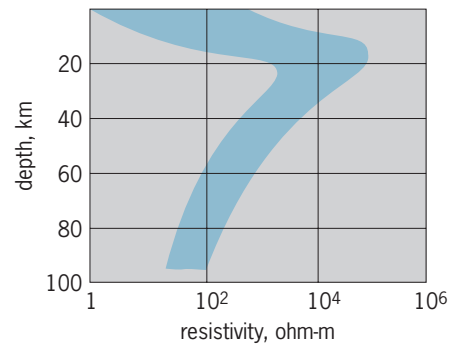
At rather moderate depths beneath the surface

Fig. 1. Probable range of electrical resistivity through the crust and into the upper mantle. 1 km = 0.6 mi.

of the second zone, resistivity begins to decrease with depth. This decrease is considered to be the result of higher temperatures, which almost certainly are present at great depths. High temperatures lead to partial ionization of the molecular structure of minerals composing a rock, and the ions render even the insulating minerals conductive. As temperature increases with depth, this effect becomes more pronounced and resistivity decreases markedly. The maximum resistivity in the Earth's crust probably occurs at depths of only 3–6 mi (5–10 km) below the surface of the crystalline basement. At depths of 18–24 mi (30–40 km), corresponding to the top of the mantle, resistivity is only about a few hundred ohm-meters. Beyond this depth, resistivity seems to decrease slightly with increasing depth, but again drops sharply at depths around 420 mi (700 km). At greater depths the resistivity is less than 1 ohm-m. A profile of resistivity as it varies downward into the Earth is shown in **Fig. 1**.

At relatively shallow depths information is gained by lowering equipment into bore holes to measure the electrical properties of rocks directly. Another approach is the use of low-frequency currents introduced into the Earth at the surface. These methods are restricted to use in studying resistivities at depths less than 6 mi (10 km). Information from greater depths is obtained by studying slow variations in the Earth's magnetic field and the currents induced in the Earth by these variations. Fluctuations in the magnetic field with periods measured in months penetrate to distances up to 600 mi (1000 km), and they may be used in estimating conductivity to these depths.

Temperature effects. Laboratory investigations have also been pursued to find how the electrical properties of rocks depend on the pressure and temperature deep in the Earth. Inasmuch as the temperature at depths of 24–36 mi (40–60 km) is probably in the range 1470–2200°F (800–1200°C) and the pressure is in the range 140,000–210,000 lb/in.^{2} (10,000–15,000 kg/cm^2), these conditions are difficult to duplicate in the laboratory. Both electrical conductivity and dielectric constant have been measured by several researchers at pressures to 700,000 lb/in.^{2} (50,000 kg/cm^2) and temperatures to 2200°F (1200°C), using dry rock samples. The most obvious conclusion from such studies is that
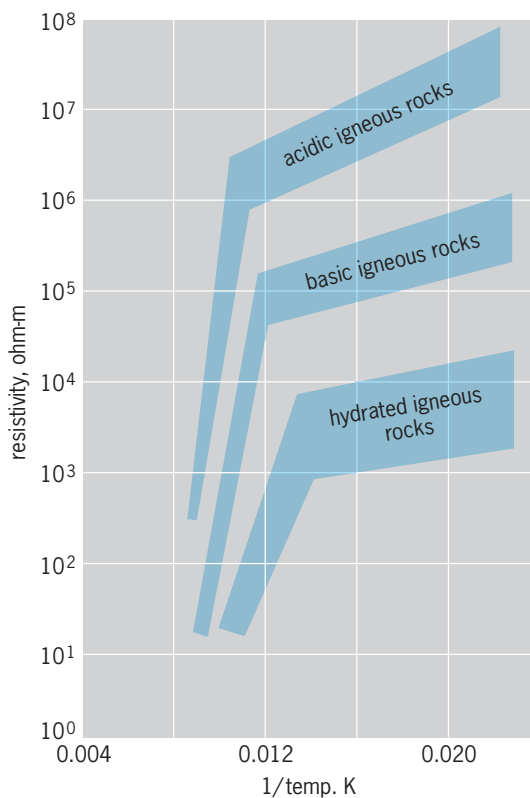
Fig. 2. **Variation of resistivity with temperature in dry igneous rocks. Hydrated samples contained only water of hydration, not free water. At the melting point, conduction increases abruptly.** $°F = (K \times 1.8) - 459.67$.

for this range of conditions the effect of temperature on electrical properties is much more profound than that of pressure. An increase in temperature from 68 to 1830°F (20 to 1000°C) will reduce the resistivity of a dry rock by a factor of 10^6. On the other hand, an increase in pressure from 0 to 700,000 lb/in.^{2} (50,000 kg/cm^2) changes the resistivity of a dry rock by less than 100%. Typical behavior for the resistivity of dry rocks as temperature is raised is shown by curves in **Fig. 2**. At temperatures up to about half the melting point, conduction increases relatively slowly with increasing temperature, and it is thought that conduction is due to impurities, which contribute ions under weak thermal excitation. At temperatures above half the melting point, conduction increases rapidly with increasing temperature, with conduction being caused by ions torn from normal lattice positions in the crystals by violent thermal agitation. At the melting point, conduction increases abruptly as a rock melts and becomes highly ionized.

It has been observed that silicic rocks, such as granite, granodiorite, and rhyolite, are poorer conductors of electricity at all temperatures than other rock types. Rocks, such as gabbro, basalt, and dunite, which are rich in ferromagnesian minerals, are more conductive than the silicic rocks by about an order of magnitude. This is believed to result from the fact that the principal ions which contribute to conduction are the small ferromagnesian ions, which are more abundant in the dark rocks.

Pressure effects. Typical curves for the behavior of resistivity and of dielectric constant with the application of pressure are shown in **Figs. 3** and **4**. The dielectric constant is found to increase by 10–20% as the pressure on a rock sample is increased to 1400–2800 lb/in.^{2} (100–200 kg/cm^2) and then to increase more slowly as the pressure is elevated further. The increase in dielectric constant is due in part to an increase in density of the rocks under pressure.

Two forms of behavior have been recognized in studies of the behavior of resistivity at pressures up to 700,000 lb/in.^{2} (50,000 kg/cm^2). With some rocks resistivity decreases uniformly with increasing pressure; in other rocks resistivity decreases at low applied pressures but increases at high pressures. A uniform decrease has been observed in rocks such as basalt, diabase, and amphibolite, while an increase in resistivity over part of the pressure range has been observed in microcline-rich rocks. The increase in resistivity may be associated with a decrease in the mobility of charge carriers as the crystal lattice is compressed.

The effect of pressure on the resistivity of rocks in which conduction is determined by water content rather than by the crystal matrix is quite different.
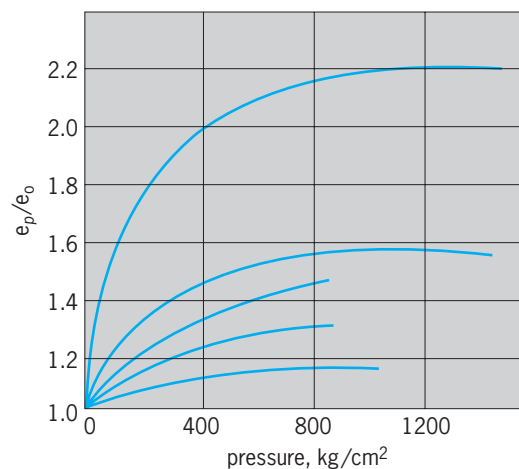

Fig. 3. **Effect of pressure on the dielectric constants of five samples of granite. The ratio e_p/e_0 is the ratio of the dielectric constant measured at elevated pressure to dielectric constant measured at zero pressure. 1 kg/cm^2 = 0.98×10^5 Pa = 14 lb/in.^{2}.**
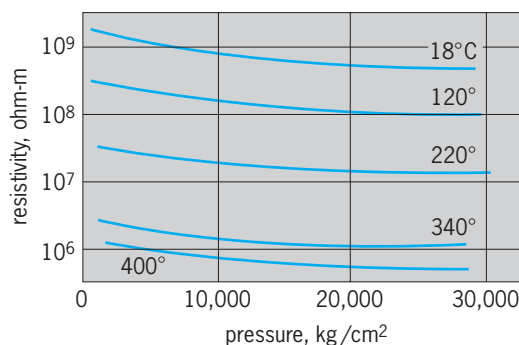

Fig. 4. **Resistivity of a basalt sample measured as a function of pressure at various temperatures. 1 kg/cm^2 = 0.98×10^5 Pa = 14 lb/in.^{2}; $°F = (°C \times 1.8) + 32$.**
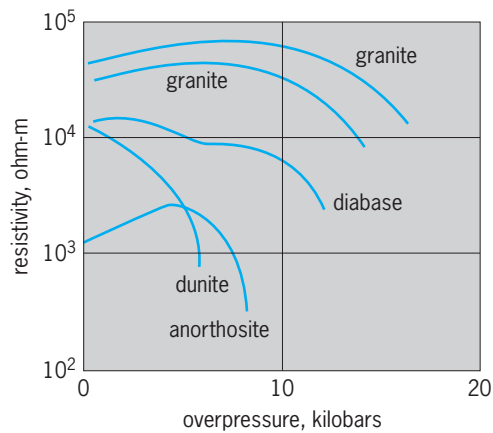
Fig. 5. Effect of greater pressure on framework of a rock than on fluid in pores in water-bearing rocks. 1 kilobar $= 10^8$ Pa.

Several investigators have studied the effect of having greater pressure on the crystal framework than on the pore fluid. As an overpressure is applied to the rock skeleton, the framework starts to close in on the void spaces, reducing their volume and increasing the resistivity. It has also been observed that, if the overpressure is made great enough, resistivity begins to increase again, a phenomenon which is attributed to the opening of small cracks prior to rock failure, with the movement of water into these cracks to contribute to conduction. The behavior of resistivity for a number of rock specimens as a function of frame overpressure is shown by the curves in **Fig. 5**. According to W. F. Brace and A. S. Orange, observations of such changes in resistivity might be used to detect minute changes in the structure of rock under stress. They observe that the buildup of strain in the Earth preceding earthquakes ought to be accompanied by changes in resistivity, and that these changes might warn of an impending earthquake. *See* HIGH-PRESSURE MINERAL SYNTHESIS; PETROLOGY.

Electrical zones in the Moon. In planning the exploration of the Moon, some thought has been given to the use of measurements of electrical properties as a means for studying the Moon's interior. Although there is little information on the electrical properties of the Moon, the knowledge gained about the Earth can be used to make an intelligent guess. The major difference is the lack of an atmosphere and moisture, which means that the surface rocks on the Moon must be quite different from those on Earth. Measurements with radar and infrared sensors have indicated that the surface of the Moon must be covered with a rock froth or soil which has a very low density—the rocks may be only about 10% solid material. Inasmuch as there is probably no moisture in this rock, it seems very likely that the electrical resistivity must be very large, perhaps 10^{10} ohm-m.

It is believed that temperature increases with depth inside the Moon, though perhaps not as rapidly as in the Earth. The temperature at the center of the Moon may be $3140°F$ (2000 K) and, as a result, the rocks in the interior can be expected to be con-

ductive, just as they are on the Earth. The resistivity at the center of the Moon may be of the order of 1–100 ohm-m.

The change in resistivity between the highly resistant surface and the conductive interior depends very much on the composition of the Moon. Some scientists believe that the interior contains as much water as the rocks within the Earth, at depths beyond which evaporization into space has been able to take place. Others believe that the Moon had a different origin from the Earth, and thus never contained very much water. If water is present, one would expect the resistivity to drop very rapidly over the first few kilometers in depth, as water contributes to electrical conduction. If the Moon is devoid of water, the resistivity should change more gradually with depth and become low only at depths of many hundreds of kilometers, where the temperatures are high. The presence of water at moderate depths would be very important to future lunar exploration—it might be used as fuel for trips deeper into space, or even to provide a tenuous atmosphere to protect humans' installations on the Moon. Therefore it is likely that electrical surveys for water on the Moon will have a high priority in the early stages of lunar exploration. *See* MOON.   George V. Keller

Bibliography. M. B. Dobrin and C. H. Savit, *Introduction to Geophysical Prospecting*, 4th ed., 1988; P. Kearey and M. Brooks, *Introduction to Geophysical Exploration*, 2d ed., 1991; G. V. Keller, *Electrical Methods of Geophysical Prospecting*, 2d ed., 1982; E. S. Robinson and C. Coruh, *Basic Exploration Geophysics*, 1988; R. E. Sheriff, *Geophysical Methods*, 1989.

Rock age determination

Finding the age of rocks based on the presence of naturally occurring long-lived radioactive isotopes of several elements in certain minerals and rocks. Measurements of rock ages have enabled geologists to reconstruct the geologic history of the Earth from the time of its formation 4.6×10^9 years ago to the present. Age determinations of rocks from the Moon have also contributed to knowledge of the history of the Moon, and may someday be used to study the history of Mars and of other bodies within the solar system. *See* RADIOACTIVITY.

Principles of dating. Many rocks and minerals contain radioactive atoms that decay spontaneously to form stable atoms of other elements. Under certain conditions these radiogenic daughter atoms accumulate within the mineral crystals so that the ratio of the daughter atoms divided by the parent atoms increases with time. This ratio can be measured very accurately with a mass spectrometer, and is then used to calculate the age of the rock by means of an equation based on the law of radioactivity.

According to this law, the number of radiogenic daughters (D) in a unit weight of rock is related to the number of radioactive parent atoms remaining

Parent-daughter pairs used for dating rocks and minerals

Parent	Daughter	Half-life, 10^9 years
Potassium-40	Argon-40	11.8
Potassium-40	Calcium-40	1.47
Rubidium-87	Strontium-87	48.8
Samarium-147	Neodymium-143	107
Rhenium-187	Osmium-187	43
Thorium-232	Lead-208	14.008
Uranium-235	Lead-207	0.7038
Uranium-238	Lead-206	4.468

(P) by Eq. (1). Here e is a mathematical constant ap-

$$\frac{D}{P} = e^{\lambda t} - 1 \qquad (1)$$

proximately equal to 2.7182, λ is the decay constant of the parent, and t may be regarded as the age of the rock or mineral. The value of the decay constant is characteristic of each radioactive atomic species, and must be known from experimental determinations. It is related to the half-life ($t_{1/2}$) of the parent by the equation $\lambda = 0.693/t_{1/2}$, where the half-life is defined as the time required for one-half of a given number of radioactive atoms to decay. The radioactive atoms used for dating rocks and minerals have very long half-lives, measured in billions of years. They occur in nature only because they decay very slowly. The pairs of parents and daughters used for dating are listed in the **table**. *See* DATING METHODS.

Rubidium-strontium method. Rubidium (Rb) is an alkali metal whose chemical properties are similar to those of potassium. Therefore, rubidium occurs in small amounts in all minerals in which potassium is a major constituent. The most common rubidium-bearing minerals are the micas (biotite and muscovite), the feldspars (orthoclase and microcline), and certain of the clay minerals (illite and glauconite). *See* CLAY MINERALS; FELDSPAR; MICA.

The radioactivity of rubidium is due to rubidium-87, which decays to stable strontium-87 (^{87}Sr) by emitting a beta particle from its nucleus. The abundance of the radiogenic strontium-87 therefore increases with time at a rate that is proportional to the Rb/Sr ratio of the rock or mineral. This relationship is expressed by Eq. (2), where ^{87}Sr/^{86}Sr and ^{87}Rb/^{86}Sr

$$\frac{^{87}\text{Sr}}{^{86}\text{Sr}} = \left(\frac{^{87}\text{Sr}}{^{86}\text{Sr}}\right)_0 + \frac{^{87}\text{Rb}}{^{86}\text{Sr}}(e^{\lambda t} - 1) \qquad (2)$$

are the ratios of these atoms in the sample to be dated and (^{87}Sr/^{86}Sr)$_0$ is the value of this ratio in the rock or mineral at the time of its formation. *See* BETA PARTICLES.

The effectiveness of the Rb-Sr method of dating is enhanced by the difference in the geochemical properties of strontium and rubidium. Strontium is an alkaline earth element that replaces calcium in minerals such as plagioclase feldspar and apatite, a calcium phosphate. The crystallization of a silicate liquid called magma may produce different varieties of igneous rocks whose Rb/Sr ratios have a wide range of values depending on whether they are composed of Rb-rich or Sr-rich minerals. Such comagmatic rocks all have the same age (t) and the same initial ^{87}Sr/^{86}Sr ratio. When their ^{87}Sr/^{86}Sr and ^{87}Rb/^{86}Sr ratios are plotted (**Fig. 1**), each rock forms a point that lies on a straight line called an isochron. The age of the rocks that form the isochron is indicated by its slope (m) because $m = e^{\lambda t} - 1$ and therefore $t = (1/\lambda) \ln (m + 1)$. The intercept of the isochron yields the value of the initial ^{87}Sr/^{86}Sr ratio. *See* MAGMA.

The whole-rock Rb-Sr isochron method outlined above has been used to date not only igneous but also metamorphic rocks. The method is particularly well suited to the dating of very old rocks such as the ancient gneisses near Godthaab in Greenland, which are almost 3.8×10^9 years old. This method has also been used to date rocks from the Moon and to determine the age of the Earth by analyses of stony meteorites.

The Rb-Sr method has had limited success in the dating of sedimentary rocks. Such rocks commonly contain fossils that indicate their geologic ages. However, only dating methods based on radioactivity give an absolute time scale. The clay mineral glauconite has yielded the only reliable dates for sedimentary rocks by the Rb-Sr and the potassium-argon (K-Ar) method. Certain other clay minerals called smectite may also be suitable for dating. Nevertheless, the dating of sedimentary rocks remains a challenge for the future.

Potassium-argon method. The decay of potassium-40 to argon-40 was one of the first methods of dating to be developed in the 1950s. It is based on the assumption that all of the atoms of radiogenic argon-40 that form within a potassium-bearing mineral accumulate within it. This assumption is satisfied only by a few kinds of minerals and rocks, because argon is an inert gas that does not readily form bonds with other atoms. In addition, the mineral or rock to be dated must not contain any excess argon-40 atoms that did not actually form by decay of potassium-40 within it. These conditions are satisfied only by the
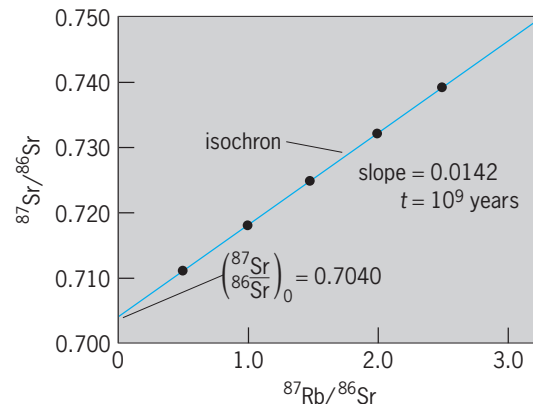

Fig. 1. Isochron diagram used for dating by the Rb-Sr method. The isochron is fitted to the data points representing rocks or minerals. The slope of the isochron yields the age of samples, and its intercept is the initial ^{87}Sr/^{86}Sr ratio. The samples in this hypothetical example are 10^9 years old.

micas, sanidine feldspar, hornblende, and the clay minerals glauconite and illite. In addition, certain fine-grained volcanic rocks of basaltic composition have been dated successfully. Age determinations by this method require measurements of the concentrations of argon-40 and potassium in the sample and are based on Eq. (3), where $\lambda = 5.305 \times 10^{-10}$ per year

$$^{40}\text{Ar} = 0.1102 \ \ ^{40}\text{K}(e^{\lambda t} - 1) \qquad (3)$$

is the decay constant of potassium-40. The alternate decay of potassium-40 to calcium-40 is not useful for dating because that atomic species of calcium has a very high natural abundance.

Potassium-argon dates of micas in igneous and metamorphic rocks indicate the time elapsed since the mineral cooled to about $300°\text{C}$ ($570°\text{F}$), when radiogenic argon is effectively trapped in the crystal lattice. Therefore K-Ar dates of micas may underestimate the crystallization age of the igneous or metamorphic rocks in which they occur. This property of K-Ar dates has been used to measure the rate of cooling of such rocks whose crystallization age can be measured by the whole-rock Rb-Sr method. Potassium-argon dates of hornblende are commonly older than those of biotite extracted from the same rocks because hornblende retains argon-40 at a higher temperature than biotite. The K-Ar method of dating has been used to establish a chronology of mountain building events in North America beginning about 2.8×10^9 years ago and continuing to the present. In addition, the method has been used to date reversals of the polarity of the Earth's magnetic field during the past 1.3×10^7 years. *See* OROGENY; PALEOMAGNETISM.

Argon-40/argon-39 method. The problem of argon loss at elevated temperatures can be avoided by the so-called argon 40-39 method. The sample to be dated is irradiated with energetic neutrons in a nuclear reactor. Some of the potassium-39 atoms in the sample are thereby converted to argon-39. The sample is subsequently heated in steps, and the ratio of argon-40 to argon-39 is measured as the gas is released. Experience has shown that argon released at low temperatures is deficient in radiogenic argon-40 compared to argon-39, and yields dates that are less than the age of the sample. Gas fractions released at higher temperatures close to the melting point of the mineral originate from the centers of grains and have argon-40 to argon-39 ratios that yield older dates approaching the crystallization age of the sample. The explanation is that radiogenic argon-40 formed in the interior of mineral grains is less likely to be lost than argon-40 formed near grain boundaries and microfractures. The $^{40}\text{Ar}/^{39}\text{Ar}$ method yields reliable results when it is used to date highly purified concentrates of potassium-bearing minerals, including plagioclase, hornblende, and pyroxene.

Uranium, thorium-lead method. All of the atoms of uranium and thorium are radioactive and decay through a series of radioactive daughters to stable atoms of lead (Pb). Minerals that contain both elements can be dated by three separate methods based on the decay of uranium-238 to lead-206, uranium-235 to lead-207, and thorium-232 to lead-208. The three dates agree with each other only when no atoms of uranium, thorium, lead, and of the intermediate daughters have escaped. Only a few minerals satisfy this condition. The most commonly used mineral is zircon (ZrSiO_4), in which atoms of uranium and thorium occur by replacing zirconium. Other minerals that have been used for dating are monazite [$(\text{Ce,La,Dy})\text{PO}_4$], titanite ($\text{CaTiSiO}_5$), apatite [$(\text{CaF})\text{Ca}_4(\text{PO}_4)_3$], uraninite ($\text{UO}_2$), and others that are less common. The necessary measurements are the concentrations of uranium, thorium, and lead and the abundances of the radiogenic lead atoms. As an example, the date based on the decay of uranium-238 to lead-206 is calculated from Eq. (4), where λ_1

$$\frac{^{206}\text{Pb}}{^{204}\text{Pb}} = \left(\frac{^{206}\text{Pb}}{^{204}\text{Pb}}\right)_0 + \frac{^{238}\text{U}}{^{204}\text{Pb}}(e^{\lambda_1 t} - 1) \qquad (4)$$

is the decay constant of uranium-238 and the atomic ratios have the same meaning as in the Rb-Sr method.

The dates obtained from uranium-bearing minerals are frequently discordant; that is, they are not in agreement, because lead atoms may be lost from the dated mineral during episodes of thermal metamorphism or by diffusion. Such discordant dates do not indicate the age of the mineral and must be corrected by means of the concordia diagram (**Fig. 2**). Minerals whose uranium-lead decay schemes have been altered lie on straight lines called discordia that intersect the concordia curve in two points. The upper intersection yields the age of the mineral, whereas the lower intersection may date the metamorphic event. The diagram can be used to treat both gain or loss of uranium and loss of lead. The points below concordia have experienced loss of lead or gain of uranium, whereas those above concordia have lost uranium. In this hypothetical example the samples are 3.8×10^9 years old and have been altered recently. Uranium-thorium concordia diagrams have also been used, but complications may arise from
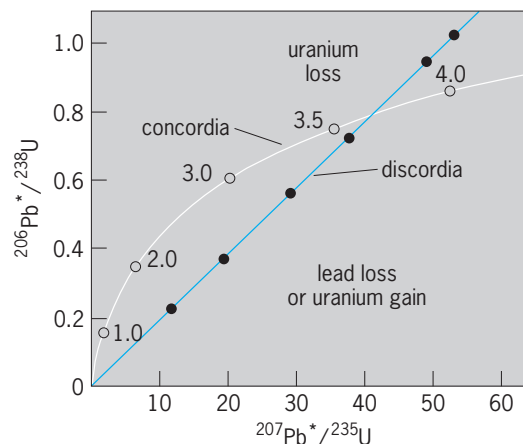

Fig. 2. Concordia diagram used for dating by the U-Pb method. Minerals that yield concordant dates plot on concordia as indicated by the dates expressed in billions of years. Discordant samples form the discordia line which intersects concordia at 3.8×10^9 years and at the origin.

the differences in the geochemical properties of uranium and thorium. When metamorphosed igneous rocks are dated by different methods, the U-Pb concordia dates of zircon usually agree with whole-rock Rb-Sr isochron dates, whereas the K-Ar dates of biotite mica agree with the time of episodic lead loss indicated by the concordia diagram. *See* LEAD ISOTOPES (GEOCHEMISTRY); RADIOACTIVE MINERALS.

Common-lead method. Lead in the common ore mineral galena (PbS) consists of primordial lead that dates from the time of formation of the Earth and varying amounts of radiogenic lead that formed by decay of uranium and thorium in the Earth. The amount of radiogenic lead that was available for mixing with primordial lead increased with time since formation of the Earth. Consequently, young galenas contain more radiogenic lead than old ones. The common-lead method has been used to date ore deposits that cannot readily be dated by other methods. In addition, the theoretical models required for the interpretation of common lead have provided insight into the early history of the solar system and into the relationship between meteorites and the Earth.

Fission-track method. Uranium-238 can decay both by emitting an alpha particle from its nucleus and by spontaneous fission with a decay constant of 8.46×10^{-17} per year. The two fragments of the nucleus leave a trail of damage in the solid through which they move as they dissipate their kinetic energies. The resulting tracks can be enlarged by etching a polished surface of the solid with a suitable reagent. They can then be counted visually by use of a microscope at a magnification of about 1000. The number of spontaneous fission tracks per square centimeter is proportional to the concentration of uranium and to the age of the sample. When the uranium content is known, the age of the sample can be calculated.

This method is suitable for dating a variety of minerals and both natural and manufactured glass. Its range extends from less than 100 years to hundreds of millions of years. Fission track dates are interpreted as cooling ages, because the tracks tend to fade at elevated temperature due to annealing. Minerals such as titanite, zircon, and apatite have different annealing temperatures and therefore yield different dates from which the cooling history of the rock can be determined. The method has also been used for dating archeological objects of glass. *See* FISSION TRACK DATING.

Thorium-230/uranium-238 method. Thorium-230 (^{230}Th) is a radioactive isotope that forms when uranium-238 (^{238}U) decays by emitting alpha particles. Since thorium-230 is an isotope of thorium whose chemical properties differ from those of uranium, the link between uranium-238 and thorium-230 may be broken in the course of geological processes such as partial melting of rocks in the mantle of the Earth or the weathering of uranium-bearing minerals in rocks at the surface of the Earth. When uranium and thorium atoms enter the oceans, uranium tends to be deposited in calcium carbonate skeletons of corals and other organisms, whereas thorium is deposited with sediment particles. Con-

sequently, modern coral reefs contain uranium-238 but lack thorium-230. As time passes, the abundance of thorium-230 in coral reefs increases because of the decay of uranium-238 until parent and daughter reestablish radioactive equilibrium as their decay rates become equal. The time-dependent change in the decay rates of thorium-230 and uranium-238 is expressed by Eq. (5), where the subscript A means

$$\left(\frac{^{230}\text{Th}}{^{238}\text{U}} \right)_{\text{A}} = 1 - e^{-\lambda t} \tag{5}$$

activity and λ is the decay constant of thorium-230 (9.217×10^{-6} per year, half-life: 7.52×10^4 years). This method of dating is useful because the calcium carbonate skeletons of corals have recorded changes in the temperature of ocean water during and after the great Pleistocene ice ages that started about 1 million years ago. *See* ALPHA PARTICLES; RADIOISOTOPE (GEOCHEMISTRY).

Samarium-neodymium method. Both samarium and neodymium belong to the rare earths and occur in low concentrations in many rock-forming minerals. Samarium-147 (^{147}Sm) is radioactive and decays by alpha emission to neodymium-143 (^{143}Nd). The rate of decay is exceedingly slow as indicated by its half-life of 1.07×10^{11} years. The Sm-Nd method of dating separated minerals or whole-rock specimens is similar to the Rb-Sr method [Fig. 1; Eq. (2)]. Specimens of comagmatic igneous rocks or their separated minerals form isochrons whose slope is used to calculate the age of the rocks or minerals. The Sm-Nd method is even more reliable than the Rb-Sr method of dating rocks and minerals, because samarium and neodymium are less mobile than rubidium and strontium.

The isotopic evolution of neodymium in the Earth is described by comparison with stony meteorites. Accordingly, this model can be used to calculate how long ago the neodymium in an igneous, metamorphic, or sedimentary rock separated from the reservoir represented by stony meteorites. *See* METEORITE.

Rhenium-osmium method. Rhenium (Re) occurs primarily in minerals of molybdenum such as molybdenite (MoS$_2$). Osmium (Os) is found as the native metal together with other platinum-group metals. The Re-Os method of dating is based on the beta decay of naturally occurring rhenium-187 to stable osmium-187. It has been used to date iron meteorites and sulfide ore deposits containing molybdenite.

Improvements in analytical techniques permit a wider use of the Re-Os method of dating for studying iron meteorites and identifying meteorite fragments in rock glasses that occur in some large terrestrial craters formed by impacts of meteorites. In addition, the isotope composition of osmium in the rocks of the crust of the Earth differs appreciably from that of young basalt derived from the mantle. Consequently, the isotope composition of osmium in volcanic rocks is a sensitive indicator of assimilation of crustal rocks by basalt magma. *See* EARTH, AGE OF; RADIOCARBON DATING.   Gunter Faure

Bibliography. G. B. Dalrymple and M. A. Lanphere, *Potassium-Argon Dating*, 1972; D. J. De Paolo, *Neodymium Isotope Geochemistry*, 1988; B. R. Doe, *Lead Isotopes*, 1970; G. Faure, *Principles of Isotope Geology*, 2d ed., 1986; D. R. Givens, *Professional Papers in Archiometric Dating*, 1982; R. E. Taylor, *Radiocarbon Dating*, 1987; D. York and R. M. Farquhar, *The Earth's Age and Geochronology*, 1972.

Rock burst

A seismic event that is caused by the mining of underground or surface openings in a high-stress environment. The Mine Safety and Health Administration (MSHA) defines a rock burst as "a sudden and violent failure of a large volume of overstressed rock, resulting in the instantaneous release of large amounts of accumulated energy." Underground, a rock burst sounds like a loud blast. On the surface, a rock burst sounds like a sonic boom, while the ground and structures on the surface shake as if a small earthquake had occurred.

Rock bursts mainly occur in deep underground hard-rock mines. They also occur in some soft-rock mines, such as coal, potash, and trona ($NaHCO_3 \cdot Na_2CO_3$). A rock burst is a very serious hazard to underground mining personnel. Worldwide, there are over 100 fatalities per year from rock bursts. Because of the violent nature of a rock burst, its occurrence often causes significant damage to nearby underground openings. *See* ORE AND MINERAL DEPOSITS; UNDERGROUND MINING.

Classification and types. In mining, rock bursts are categorized as strain, pillar, and fault-slip bursts. Strain bursts are caused by local high stresses at the edge of a mine opening. Generally, the Richter magnitude of these bursts is less than 2.0, and the resulting damage to the opening is less than 100 tons of displaced rock. Pillar bursts occur when the stress or load on a support pillar exceeds its strength (see **illustration**). Pillar bursts can have Richter magnitudes up to 3.5, and damage to the pillar and adjacent openings can be in the hundreds of tons. A fault-slip burst occurs when the shear stress along a

Rock burst damage to a mine opening from a small pillar burst. Note the track and rail heaved up almost 8 ft (2.5 m).

nearby geologic structure is influenced by the overall mine geometry, resulting in a movement along the structure, generally toward the mined-out opening. Richter magnitudes of fault-slip bursts can be very large, up to 5.2, and the resulting damage can be spread over large areas of the mine, involving thousands of tons of displaced rock. *See* MINING; ROCK MECHANICS; SEISMOLOGY; STRESS AND STRAIN.

Causes. As an excavation progresses, the stresses change and the rock deforms around the mine opening. For any type of seismic phenomena to occur, including rock bursting, the stress changes caused by the excavation must result in an unstable equilibrium of the rock mass.

Most hard-rock mining districts throughout the world that experience rock bursts have many similarities. They are in an initially high-stress environment, because of depth, tectonic forces, or both. The shape of the orebody is tabular and the thickness of the vein or width of the seam is usually less than 5 m (16 ft). A very high percentage (over 80%) of the orebody is mined. The wall rocks, and often the vein or seam material, are very hard, brittle, and strong, as well as very old geologically—greater than a billion years old. This combination of high stresses, narrow veins, high extraction ratios, and strong, brittle rocks in a complex geologic environment almost always leads to rock bursting in deep mines.

Dealing with rock bursts. While there is no practical way to eliminate rock bursting, measures have been developed to minimize both the occurrence and effects of bursting. The most usual means of minimizing high stresses around mining openings is by using geometries and excavation sequences that do not create burst-prone pillars and that transfer stress to abutments in a manner that reduces stress concentrations. In practice, this is achieved by using a long-wall-type mining geometry. As mining experience in an orebody or district is gained, the relationship between rock bursting and certain geologic features become evident. Mining through or along such features should be done so that stress increases are minimized or dealt with. Mining rates can be adjusted to allow time for the ground around an advancing opening to settle after blasting. Burst-prone faces are usually mined on a single-shift basis. Most mines in North America have converted to cut-and-fill mining to minimize the geometry change caused by the daily face advance. More recently, mines have converted to long-hole mining to reduce miner exposure at the face and increase productivity.

Standard roof and wall support for an opening will not hold up under the effects of a nearby rock burst. Such reinforcement is designed to support dead loads and not the dynamic shock load resulting from a rock burst. Yielding support, which absorbs energy during deformation, has been found to best contain rock-burst damage. Containment of damage maintains the integrity of the opening and prevents miners from being buried.

In mining, it is not always possible to prevent a rock burst. For some special geometric relationships, it has been possible to avoid the occurrence of

or trigger a burst by destressing. The idea of destressing is to blast/fracture the rock in a high-stress or potential rock-burst zone to reduce the stress. Destress blasting reduces the stress and dissipates the stored strain energy in the rock by inducing fractures that allow yielding.

Most mines that experience rock bursting have installed some type of seismic or microseismic monitoring system. Data from seismic monitoring are used to assess the relative seismicity and potential rock-burst hazard associated with advancing mining faces. Knowledge of a problem area allows rock-burst-control measures to be taken to minimize the occurrence and effects of a burst. *See* SEISMOGRAPHIC INSTRUMENTATION.

During the last 40 years, numerous government-industry cooperative research groups have investigated rock bursts in the United States, Canada, South Africa, Poland, and Australia. This research has resulted in mines being better able to manage their rock-burst problems. However, rock bursts are still the most serious and least understood problem facing deep-mining operations worldwide.

Wilson Blake

Bibliography. W. Blake and D. G. F. Hedley, *Rockbursts: Case Studies from North American Hard-Rock Mines*, 2003; D. G. F. Hedley, *Rockburst Handbook for Ontario Hardrock Mines*, 1992; M. Hudyma, *Mining-Induced Seismicity in Underground Mines: Results of World Wide Survey*, 2004.

Rock cleavage

A secondary, planar structure of deformed rocks. A cleavage is penetrative and systematic, as opposed to fractures and shear zones which may occur alone or in widely spaced sets. It is generally better developed in fine grained rocks than in coarse ones. Application of the term derives from the ability to split rocks along the structure.

Simple cleavages are generally parallel to the axial surfaces of folds. This is true for folds formed by uniform flow of the rock mass where primary layering behaves as a passive marker. When rock layers are buckled, the primary layering behaving as mechanical discontinuities, cleavage that develops early may be fanned by subsequent growth of the fold. The mean orientation of the fanned structures remains a measure of the orientation of the axial surface. Cleavage orientation may change from layer to layer in folds when the rock layers possess contrasting mechanical properties. This phenomenon, known as refraction of cleavage, probably results from local variations in stress orientation because of interlayer shear or from cleavage forming at different times in the various layers during fold growth. The angular relationship between bedding planes and cleavage is an important indicator for the interpretation of fold geometry. In normal situations, cleavage dips more steeply than the bedding planes of the layers in which it occurs. If the cleavage is the more steeply dipping structure, the strata are overturned. In certain completely overturned structures, the structural facing may be determined by examining the younging direction of the strata as observed on the cleavage surface. Cleavage is also associated with faults. *See* FAULT AND FAULT STRUCTURES; FOLD AND FOLD SYSTEMS.

Continuous (microscopically penetrative) cleavages, as in slates, are the earliest tectonic fabric elements that can be recognized in rocks. They may be deformed subsequently by disjunctive (mesoscopically penetrative) structures called crenulation cleavage, a periodic corrugation of the previous foliation. These domainal structures can occur in several generations, reflecting different episodes of deformation.

H. C. Sorby observed the microscopic preferred orientation of phyllosilicate grains in slates. He suggested that platy grains in the shale protolith were rotated into the cleavage plane. A. March showed theoretically that preferred orientation derived in this manner is a function of the strain in the rock. Other early workers reasoned that minerals should grow with their most compliant directions parallel to the maximum compressive stress; phyllosilicate grains would align their basal planes perpendicular to the maximum compression. D. S. Wood has shown that cleavage develops perpendicular to the maximum shortening direction, and that the degree of preferred orientation reflects the strain as predicted by March. It appears that both growth and reorientation may play important parts. As temperature and duration of metamorphism increase, growth processes become dominant, and slaty cleavage is transformed into the coarser foliation of schists. *See* SCHIST; SLATE.

Cleavage also develops in rocks of lower deformational and metamorphic grade than slates. In these rocks, the cleavage occurs as discrete surfaces or seams, often coated with a film of clay or carbonaceous material; the cleavage surfaces are separated by zones of undeformed sedimentary rock. The surfaces may be smooth, anastomosing, or stylolitic. Pressure solution is a major cleavage-forming process in these rocks. Because no macroscopic inhomogeneity exists, these cleavages were thought to result from brittle failure and were termed fracture cleavage. However, it has been determined that compressive stresses cause elastic strains which allow preferential solution that is roughly perpendicular to the direction of the applied stress. *See* METAMORPHISM; ROCK MECHANICS.   David B. Bieler

Bibliography. A. J. Barker, *Metamorphic Textures and Microstructures*, 1990; G. J. Borradaile, M. B. Bayly, and C. McA. Powell, *Atlas of Deformational and Metamorphic Rock Fabrics*, 1982; G. H. Davis, Cleavage, foliation and lineation, *Structural Geology of Rocks and Regions*, 2d ed., 1996; F. J. Turner and L. E. Weiss, *Structural Analysis of Metamorphic Tectonites*, 1963; D. S. Wood, Current views of the development of slaty cleavage, *Annu. Rev. Earth Sci.*, 2:1–35, 1974; D. S. Wood et al., Strain and anisotropy in rocks, *Phil. Trans. Roy. Soc. London*, A283:27–42, 1976.

Rock magnetism

The permanent and induced magnetism of rocks and minerals on scales ranging from the atomic to the global, including applications to magnetic field anomalies and paleomagnetism. Natural compasses, concentrations of magnetite (Fe_3O_4) called lodestones, are one of humankind's oldest devices. W. Gilbert in 1600 discovered that the Earth itself is a giant magnet, and speculated that its magnetism might be due to subterranean lodestone deposits. Observations by B. Brunhes in 1906 that some rocks are magnetized reversely to the present Earth's magnetic field, and by M. Matuyama in 1929 that reversely and normally magnetized rocks correspond to different geological time periods, made it clear that geomagnetism is dynamic, with frequent reversals of north and south poles. Nevertheless, permanent magnetism of rocks remains important because it alone provides a memory of the intensity, direction, and polarity of the Earth's magnetic field in the geological past. From this magnetic record comes much of the evidence for continental drift, sea-floor spreading, and plate tectonics. *See* CONTINENTAL DRIFT; GEOMAGNETISM; MAGNET; MAGNETISM; PALEOMAGNETISM; PLATE TECTONICS.

Origin of rock magnetism. The magnetism of rocks arises from the ferromagnetism or ferrimagnetism of a few percent or less of minerals such as magnetite. The magnetic moments of neighboring atoms in such minerals are coupled parallel or antiparallel, creating a spontaneous magnetization M_s. All magnetic memory, including that of computers, permanent magnets, and rocks, is due to the spontaneous and permanent nature of this magnetism. Spontaneous magnetization requires no magnetic field to create it, and cannot be demagnetized.

The magnetism can be randomized on the scale of magnetic mineral grains because different regions of a crystal tend to have their M_s vectors in different directions. These regions are called magnetic domains. Grains so small that they contain only one domain (single-domain grains) are the most powerful and stable paleomagnetic recorders. Larger, multidomain grains can also preserve a paleomagnetic memory, through imbalance in the numbers, sizes, or directions of domains, but this memory is more easily altered by time and changing geological conditions. *See* FERRIMAGNETISM; FERROMAGNETISM; MAGNETISM; MAGNETITE.

Magnetization processes. By applying a large enough magnetic field, domains can be enlarged or reoriented. When the magnetic field is reduced to zero, the original domain sizes and directions are usually not recovered. This phenomenon is called hysteresis (**Fig. 1**). It leaves the entire rock sample with a permanent magnetism, called remanent magnetization M_r. Although M_r is much smaller than M_s of an individual domain, it is large enough to be measured easily with modern magnetometers. *See* MAGNETOMETER.

The permanent magnetism of rocks does not normally respond to a magnetic field as small as that of the Earth, except for a small component called viscous remanent magnetization (VRM) which is acquired slowly over very long times. Instead, the main part of an igneous rock's magnetic memory is acquired as a result of temperature changes, and is called thermoremanent magnetization (TRM). As a rock is heated, its hysteresis loop shrinks (Fig. 1). The magnetization decreases and eventually disappears at the Curie temperature. At the same time, the loop shrinks along the field axis, until at sufficiently high temperatures even a field as small as the Earth's can redirect the domains and create a new M_r. By cooling from this temperature, called the blocking temperature, M_r is stabilized against further field changes. *See* CURIE TEMPERATURE; MAGNETIZATION.

Laws of partial TRM. The thermoremanent magnetization creation process was demonstrated by J. Koenigsberger, E. Thellier, and T. Nagata in the 1930s and 1940s. Thellier in particular described some
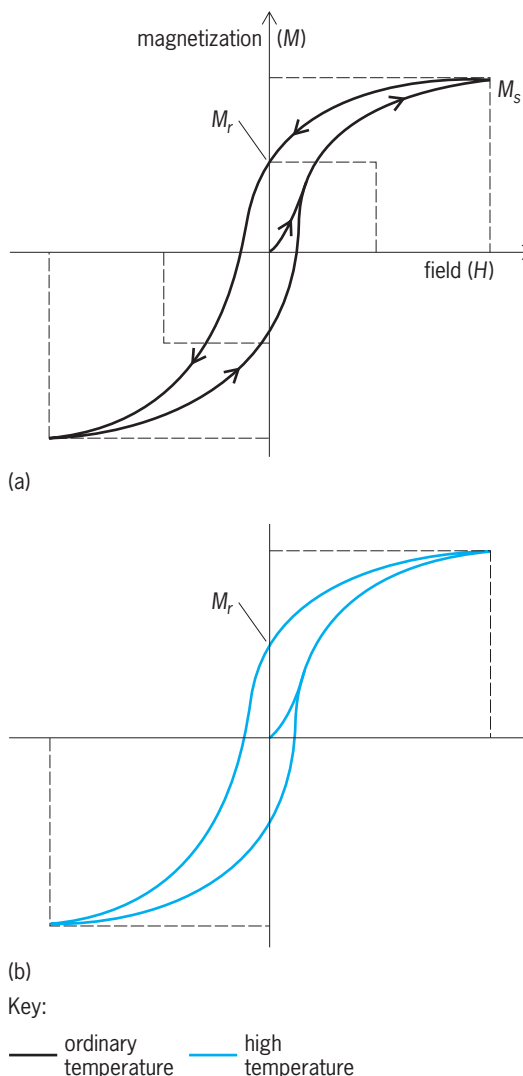

(a)

(b)

Key:

———— ordinary temperature   ———— high temperature

Fig. 1. Magnetic hysteresis curves at ordinary temperature and high temperature approaching the Curie point. M_r = remanent magnetization; M_s = saturation magnetization. The *M-H* curve contracts along both the *M* and *H* axes with heating.
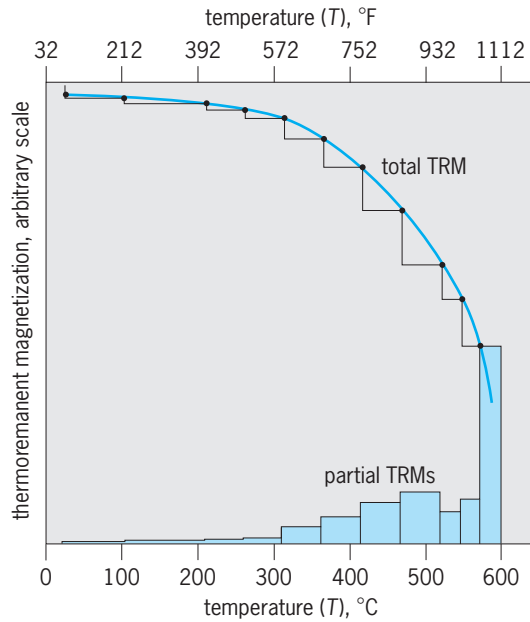
Fig. 2. Illustration of the Thellier laws. The partial thermoremanent magnetizations sum up to give the total thermoremanent magnetization. When the total thermoremanent magnetization is demagnetized by heating in steps, the decrements match the partial thermoremanent magnetizations over corresponding temperature intervals.

remarkable experimental properties of thermoremanent magnetization. The magnetic field may change during cooling. For example, the geomagnetic field may reverse during slow cooling of a rock. Thellier discovered that each blocking temperature interval over which the field is constant, has associated with it a partial thermoremanent magnetization that is independent of all other partial thermoremanent magnetizations created in other intervals. The partial thermoremanent magnetizations can be measured, then erased by heating in a zero field to randomize the domains, without affecting one another. They sum together to give the total thermoremanent magnetization created by cooling from the Curie temperature in a single field (**Fig. 2**).

These observations, called the Thellier laws, are the basis of the paleomagnetic method. They allow paleomagnetists to unravel the often complicated magnetic history of a rock that may have cooled through a series of field reversals, or have been reheated in nature at different times, by stripping away the successive generations of partial thermoremanent magnetizations and isolating the M_r vector of each one. By observing how the natural thermoremanent magnetization of the rock (its natural remanent magnetization, or NRM) compares with the partial TRMs imparted in laboratory experiments, paleomagnetists can deduce the intensity of ancient magnetic fields as well as their vector directions.

The elegant simplicity of Thellier's laws were explained by L. Néel in 1949, the year after his Nobel prize-winning work on ferrimagnetism and anti-

ferromagnetism. Néel showed that a single-domain grain has one and only one blocking temperature. Each partial thermoremanent magnetization is associated with a separate population of single-domain grains in a rock. The independence and additivity of partial thermoremanent magnetizations is made clear by this simple picture. However, the thermoremanent magnetization of multidomain grains, which also approximately obey the Thellier laws, is not so easily understood because the domains in any grain can contribute to partial thermoremanent magnetizations in several different blocking temperature intervals. *See* ANTIFERROMAGNETISM.

Induced magnetization and magnetic anomalies. Although rock magnetism emphasizes remanent or permanent magnetization, the temporary magnetization induced by the Earth's field is also of importance. For most types of rocks in the Earth's crust, induced magnetization outweighs natural remanent magnetization. This is sometimes described by a Koenigsberger ratio of remanent to induced magnetization, Q, of less than 1. Another reason for the importance of induced magnetization is that it remains constant or increases with heating, whereas M_r decreases with heating (Fig. 1). Rocks in the deep crust, where temperatures approach the Curie point, have smaller Q values, that is, more important induced magnetizations, than similar rocks at the surface. *See* GEOPHYSICAL EXPLORATION.

The large magnetic anomalies recently discovered by the *Mars Global Surveyor* satellite cannot be due to induced magnetization because the present global magnetic field of Mars is very small compared to that of the Earth. Remanent magnetization must be responsible. The magnetic mineral or minerals involved have not yet been identified.

Magnetic minerals. Five groups of magnetic minerals are significant sources of induced and remanent magnetization. On the Moon and in most meteorites, metallic iron is most important. On Earth, titanomagnetites ($Fe_{3-x}Ti_xO_4$) or their oxidized equivalents, titanomaghemites, are usually dominant. Fresh basaltic lavas erupting at mid-ocean ridges, the source of new sea floor, contain titanomagnetites with about 60% titanium (Ti). Older sea floor and continental rocks tend to contain almost titanium-free magnetite (Fe_3O_4) or its oxidized equivalent, maghemite (γFe_2O_3). Titanohematites ($Fe_{2-y}Ti_yO_3$) occur as primary minerals in some lavas and plutonic rocks. Almost titanium-free hematite (αFe_2O_3) is a common oxidation product in rocks, sediments, and soils, as are iron oxyhydroxides such as goethite (FeOOH), which ultimately dewaters to hematite. Both hematite and goethite are weakly magnetic antiferromagnets, but are significant sources of natural remanent magnetization because they occur in such large quantities in weathered materials. The iron sulfides pyrrhotite (Fe_7S_8) and greigite (Fe_3S_4) are fairly common in terrestrial igneous and sedimentary rocks. Troilite (FeS) occurs in lunar rocks and meteorites. *See* GOETHITE; HEMATITE; IGNEOUS ROCKS; METEORITE; PYRRHOTITE; SEDIMENTARY ROCKS.

Not all magnetic minerals are of inorganic origin. Magnetotactic bacteria produce chains of single-domain magnetite or greigite crystals, which they use as compasses. These crystals become magnetic fossils after the bacteria die. When deposited in soils and sediments, the crystals are an important source of stable natural remanent magnetization of an origin quite different from thermoremanent magnetization. Many other organisms, among them bees, birds, and fish, produce magnetic crystals, but their navigational or other functions are not always clear. *See* BIOMAGNETISM.

Sea-floor spreading and linear magnetic anomalies. When lavas erupt at mid-ocean ridges, the titanomagnetites they contain cool from above the Curie temperature. The geomagnetic field produces thermoremanent magnetization during this cooling (**Fig. 3**). Periodically, at intervals of about 1 million years, the geomagnetic field reverses. The thermoremanent magnetization acquired by lavas erupted after such a reversal is opposite in direction to thermoremanent magnetization acquired before the reversal. Because the sea floor is slowly spreading away from mid-ocean ridges, older lavas are found farther from ridges than are recently erupted lavas. For this reason, the spreading sea floor acts like a tape recorder, with successive bands of oppositely directed thermoremanent magnetization recording successive field reversals (Fig. 3). Mid-ocean ridges are hundreds or thousands of kilometers long while the bands are typically only tens of kilometers wide, because spreading rates are only a few centimeters per year. The bands therefore form "magnetic stripes" paralleling the ridge axis on either side.

This picture of sea-floor magnetism was proposed by F. Vine and D. Matthews in 1963. Their main evidence was the widespread linear pattern of magnetic field anomalies measured over the oceans, which is quite different from the irregular shapes of anomalies over the continents. The linear magnetic anomalies arise from fields produced by the magnetic stripes on the sea floor, which alternately add to or subtract from the main geomagnetic field, producing measurable anomalies. This picture convinced most geologists of the reality of sea-floor spreading and ushered in the era of plate tectonics. *See* MARINE GEOLOGY; MID-OCEANIC RIDGE.

Paleomagnetic evidence for continental drift and plate tectonics. Once geologists were convinced that thermoremanent magnetization of sea-floor basalts faithfully records sea-floor spreading, and therefore represents a trustworthy magnetic memory of the geomagnetic field millions of years earlier when the basalts formed, they looked more favorably on paleomagnetic measurements of the natural remanent magnetization of continental rocks. This natural remanent magnetization is much more varied than that of the sea floor. It includes thermoremanent magnetization of igneous rocks and depositional remanent magnetization (DRM) of sedimentary rocks produced by partial alignment of magnetized grains
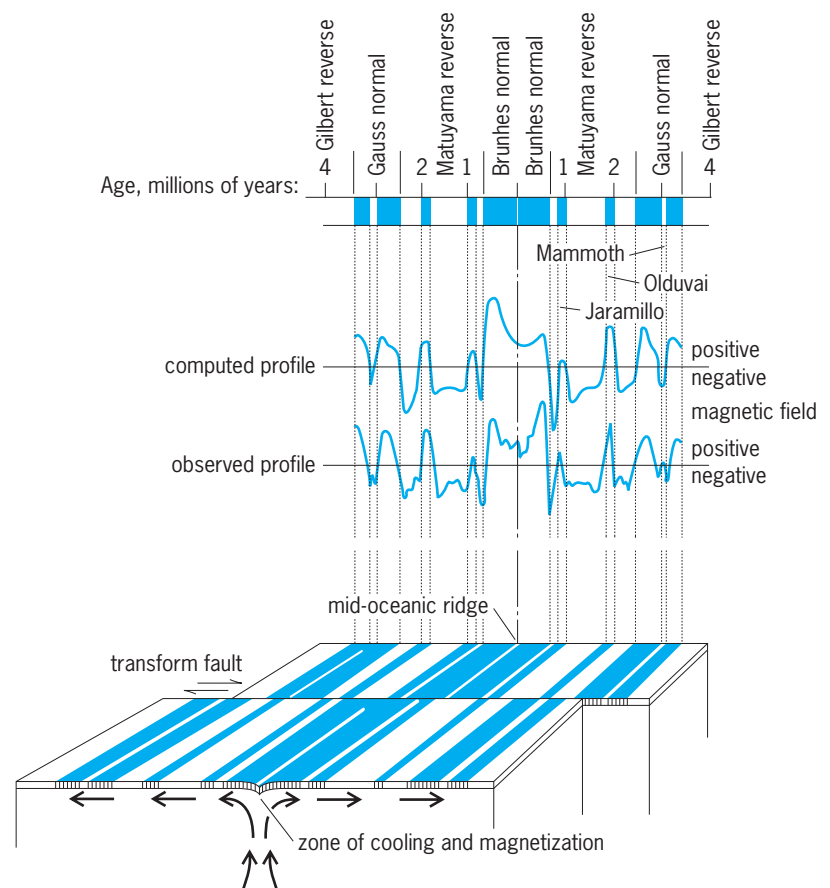

Fig. 3. Vine and Matthews' model of linear magnetic anomalies. Lavas cooling at a mid-ocean ridge acquire thermoremanent magnetization. The combination of spreading away from the ridge and periodic reversals of the geomagnetic field imprints magnetic stripes, alternately normal and reverse, on the sea floor. At the sea surface, the magnetic stripes create alternating positive and negative magnetic field anomalies, shown in the profiles.

in wet sediments. These are primary natural remanent magnetizations dating from the formation of the rock. Rocks of all types can also acquire secondary natural remanent magnetizations. Partial thermoremanent magnetization due to later burial and reheating and chemical remanent magnetization (CRM) resulting from changes in the magnetic minerals are the most common types. Metamorphic rocks which have experienced a turbulent tectonic history may have several generations of partial thermoremanent magnetization and chemical remanent magnetization. *See* METAMORPHIC ROCKS.

Part of the earlier skepticism about paleomagnetic data arose from the complexity of natural remanent magnetization in continental rocks. The paleomagnetist's task is to separate and successively erase different layers of natural remanent magnetization, associating an age with each. The usual methods for such "cleaning" are demagnetization, in steps, of the natural remanent magnetization by either heating or alternating magnetic fields in the laboratory. *See* DEMAGNETIZATION.

The methodology is different from that used for measuring sea-floor spreading recorded by magnetic
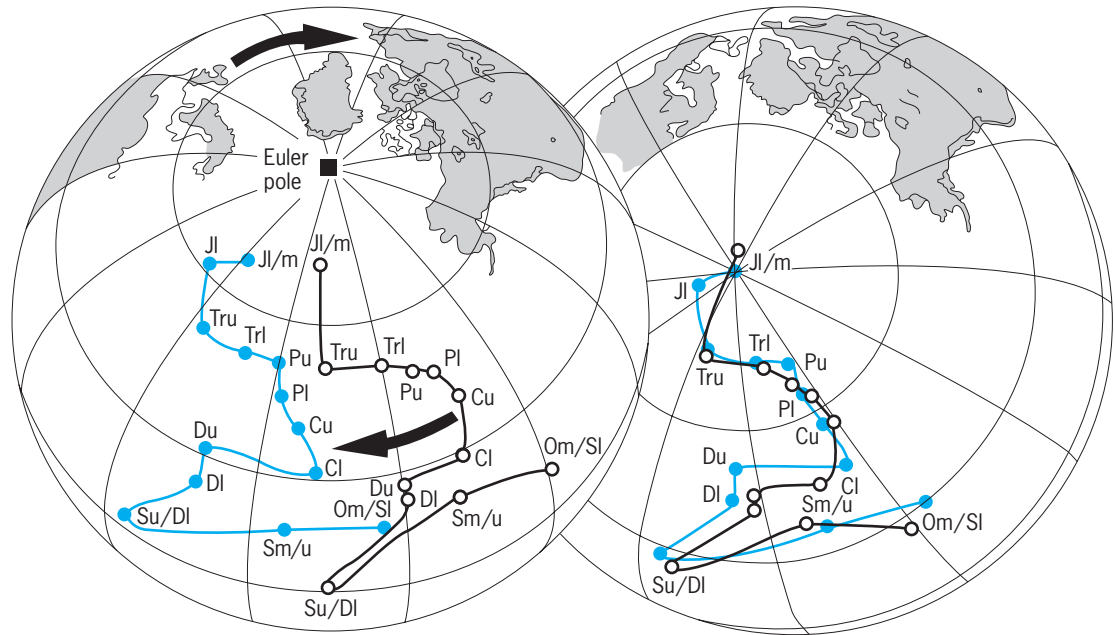
Fig. 4. Tracks of the positions of paleomagnetic poles for European and North American rocks with these continents in their present positions (left) and after closing the Atlantic Ocean (right). The polar tracks diverge before closure but become coincident with the continents in their Pangea configuration. The symbols along the tracks refer to geological time periods: J, Jurassic; Tr, Triassic; P, Permian; C, Carboniferous; D, Devonian; S, Silurian; O, Ordovician; u, upper; m, middle; l, lower. (*After R. Van der Voo, Phanerozoic paleomagnetic poles from Europe and North America and comparisons with continental reconstructions, Rev. Geophys., 28:167–206, 1990*)

field anomalies. Instead of remote sensing of a magnetic signal, in paleomagnetism the magnetized rocks themselves are sampled and their magnetic moments measured. Each sample acts as a fossil compass, recording both the direction and the intensity of the geomagnetic field. Except for very recent rocks, the natural remanent magnetization is generally not parallel to the present geomagnetic field vector. Either the magnetic poles of the Earth have moved (polar wander) or the continents have moved relative to fixed poles (continental drift). Rocks from a single continent define a consistent magnetic pole track with increasing age, but different continents have different pole tracks. However, when the continents are fitted together into a Pangea supercontinent, as suggested by A. Wegener in 1912, large sections of the pole tracks for different continents come together as well (**Fig. 4**).

Continental drift and sea-floor spreading are different manifestations of plate tectonics. Each continent and its adjacent sea floor form a lithospheric plate, which moves rigidly as a unit over the Earth's surface. The plate boundaries do not necessarily correspond to continental margins. Instead they form a continuous linkage of mid-ocean ridges, subduction zones, and transform faults. *See* CONTINENTS, EVOLUTION OF; LITHOSPHERE; SUBDUCTION ZONES; TRANSFORM FAULT.

Magnetic field reversals and magnetostratigraphy. Reversals of the Earth's magnetic field are synchronous worldwide. This fact was first shown by A. Cox, R. Doell, and B. Dalrymple in the 1960s by dating lava flows from different parts of the world

recording the same sequences of reversals. Reversals therefore are time markers in the geological record and can be used for correlation of rocks and sediments locally, regionally, or globally. This branch of paleomagnetism is called magnetostratigraphy. Long sections of ocean sediments cored by drilling ships and similar sections of older sedimentary rocks now exposed on the continents have yielded sequences of normal and reversed natural remanent magnetizations that match the magnetic stripe sequence from linear magnetic anomalies (**Fig. 5**). Many important geological boundaries, including the Cretaceous-Tertiary boundary around 65 million years ago, at which the dinosaurs became extinct, and the Brunhes-Matuyama boundary (the most recent geomagnetic reversal, 780,000 years ago), an important time marker in human evolution, are now much more accurately dated geomagnetically than they were previously from fossil records. *See* CRETACEOUS; GEOLOGIC TIME SCALE; TERTIARY.

Magnetic records from lake sediments, soils, and loess. When sedimentary deposition is rapid, as in lakes, soil formation, and windblown loess deposits, it is possible to measure the fine-scale variation of geomagnetic field direction and intensity, called paleosecular variation. In some rapidly deposited marine sediments, it is also possible to measure magnetic field variations between and even during field reversals. The geomagnetic signal is often modulated by climatic influences such as paleotemperature, rainfall variation, and paleowind direction. Therefore, paleomagnetic records from sediments and soils are

proving to be valuable archives of information about global climate changes in the past. *See* CLIMATE HISTORY; DEPOSITIONAL SYSTEMS AND ENVIRONMENTS; LOESS.

Paleointensity variation. Although depositional remanent magnetization of sediments gives some indication of the past variation of the Earth's magnetic field intensity, precise records rely on the thermoremanent magnetization of rocks or baked archeological materials such as pottery, bricks, and kilns. Such records are a method of archeological and geological dating. Another important purpose of paleointensity studies is to take the pulse of the geodynamo in the Earth's core, which is believed to be the source of most of the geomagnetic field. The changing intensity of the field before, during, and after reversals helps advance the understanding of how the geodynamo works. *See* DATING METHODS; GEODYNAMO; MAGNETOHYDRODYNAMICS.

Paleomagnetism of other planets. The natural remanent magnetizations of meteorites bring to Earth a record of magnetic fields from other times and places in the solar system. Direct sampling of rocks on the Moon during the *Apollo* scacecraft missions of the 1970s demonstrated that many carry very ancient natural remanent magnetizations that record a declining lunar magnetic field strength over time (**Fig. 6***a*). On the basis of magnetic field anomalies measured by the *Mars Global Surveyor* satellite, portions of the crust of Mars are intensely magnetized, about ten times more strongly than on Earth, with a lineated pattern reminiscent of magnetic stripes in the Earth's oceans. As with the Moon, the present global field of Mars is very small. Only the Earth, among the terrestrial planets, has preserved a vigorous geodynamo as evidenced by its paleointensity

(a)

(b)

Fig. 6. Paleointensities recorded by meteorites, lunar samples, and terrestrial rocks throughout geological history. The Earth's magnetic field has fluctuated with time, but its present strength is comparable to the range of values 2.5–3 billion years ago. The Moon's global field has decreased with time and is negligible today. For comparison between the two graphs, the present field of the Earth is about 30×10^6 tesla near the Equator and about 60×10^{-6} tesla near the Poles, similar to average meteorite and early lunar values.

Fig. 5. Sequence of reversals of the geomagnetic field in late Cenozoic time.

record covering practically all of geological history (Fig. 6*b*).   David J. Dunlop

Bibliography. R. F. Butler, *Paleomagnetism*, Blackwell, 1992; J. E. P. Connerney et al., Magnetic lineations in the ancient crust of Mars, *Science*, 284:794–798, 1999; A. Cox, G. B. Dalrymple, and R. R. Doell, Reversals of the Earth's magnetic field, *Sci. Amer.*, 216:44–54, 1967; D. J. Dunlop and Ö. Özdemir, *Rock Magnetism: Fundamentals and Frontiers*, Cambridge University Press, 1997; W. Gilbert, *De Magnete*, Dover Reprints, 1958; B. A. Maher and R. Thompson (eds.), *Quaternary Climates, Environments and Magnetism*, Cambridge University Press, 2000; R. T. Merrill, M. W. McElhinny, and P. L. McFadden, *The Magnetic Field of the Earth*, Academic Press, 1996; L. Néel, Some theoretical aspects of rock magnetism, *Adv. Phys.*, 4:191–243, 1955; N. D. Opdyke and J. E. T. Channell, *Magnetic Stratigraphy*, Academic Press, 1996.

Rock mechanics

The scientific discipline that deals with understanding how rocks deform and fail due to natural and human-induced forces at scales ranging from micrometers to kilometers. Rock mechanics is also the engineering discipline for designing and stabilizing underground and surface excavations and rock foundations.

The deformation and failure of rocks influence our lives in many ways. Structures that are built with or into rocks include highways, dams, bridges, tunnels, mines, and water and petroleum wells, to name some. These structures must be designed to remain stable over their expected lifespan. Natural occurrences of rock failure include earthquakes, volcanic eruptions, and landslides. These geologic hazards must be understood and methods for predicting rock failure events must be developed. *See* ENGINEERING GEOLOGY; GEODYNAMICS.

Intact rock, discontinuities, and rock mass. Rock, such as granite, sandstone, or gneiss, is defined as an aggregate of minerals. Small samples of rock that are tested in the laboratory are referred to as intact rock, in contrast to the larger volumes of rock that accommodate highways, tunnels, and other engineering structures. These larger volumes of rock are referred to as rock masses and contain discontinuities as well as intact rock. Discontinuities in rock include joints, faults, bedding planes, and other types of large-scale rock fractures. A typical rock mass is shown in **Fig. 1**. Joints are the most common type of discontinuity and are found in almost all rocks. Joints usually occur in sets defined by their common orientation. Several joint sets can be seen in Fig. 1 (one set sloping to the right and one set sloping to the left). A stereonet is often used to display joint ori-

Fig. 1. Typical rock mass containing numerous discontinuities. The person in picture is conducting a scanline survey, where a tape is placed across the rock and detailed information is gathered on the fractures that intersect the scanline.

entation information and assist in identifying joint sets. *See* JOINT (GEOLOGY); PETROFABRIC ANALYSIS; PETROLOGY; ROCK.

Rock-mass characterization and classification. Discontinuities cause the strength of a rock mass to be much weaker than the intact rock. Before an engineering structure is made in a rock mass, it is very important to characterize the properties of the discontinuities, a process called rock-mass characterization. Rock-mass characterization involves the careful measurement of discontinuity type, orientation, spacing, length, roughness, fill, and other properties. This information can be collected in the field from existing rock outcrops or from borehole drill core. Figure 1 shows one method of collecting rock-mass information in the field, referred to as a scanline survey. *See* DRILLING, GEOTECHNICAL.

Rock-mass classification is another aspect of rock-mass characterization. Rock-mass classification uses subsets of the rock characterization data to determine numerical or descriptive ratings. These ratings can then be used to estimate underground support requirements, rock-mass strength and deformation, and other engineering design criteria. Rock-mass classification schemes represent a convenient way to assimilate the rock characterization data into a single number or distribution of numbers. The rock quality designation (RQD) considers one type of information: the total percent of a rock core containing pieces of core greater than 10 cm (4 in.). It gives a rating from 0 to 100 (worst to best). The rock mass rating (RMR) system considers 10 types of information: intact rock strength, rock quality designation, joint spacing, joint length, joint roughness, joint fill, joint weathering, joint aperture, water, and joint orientation. Individual ratings are determined for each information type. The rock-mass rating, which is the sum of these individual ratings, is a number that ranges from 0 to 100 (worst to best). Other popular rock-mass classification schemes used today include the mining rock-mass rating (MRMR), the slope mass rating (SMR), the geologic strength index (GSI), the rock quality index (Q), and the rock-mass index (RMi).

Stress and strain. Rocks deform and fail due to forces, including gravity, tectonic, seismic, water, and thermal. Force is not a useful measure of rock strength because it depends on the size of the rock specimen subjected to the force; that is, it takes a small force to break a small rock sample and a larger force to break a larger sample of the same rock type. Instead, we use stress as a measure of rock strength. Rock failure will occur at a point in a rock when the stress at that point exceeds the strength of the rock. Stress σ has the unit of force per unit area. In SI units the stress unit is pascals (newtons/m^2) and in English units the stress unit is psi (pounds/in.^{2}).

Strain is a measure of the deformation of a rock. Strain is defined as the change in length divided by the original length, and is unitless. The symbol ε is normally used for strain. Under conditions in the shallow crust, rock is brittle and cannot strain very much before failure. A strain of 0.05 (0.5%) is usually

sufficient to cause most rocks to fail. *See* STRESS AND
STRAIN.

Strength of intact rock. The strength of a small
sample of rock can be determined from labora-
tory tests. Many tests usually are done to deter-
mined the strength of a particular rock type, because
the strength of rock in tension is different than its
strength in compression and because the strength
varies from one location to another (heterogeneity).
Common laboratory tests to determine the strength
of rock include the uniaxial and triaxial compression
tests, the Brazilian disk tension test, and the point
load test. **Figure 2** shows a point load test.

In the uniaxial compression test, a force is applied
in only one direction, as shown in **Fig. 3***a*. The rock
strength under this type of loading is referred to as
the uniaxial compressive strength (UCS). The uniax-
ial compressive strength is the most common index
property for intact rock strength. **Table 1** shows
the uniaxial compressive strength for different rock
types and UCS ranges for the categories of very weak,
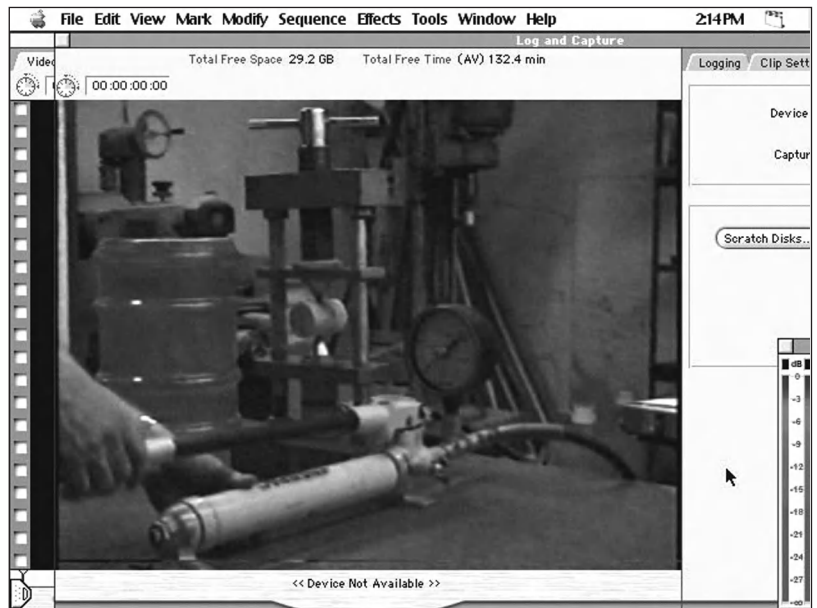weak, medium, strong, and very strong rock.

Fig. 2. Point load test on a core sample of rock. The results of the point load test can be
used to estimate the uniaxial compressive strength (UCS) of the intact rock.

TABLE 1. Classification of rock hardness

Strength classification	UCS (MPa)	Typical rock types
Very weak	10–20	weathered and weakly compacted sedimentary rocks
Weak	20–40	weakly cemented sedimentary rocks, schist
Medium	40–80	competent sedimentary rocks; some low-density coarse-grained igneous rocks
Strong	80–160	competent igneous rocks; some metamorphic rocks and fine-grained sandstones
Very strong	160–320	quartzites; dense fine-grained igneous rocks

Adapted from P. B. Attewell and I. W. Farmer, *Principles Of Engineer-
ing Geology*, Chapman and Hall, London, 1976.

The uniaxial compression test does not adequately
represent the actual stresses in the ground. In an ac-
tual rock mass, the stress state is a result of forces in
more than one direction. The triaxial compression
test is a better representation for the actual stresses
in the ground. In the triaxial test, a confining stress
is applied in addition to the axial stress as shown in
Fig. 3*b*. During the test, the confining stress is held at
a constant value and then the axial stress is increased
until the sample fails. Rock strength increases with
increasing confining stress and the friction angle ϕ
is the rock parameter that describes the increase in
strength with increasing confining stress. A typical
friction angle for a strong rock is 40–60° but can be as
low as 20° for weak rock. With the two strength pa-
rameters UCS and friction angle, the Mohr-Coulomb
criterion predicts a linear relationship between the
triaxial compressive strength (σ_1) and the confining
stress (σ_3), as shown in Eq. (1).

$$\sigma_1 = \text{UCS} + \sigma_3 \tan^2(45 + \phi/2) \qquad (1)$$

Strength of discontinuities. Discontinuities repre-
sent preexisting weakness planes in the rock. Slid-
ing will occur along discontinuities when the shear
stress τ along the discontinuity plane exceeds the
shear strength of the discontinuity. The strength of a
discontinuity is frictional and depends on the nor-
mal stress acting perpendicular to the plane, σ_n.
The Coulomb criterion predicts a linear relationship
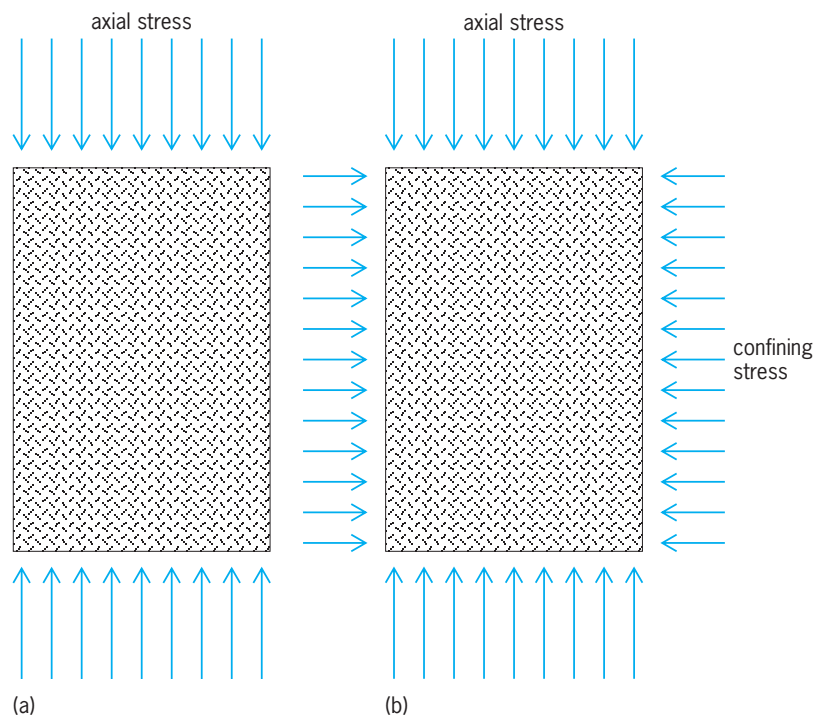between the shear strength of a discontinuity (τ)

Fig. 3. (*a*) Uniaxial compression test, where an axial stress is applied to a core sample of
rock and increased until failure. (*b*) Triaxial compression test, where a constant confining
stress is first applied, then the axial stress is increased until failure.

and the normal stress (σ_n) as shown in Eq. (2).

$$\tau = \sigma_n \tan \phi_j \qquad (2)$$

Note that the friction angle for a discontinuity is not the same as that for the intact rock, and thus the symbol ϕ_j is used. A typical friction angle for a smooth unweathered discontinuity is $30°$. Faults often have frictional angles as low as $10°$, which can result in stability problems encountered in excavations or foundations. The roughness of the discontinuity can significantly increase the friction angle for a discontinuity to values above $50°$. However, under very high normal stresses, the roughness does not contribute as much to the friction angle since the asperities that make up the roughness can be sheared off during sliding.

Strength of rock mass. Both the strength of the intact rock and the strength of the discontinuities influence the strength of a rock mass. If only one or two sets of joints occur in a rock mass, the strength of the rock mass may be anisotropic, which means that the strength is orientation dependent. Shale is an example of a commonly encountered rock that has anisotropic strength properties. If three or more joint sets occur in a rock mass, then it is usually assumed that the rock-mass strength is isotropic (same strength in all directions). The strength of a rock mass will depend on all the information that was collected during rock-mass characterization, such as the number of joint sets, joint spacing, joint roughness, joint fill, and other properties. Because of this, empirical relationships for rock-mass strength often use the rock-mass classification directly. For instance, the 1992 version of the popular Hoek and Brown rock-mass strength criterion uses the rock mass rating (RMR), as given in Eq. (3), where m are parameters that depend on the rock type and condition.

$$\frac{\sigma_1}{\text{UCS}} = \frac{\sigma_3}{\text{UCS}} + \sqrt{m\frac{\sigma_3}{\text{UCS}} + s} \qquad (3)$$

Because the strength of a rock mass cannot be measured directly, empirical methods for estimating rock-mass strength are widely used today. With the advent of sophisticated numerical methods, it is now possible to estimate rock-mass strength using numerical modeling.

In-situ stress. To determine how a rock mass will behave in an excavation or foundation, it is necessary to estimate the stresses in the ground prior to engineering works, referred to as in-situ stresses. In general, the stresses in the ground increase with depth below the surface, and the vertical stress can be estimated by the weight of the material above a location of interest. Most rocks have a density of about .025 MN/m³ (which corresponds to 0.025 MPa per m depth), and therefore at a depth of 1000 m (3300 ft), the vertical stress will be about 25 MPa. The horizontal stresses can be less than or greater than the vertical stress at a particular depth. Normal faulting tectonic regimes are associated with stress states where the horizontal stress is less than the vertical stress,

and thrust faulting regimes are associated with stress states where the horizontal stress is greater than the vertical stress. In addition, high horizontal stresses can occur in regions where a large amount of geologic removal of rock has occurred. There are many techniques available for measuring the in-situ stress state, including overcoring, hydraulic fracturing, and borehole breakouts.

Induced stresses around underground excavations. An excavation made in a rock mass will perturb the in-situ stress state. It will cause a concentration of stress at the boundaries of the excavation, particularly in the case of an underground excavation [more precisely, referring to Eq. (1) or (3), it results in an increase in σ_1 and a decrease in σ_3 around the boundary]. These induced stresses may result in rock failure. Failure around an underground excavation usually starts at the boundary and progresses into the rock mass, as shown in **Fig. 4**a. In some cases, the
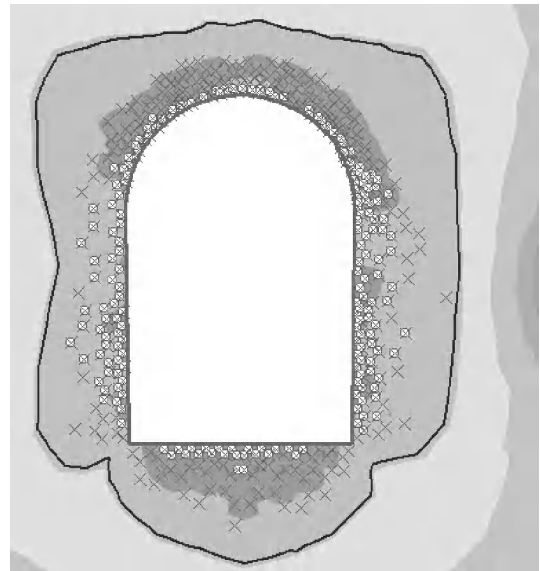
(a)

(b)

Fig. 4. (a) Example of the failure zone around an underground excavation (×, shear failure; o, tensile failure), calculated using the Rocscience Phase2 computer program. (b) Example of a shallow underground excavation where the failure mode is either gravity-induced block falls or block sliding along discontinuities.

TABLE 2. Recommendations for rock-support and excavation sequence based on the rock-mass rating (RMR) for the rock mass

Rock-mass class	Excavation	Rock bolts (20 mm diameter, fully grouted)	Shotcrete	Steel sets
I. Very good rock RMR: 81–00	Full face, 3 m advance	Generally no support required except spot bolting		
II. Good rock RMR: 61–80	Full face, 1–1.5 m advance; complete support 20 m from face	Locally, bolts in crown 3 m long, spaced 2.5 m with occasional wire mesh	50 mm in crown where required	None
III. Fair rock RMR: 41–60	Top heading and bench 1.5–3 m advance in top heading; commence support after each blast; complete support 10 m from face.	Systematic bolts 4 m long, spread 1.5–2 m in crown and walls with wire mesh in crown	50–100 mm in crown and 30 mm in sides	None
IV. Poor rock RMR: 21–40	Top heading and bench 1.0–1.5 m advance in top heading; install support concurrently with excavation, 10 m from face.	Systematic bolts 4–5 m long, spaced 1–1.5 m in crown and walls with wire mesh	100–150 mm in crown and 100 mm in sides.	Light to medium ribs spaced 1.5 m where required
V. Very poor rock RMR: < 20	Multiple drifts 0.5–1.5 m advance in top heading; install support concurrently with excavation; Shortcreate as soon as possible after blasting	Systematic bolts 5–6 m long, spaced 1–1.5 m in crown and walls with wire mesh; bolt invert	150–200 mm in crown, 150 mm in sides, and 50 mm on face	Medium to heavy ribs spaced 0.75 m with steel lagging and forepoling if required; close invert

Adapted from http://www.rocscience.com/hoek/PracticalRockEngineering.asp.

induced stresses may increase the stability of the excavation. This is often the case in relatively competent rock mass in a shallow excavation where the dominant mode of failure is gravity, causing blocks of rock to fall from the roof or slide from the walls, as shown in Fig. *4b*. In this case, the induced stresses squeeze the rock blocks together and prevent them from sliding [more precisely, referring to Eq. (2), it results in an increase in J_b on the discontinuities]. The induced stresses around the opening depend on the in-situ stresses and the shape of the excavation. Sharp corners have particularly high stress concentrations, as do tunnel intersections. It is often possible to increase the stability of an underground excavation by changing its size, shape, or orientation. A technique used in weak rocks is to excavate multiple drifts rather than a single "full face" excavation to minimize induced stresses and to allow rock support to be installed between excavations. For simple geometries, closed- form solutions are available for the induced stresses. One of the most-used formulas in rock mechanics is the Kirsch solution for the induced stresses around a circular underground opening. For more complicated geometries (for example, the three-dimensional geometry formed by two intersecting tunnels), numerical methods are used to estimate induced stresses and subsequent rock failure.

Supporting unstable rock masses. In many circumstances, a rock engineering design will be unstable without artificial rock support. Rock support technologies include rock bolts, cable bolts, shotcrete, concrete, steel sets, wire mesh, and other methods. Rock bolts are the most popular method for underground stabilization. They consist of steel or aluminum rods 1–6 m (3–20 ft) in length inserted into predrilled holes, with bolt spacing ranging 0.5–3 m (1.5–10 ft). Rock-bolt-support forces range 0.04–0.3 MN per bolt. Friction bolts support rock though frictional forces between the bolt and the drill hole.

Mechanical anchor bolts are anchored to the rock at the end of the drill hole and often grouted for additional support. Resin bolts use a quick-setting resin and steel dowels to support the rock. Wire mesh is often used in addition to bolts to contain small rock blocks between the bolts. Shotcrete is a form of sprayed concrete that is very effective in stabilizing rock excavations. The shotcrete mix contains cement, aggregate, and various agents to improve strength and ductility. Shotcrete thickness varies 25–300 mm (1–12 in.) and has a final (28 day) strength of 25–35 MPa. Various techniques are available to determine the type and amount of support required to stabilize a specific surface and underground design. Empirically based tables are commonly used to determine the support requirements, and most of these tables are based on rock-mass classification schemes. An example of an underground-excavation support table using the rock mass rating is shown in **Table 2**. The other method for estimating the support requirements for a particular rock engineering design is using numerical methods. Modern numerical programs allow detailed support designs to be modeled and the resulting deformation and failure of the rock mass to be analyzed. *See* MODEL THEORY.

Modern technologies and research. New technologies and research are essential elements of the field of rock mechanics. The most important new technologies that have been developed in the past 20 years are sophisticated numerical programs for modeling rock-mass deformation and failure. The numerical programs that are used in rock mechanics include finite-element models, boundary-element models, distinct-element models, finite difference models, micromechanical models, and many others. These include two- and three-dimensional models, constitutive relationships such as strain hardening and strain softening, large-strain, and coupled hydrothermomechanical models. New technologies for rock-mass characterization are just emerging.

In general, these are existing technologies that are being applied to rock-mass characterization for the purpose of increasing accuracy, safety, access, and human bias. These technologies include digital image processing, ground-based lidar, Geographical Information Systems (GIS), Geographical Positioning Systems (GPS), remote sensing, and geophysics. The important research areas include neural networks, scale effects (including fractals), coupled mechanical/hydraulic/thermal/chemical models, fracture mechanics, fracture flow, block theory, constitutive modeling, wellbore stability, contaminant transport/multiphase flow, and rock mechanics for long-term underground nuclear waste storage. *See* FINITE ELEMENT METHOD; GEOGRAPHIC INFORMATION SYSTEMS; GEOPHYSICS; IMAGE PROCESSING; LIDAR; REMOTE SENSING.

John M. Kemeny

Bibliography. B. H. G. Brady and E. T. Brown, *Rock Mechanics for Underground Mining*, 3d ed., Allen and Unwin, London, 2005; R. E. Goodman, *Introduction to Rock Mechanics*, 2d ed., Wiley, 1989; E. Hoek and E. T. Brown, *Underground Excavations in Rock*, Institution of Mining and Metallurgy, London, 1980; E. Hoek and J. W. Bray, *Rock Slope Engineering*, 3d ed., Institution of Mining and Metallurgy, London, 1981; E. Hoek, P. K. Kaiser, and W. F. Bawden, *Support of Underground Excavations in Hard Rock*, Balkema, Rotterdam, 1995; J. A. Hudson and J. P. Harrison, *Engineering Rock Mechanics: An Introduction to the Principles*, Pergamon, 1997; J. C. Jaeger and N. G. W. Cook, *Fundamentals of Rock Mechanics*, 2d ed., Chapman and Hall, London, 1979; S. D. Priest, *Discontinuity Analysis for Rock Engineering*, Chapman Hall, London, 1993.

Rock varnish

A dark to dark-brown coating on rock surfaces exposed to the atmosphere (**Fig. 1**). Rock varnish is probably the slowest-accumulating sedimentary deposit, growing at only a few micrometers to tens of micrometers per thousand years. Its thickness ranges from less than 5 μm to over 600 μm, and is typically 100 μm or so. Although found in all terrestrial environments, varnish is mostly developed and well preserved in arid to semiarid deserts; thus, another
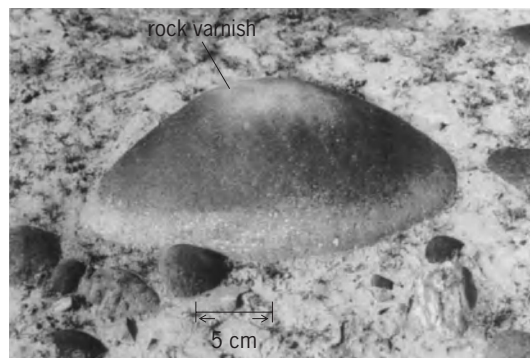

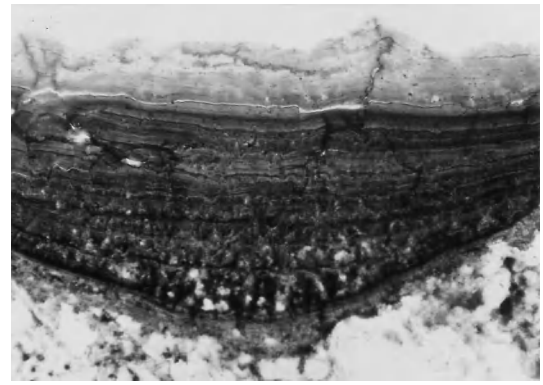Fig. 1. Varnished river cobble in western China.

Fig. 2. Microlaminations seen in a varnish ultrathin section under a light microscope. Maximum varnish thickness is 150 μm.

common name is desert varnish. Rock varnish is composed of about 30% manganese and iron oxides, up to 70% clay minerals, and over a dozen trace and rare-earth elements. The building blocks of rock varnish are mostly blown in as airborne dust. Although the mechanism responsible for the formation of rock varnish remains unclear, two hypotheses have been proposed to explain the great enrichment in manganese within varnish (typically over 50 times compared with the adjacent environment such as soils, underlying rock, or dust). The abiotic hypothesis assumes that small changes in pH can concentrate manganese by geochemical processes. The biotic hypothesis posits that bacteria, and perhaps other microorganisms, concentrate manganese; this is supported by culturing experiments and direct observations of bacterial enhancement of manganese. *See* SEDIMENTARY ROCKS.

Rock varnish has been scientifically studied since the 1950s largely because of the possibility of using its development as an indicator of time or environmental change. One outstanding technique in recent research is varnish microlaminations. Microlaminations are seen only when the varnish is shaved thin enough (5–10 μm) to see through with a light microscope (**Fig. 2**). The dark layers are rich in manganese (but poor in silicon, aluminum, and iron), while the lighter layers are poor in manganese (but rich in silicon, aluminum, and iron). Manganese-rich dark layers in varnish appear to form under relatively humid climate, and manganese-poor lighter layers under relatively dry climate. These two types of layers alternate to form distinct sequences. Thus, rock varnish can be used to reconstruct the moisture history of deserts where other records of climatic change are often hard to obtain or difficult to read. *See* VARNISH.

Since rock varnish records environmental, especially climatic, events that are regionally or even globally synchronous, varnish microstratigraphy can be used as a tephrachronology-like tool for age dating. Without radiometric calibration, varnish microstratigraphy itself may be used to estimate relative ages of varnished geomorphic or archeological features in deserts. Once calibrated, varnish microstratigraphy can provide numerical age estimates for geomorphic and archeological features. Specifically, for

petroglyphs and geoglyphs, the layering patterns of rock varnish hold the greatest potential for assigning ages. *See* ARCHEOLOGICAL CHRONOLOGY; DATING METHODS; ROCK AGE DETERMINATION.   Tanzhuo Liu

Bibliography. R. I. Dorn and T. M. Oberlander, Rock varnish, *Progr. Phys. Geog.*, 6:317–367, 1982; C. D. Elvidge and C. B. Moore, A model for desert varnish formation, *Geol. Soc. Amer. Abstr. Prog.*, 11:271, 1979; M. Fleisher et al., A clue regarding the origin of rock varnish, *Geophys. Res. Lett.*, 26:103–106, 1999; T. Liu et al., Terminal Pleistocene wet event recorded in rock varnish from southern Nevada, western United States, *Geol. Soc. Amer. Abstr. Prog.*, 30(7):361, 1998; R. M. Potter and G. R. Rossman, Desert varnish: The importance of clay minerals, *Science*, 196:1446–1448, 1977.

Rocket

Either a propulsion system or a complete vehicle driven by such a propulsive engine. A rocket engine provides the means whereby chemical matter is burned to release the energy stored in it and the energy expended, in this example, by ejection at high velocity of the products of combustion (the working fluid). The ejection imparts motion to the vehicle in a direction opposite to that of the ejected matter. A rocket vehicle is propelled by rocket reaction and includes all components necessary for such propulsion, a payload such as an explosive charge, scientific instruments, or a human crew. A rocket vehicle also includes guidance and control equipment mounted in a structural airframe or spaceframe. *See* ROCKET PROPULSION; ROCKET STAGING; SATELLITE (SPACECRAFT); SPACE FLIGHT; SPACE PROBE.   George P. Sutton

Rocket astronomy

The discipline that makes use of sounding rockets (**Fig. 1**) that fly near-vertical paths carrying scientific instruments to altitudes ranging from 25 to more than 900 mi (40 to 1500 km). Altitudes up to 30 mi (48 km) can be reached by balloons, so sounding rockets are typically used for higher altitudes in order to measure emissions from the Sun or other celestial sources that do not penetrate the Earth's atmosphere. Sounding rockets do not achieve orbital velocity; after completion of the launch phase, the payload follows a ballistic trajectory that permits 5–15 min of data taking before reentry. *See* ROCKET.

Mission profile. Rocket astronomy began in the United States in 1946, using captured German V-2 rockets with no pointing capability and very simple instrumentation. A typical modern launch is comparatively sophisticated, involving engine ignition, vehicle spinup, payload separation and despin, target acquisition, attitude control system (ACS) operation, telemetry and data command via the ground station, door closure or boom retraction, reentry and chute deployment, impact and recovery, data retrieval, and reuse of the scientific payload.
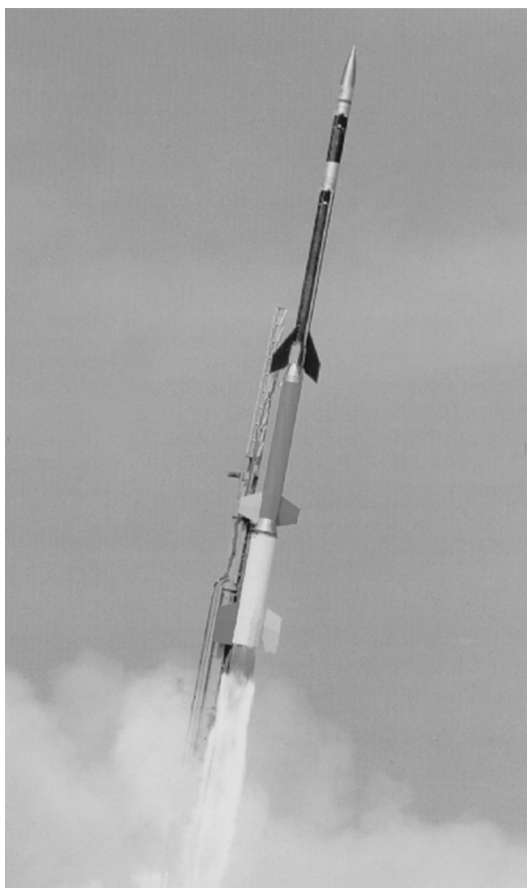

Fig. 1. Launch of a NASA Black Brant XII sounding rocket. (*NASA Goddard Space Flight Center*)

In comparison with experiments launched on satellites, rockets offer the advantages of simplicity, in that the payload is usually composed of a single instrument package; relatively frequent access to launch opportunities, typically once per year in solar physics; a shorter time scale from conception to reality, 3–5 years for a rocket versus 5–20 years for satellites; lower cost; and recoverability of the payload and the possibility of postflight instrument calibration, refurbishment, and reflight. Major disadvantages are short observing time, minutes versus months or years for satellites; localized coverage, restricted to an area adjacent to the launch site; and size and weight restrictions on payloads. *See* SATELLITE ASTRONOMY.

Over a dozen different launch vehicles are in common use (**Fig. 2**), ranging in length from 10 to 65 ft (3 to 20 m) and in diameter from 4.5 to 22 in. (11 to 56 cm). Payloads up to 2200 lb (1000 kg) can be carried to altitudes that are typical of those at which the space shuttle operates, 150–200 mi (250–320 km). For each launch vehicle the maximum altitude is strongly dependent upon payload weight (**Fig. 3**). For instance, the Black Brant X can carry a 200-lb (91-kg) payload to an altitude of 620 mi (1000 km) and can boost an 800-lb (364-kg) payload to 250 mi (400 km).

Sounding rockets are routinely launched in the United States from Wallops Island, Virginia; Poker
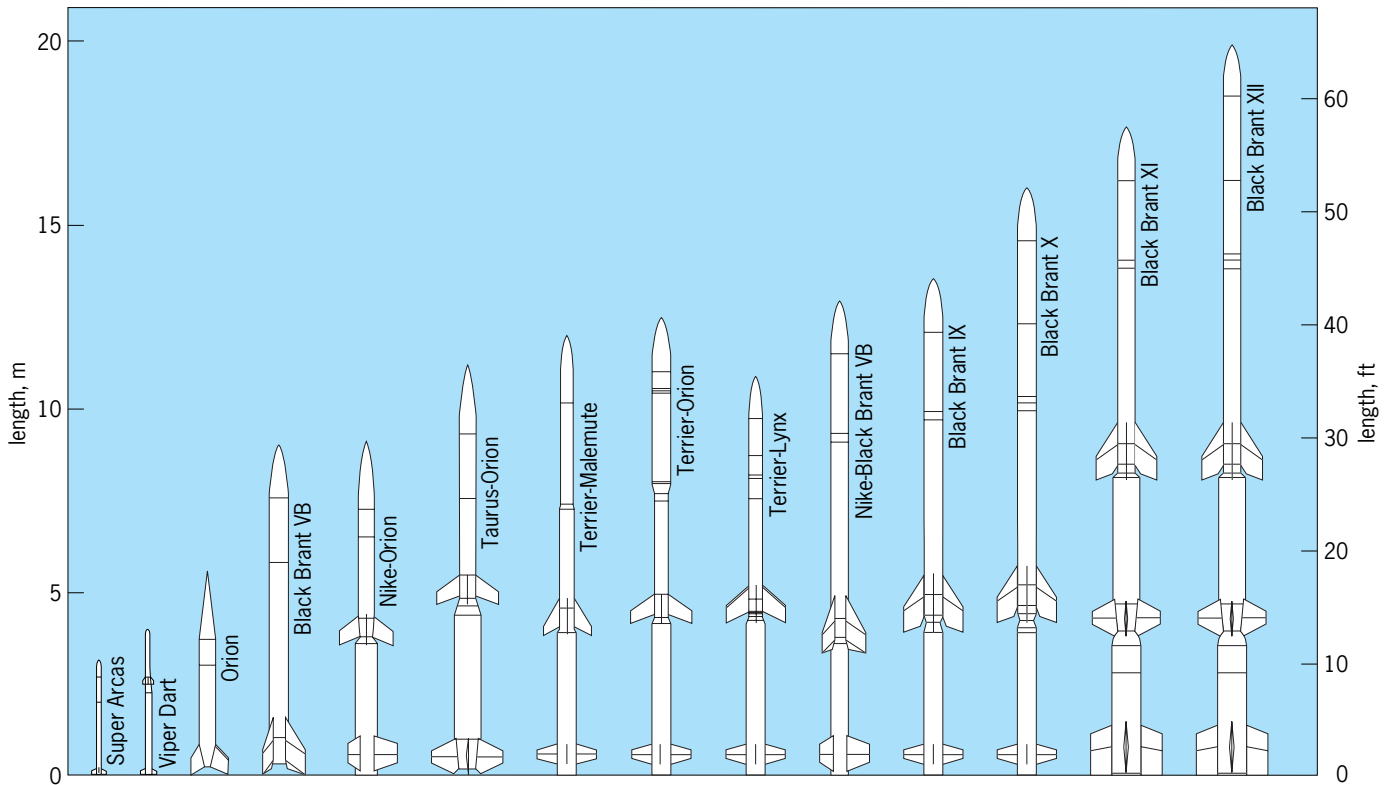
Fig. 2. NASA sounding rockets. (*NASA Sounding Rockets Working Group*)

Flats Research Range, Alaska; and White Sands Missile Range, New Mexico. Sites in Norway, Sweden, and Canada are also used. Temporary launch ranges are arranged when needed, and scientific requirements may dictate heavy usage of a particular facility. This was the case in 1988 and 1989 when numerous observations of Supernova 1987*a* were carried out from Woomera, in Australia. The United States sounding rocket program makes extensive use of surplus military rockets as an economical source of engines. NASA's Orion was formerly a Hawk ground-to-air missile, while Nike, Taurus, and Terrier engines are used as boosters in two- and three-stage flights. These components are usually supplied without guidance control systems.

The United States, Canada, Japan, and many European countries, including Germany, France, Sweden, Norway, Switzerland, and the United Kingdom, maintain vigorous scientific rocket programs. These programs focus on the disciplines of aeronomy, magnetospheric physics, meteorology, and material sciences, as well as astronomy and astrophysics. The ability to carry out vertical profile measurements of relevant atmospheric parameters at heights of 25–125 mi (40–200 km) is essential in many of these scientific disciplines. To the extent that the Sun influences or controls conditions in the upper atmosphere and magnetosphere of the Earth, there is a strong connection between solar astronomy and the more local research areas. *See* AERONOMY; IONOSPHERE; MAGNETOSPHERE.

Solar rocket astronomy. The Earth's atmosphere is opaque to wavelengths shorter than ~300 nanometers. Observations in the ultraviolet and x-ray regions of the spectrum therefore require that instruments be placed at altitudes exceeding 25–87 mi (40–140 km), depending upon wavelength. The outer portion of the Sun's atmosphere, the corona, is at a substantially higher temperature than is the photosphere, several million kelvins versus 6000 K. The emission from this region is dominated by ultraviolet and x-ray photons, whose total luminosity is very weak in comparison to the Sun's visible light output because of the extremely low density of the corona. During the late 1950s and early 1960s, techniques were developed for focusing x-rays and thereby providing direct imaging of the corona. These early studies revealed the highly structured nature of the atmosphere, with approximately semicircular loops of hot plasma outlining the shape of the underlying magnetic field, which confines the hot gas. *See* X-RAY TELESCOPE.

Areas of the solar surface at which strong magnetic fields emerge are known as active regions; in addition to producing strong x-ray and ultraviolet emission, they are the sites of explosive solar flare events. In such events the plasma is temporarily heated to temperatures exceeding 10^7 K by the rapid release of stored magnetic energy; in addition, material is often ejected into interplanetary space, and these events may have associated with them bursts of high-energy particles, microwave emission, gamma rays, and magnetic storms. Flares have long been known to cause disturbances of the Earth's atmosphere, and the largest flares pose a serious threat to astronauts situated above the atmosphere; forecasting of flares
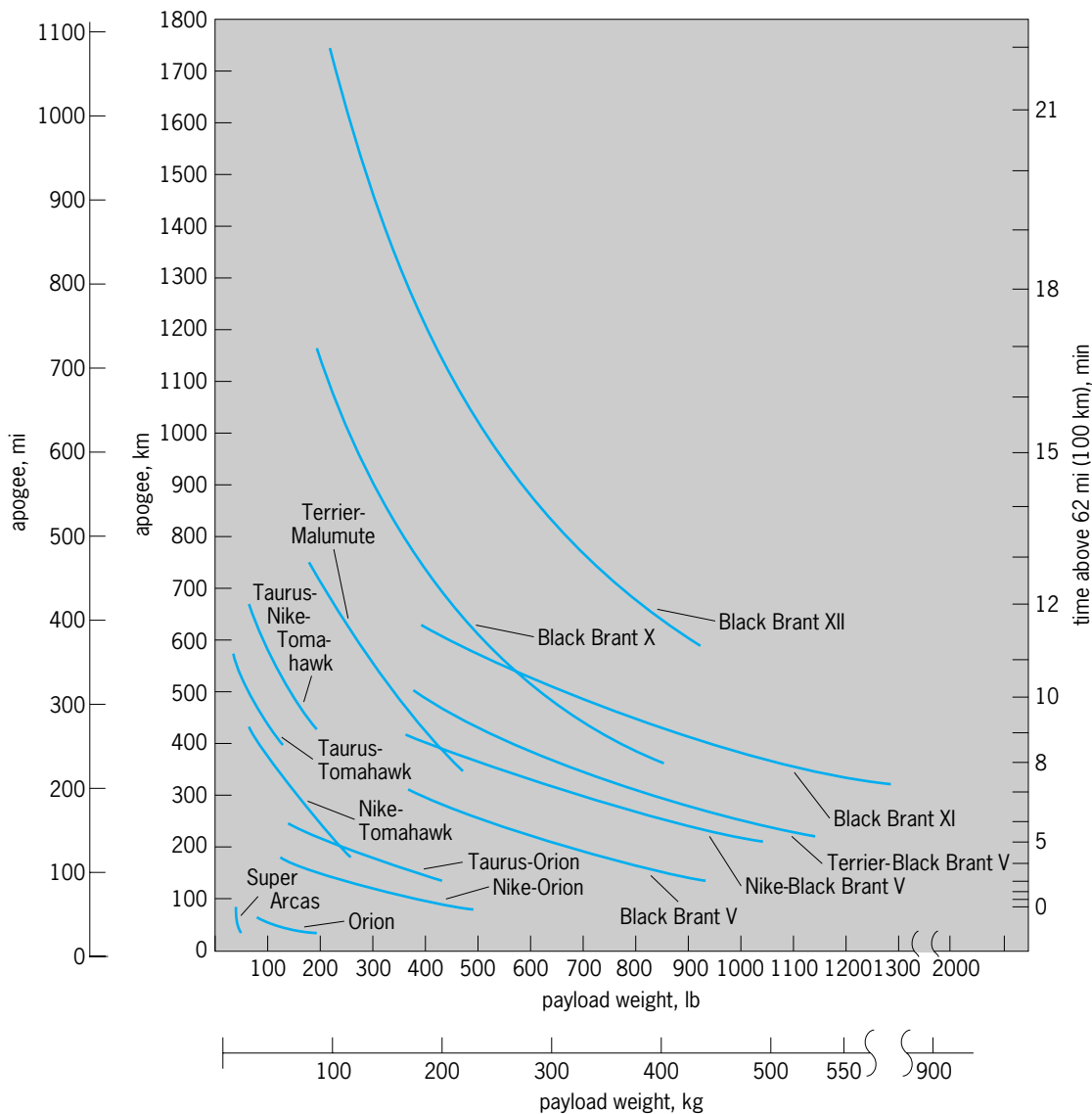
Fig. 3. Lift capability of NASA sounding rockets. Apogee is shown as function of payload weight.

is thus a major area of research. Analysis of the physical conditions leading to and during a flare involves observation of the entire three-dimensional atmospheric structure before a flare and the location and evolution of the hot plasma during the flare. Such studies require recording of all temperature regimes from 5000 to 10^7 K, which implies the ability to observe wavelengths from the infrared, through the ultraviolet and soft x-ray, into the hard x-ray (10-keV) regime, plus microwave and other radio emission. Such data cannot be obtained by a single technique, and solar research satellites are therefore composed of multiple co-observing instruments combining both imaging and spectroscopy, and including ground-based observations as a necessary part of the overall scientific program.

Among the notable advances in coronal studies from rockets has been the development of a new technique for x-ray imaging, with concomitant improvements in imaging spectroscope methods. The use of multilayer coatings for enhanced x-ray re-

flectivity may be viewed as depositing an artificial Bragg crystal onto a figured substrate. This technique offers the advantage of substantially higher image sharpness, and provides as well simultaneous spectroscopy because of the narrow x-ray bandwidth of the coatings. The major technological limitation of this technique is the inability to produce adequate normal-incidence reflectivity at short wavelengths (less than 2 nm); in this spectral region, grazing-incidence methods continue to be the most appropriate. *See* SUN; X-RAY DIFFRACTION.

Other celestial sources. Observations of nonsolar sources from sounding rockets encounter two major difficulties: the low intensity of the emission and the technological problem of pointing the payload at the source during the flight. The two problems are related in that the low flux levels generally lead to a requirement for long exposure or integration times, and the faintness of the source leads to difficulty in constructing a sensor that can provide the appropriate detection signal to an attitude control system.

The development of a gas-jet attitude control system for Aerobee rockets in 1964 and the addition of gyroscopic stabilization in 1965 permitted ultraviolet spectroscopy of stars with high enough dispersion to record interstellar absorption features in the 130–160-nm range. Such data permit measurement of interstellar hydrogen, carbon, nitrogen, oxygen, silicon, and sulfur. Spectra in the 100–250-nm region permit studies of stars with surface temperatures of 10,000–30,000 K; in addition to total-luminosity determination, a major discovery from line-profile measurements of hot-star absorption features was the finding that such stars have large mass loss due to high-velocity (greater than 600 mi/s or 10^3 km/s) stellar winds. In 1970, a sounding rocket detected the absorption line of interstellar molecular hydrogen; this detection was the forerunner of later extensive studies from the *Copernicus* satellite, which have had a major impact on the field of interstellar chemistry. *See* ASTRONOMICAL SPECTROSCOPY; COSMOCHEMISTRY; INTERSTELLAR MATTER; SPACECRAFT PROPULSION; ULTRAVIOLET ASTRONOMY.

In 1962, a rocket experiment detected nonsolar x-rays for the first time, at a level many orders of magnitude higher than would be produced by an equivalently placed solar-strength source. Subsequent flights confirmed the initial result and discovered other x-ray sources, as well as a diffuse x-ray background and an emission source in the Crab Nebula. The field of x-ray astronomy is now dominated by observations from satellites, which have the ability to carry out all-sky surveys and to integrate for many hours in order to study low-intensity sources. *See* X-RAY ASTRONOMY.   Leon Golub

Bibliography. American Institute of Aeronautics and Astronautics, *Proceedings of the 7th Conference on Sounding Rockets, Balloons, and Related Space Systems*, 1986; G. V. Groves (ed.), *Dynamics of Rockets and Satellites*, 1965; H. E. Newell, Jr., *Beyond the Atmosphere*, NASA Publ. SP-4211, 1980; H. E. Newell, Jr., *Sounding Rockets*, 1959; A. I. Skoog (ed.), *History of Rocketry and Astronautics*, American Astronautical Society, 1990.

Rocket propulsion

The process of imparting a force to a flying vehicle, such as a missile or a spacecraft, by the momentum of ejected matter. This matter, called propellant, is stored in the vehicle and ejected at high velocity. In chemical rockets the propellants are chemical compounds that undergo a chemical combustion reaction, releasing the energy for thermodynamically accelerating and ejecting the gaseous reaction products at high velocities. Chemical rocket propulsion is thus differentiated from other types of rocket propulsion, which use nuclear, solar, or electrical energy as their power source and which may use mechanisms other than the adiabatic expansion of a gas for achieving a high ejection velocity. *See* ELECTROTHERMAL PROPULSION; INTERPLANETARY PROPULSION; ION PROPULSION; PLASMA PROPULSION; PROPULSION; SPACECRAFT PROPULSION.

In this article the performance criteria, the basic features, the application, and the preparation of various rocket propulsion systems that are driven by the combustion energy of chemicals are described. Propulsion systems using liquid propellants (such as kerosine and liquid oxygen) have traditionally been called rocket engines, and those that use propellants in solid form have been called rocket motors.

Performance. The performance of a missile or space vehicle propelled by a rocket propulsion system is usually expressed in terms of such parameters as range, maximum velocity increase of flight, payload, maximum altitude, or time to reach a given target. Propulsion performance parameters (such as rocket exhaust velocity, specific impulse, thrust, or propulsion system weight) are used in computing these vehicle performance criteria. **Table 1** gives typical performance values.

The momentum imparted to the vehicle by a rocket depends directly on the exhaust velocity v at which the hot gases are expelled from the rocket in a supersonic nozzle. This supersonic exhaust velocity is a good indicator of the effectiveness of any one propellant combination and the gas expansion in the nozzle. It is given by Eq. (1), where R is the

$$v = \sqrt{\frac{2gkR}{k-1}\frac{T}{M}\left[1 - \left(\frac{p_2}{p_1}\right)^{(k-1)/k}\right]} \qquad (1)$$

universal gas constant, k is the specific heat ratio of the reaction product gases, and g is the acceleration of gravity (32.2 ft/s^2 or 9.8 m/s^2). When SI units are used, the factor g can be deleted. The pressure of the nozzle exit at optimum nozzle expansion area ratio is p_2 and the chamber pressure is p_1. Because of this

---

TABLE 1. Typical performance values of rocket propulsion systems*

Propulsion system parameter	Typical range of values
Specific impulse at sea level	180–390 s
Specific impulse at altitude	215–470 s
Exhaust velocity at sea level	5800–15,000 ft/s (1800–4500 m/s)
Combustion temperature	4000–7200°F (2200–4000 °C)
Chamber pressures	100–3000 lb/in.^{2} (0.7–20 MPa)
Ratio of thrust to propulsion system weight	20–150
Thrust	0.01–6.6 × 10^6 lb (0.05–2.9 × 10^7 N)[†]
Flight speeds	0–50,000 ft/s (0–15,000 m/s)

*Exact values depend on application, propulsion system design, and propellant selection.
[†]Maximum value applies to a cluster; for a single rocket motor it is 3.3 × 10^6 lb (14,700 kN).

dependence on the pressure ratio, there is a slight increase (10–20%) in exhaust velocity as the exit pressure is decreased (increase in altitude) or as the chamber pressure is increased. The exhaust velocity increases as the molecular weight of the combustion gases M decreases and the combustion temperature T increases, but exhaust velocity also increases if the pressure ratio across the exhaust nozzle increases.

The specific impulse is a key rocket propulsion performance parameter, and it can be considered the thrust obtained from an equivalent rocket propulsion system that has a propellant mass flow of unity. It is expressed in pounds of force per pound of mass per second, or the exhaust velocity v divided by the acceleration of gravity. In the foot-pound-second system of units it is sometimes expressed simply as seconds. In SI units it is expressed in newtons per kilogram per second. *See* SPECIFIC IMPULSE.

Exhaust velocities and specific impulses can be calculated from thermochemical relations; the values so obtained are known as the theoretical specific impulses and theoretical nozzle exhaust velocities. Actual values are 92–98% of their theoretical values, because the losses due to friction, nozzle exhaust, divergence, and incomplete combustion are relatively small.

Thrust F is defined by Eq. (2) in terms of a momen-

$$F = \dot{m}v + (p_2 - p_3)A_e \qquad (2)$$

tum and a pressure term, where $\dot{m}$ is the mass flow through the nozzle, v is the average nozzle exit velocity, A_e is the nozzle exit area, and pressure difference $(p_2 - p_3)$ is that between nozzle exit pressure p_2 and ambient pressure p_3. As the mass flow increases, so does the thrust. Because p_2 is usually close to p_3, the second term of the thrust equation is small, and the thrust is roughly equal to the product of mass flow $\dot{m}$ and exhaust velocity v. As the altitude increases (p_3 decreases), the velocity and the thrust increase (**Fig. 1**).

The nozzle area ratio is the nozzle exit area divided by the nozzle throat area. If this ratio is too large, then the gas flow in the nozzle overexpands, the value of p_2 is less than p_3 (the last term in the thrust equation becomes negative), and the thrust decreases. The values of nozzle area ratio are low (between 3 and 20) for rocket propulsion systems operating at
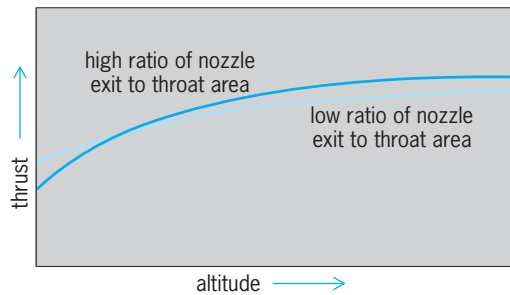

Fig. 1. Typical variation of thrust with altitude.

low altitudes. For high-altitude operation, where p_3 is close to zero, a high area ratio (between 20 and 400) allows a 10–25% increase in performance over sea level; this means a relatively large exit nozzle. *See* NOZZLE.

Clever designs and highly stressed materials are used to make propulsion system weights as low as possible so as to minimize the amount of energy necessary to accelerate inert mass. The best vehicle performance is obtained when the propellant mass is a maximum and the inert mass of the propulsion system, the residual propellants, structure, controls, nonpropulsive upper stage, or payload is a minimum.

Applications. Rocket propulsion is used for different military missiles or space-flight missions. Each requires different thrust levels, operating durations, and other capabilities. **Tables 2** and **3** show some of these and also the preference for liquid or solid propellants, but in some applications both types are used. In addition, rocket propulsion systems are used for rocket sleds, jet-assisted takeoff, principal power plants for experimental aircraft, or weather sounding rockets. For some space-flight applications, systems other than chemical rockets are used or are being investigated for possible future use. *See* GUIDED MISSILE; MISSILE; SATELLITE (SPACECRAFT); SPACE FLIGHT; SPACE PROBE.

The space shuttle is an example of a multistage vehicle that uses a variety of propulsion systems. It has 67 individual propulsion units, located in different places on the vehicle (**Fig. 2**). Each of the several rocket propulsion systems has its own application requirements (**Table 4**). *See* ROCKET STAGING; SPACE SHUTTLE.

TABLE 2. Space flight application of rocket propulsion systems			
Type	Liquid (L) or solid (S) propellant	Thrust range, lbf (N)	Special capabilities
Launch vehicle			
Booster	S/L	20,000–7,000,000 (90,000–31,000,000)	Reuse
Sustainer(s)	L/S	2000–1,000,000 (9000–4,500,000)	Throttling, reuse
Attitude control	L	1–1000 (4.5–4500)	Many restarts, reuse
Separation	S	200–20,000 (900–90,000)	
Spacecraft*			
Main propulsion	L	200–15,000 (900–67,000)	Throttling (variable thrust)
Attitude/maneuver	L	0.01–300 (0.045–1350)	Many restarts
Planetary/lunar landing	L	100–10,000 (450–45,000)	Throttling, restarts

*For reusable, orbiting space shuttle, see Table 4.

TABLE 3. Military applications of rocket propulsion systems

Type	Liquid (L) or solid (S) propellant	Thrust, lbf (N)	Maximum acceleration* and special capabilities	Example
Surface-to-surface long-range or antisatellite (1, 2, or 3 stages)				
Booster	S/(L)	20,000–2,500,000	1.2–6 g	Titan (2 or 3 stages)
Sustainer(s)	S/(L)	2000–500,000		Minuteman (3 stages)
		(90,000–11,000,000)	0.5–5 g	Poseidon (2 stages)
Attitude/maneuvers	L	1–3000		Pershing (2 stages)
		(4.5–13,500)	0.01–1 g	
Air-to-surface (1 or 2 stages)				
Boost/sustain	S	1000–35,000	Two thrust levels	SRAM, Shrike
		(4500–155,00)		
Surface-to-air (1 or 2 stages)				
Boost/sustain	S	800–30,000	Regressive up to 15g	Redeye, Chapparell,
		(3600–135,000)		Standard
Air-to-air				
Boost/sustain	S	1000–35,000	Two thrust levels	Phoenix, Sparrow,
		(4500–155,000)		Falcon, Sidewinder
Infantry support				
Antitank	S	500–8000	1–15 g	Lance (L), Shillelagh,
		(2200–36,000)		
Short range	S/(L)	500–15,000		Hellfire
		(2200–67,000)		
Ballistic missile defense				
Booster	S	50,000–500,000	3–100 g	Being developed
		(220,000–2,200,000)		
Sustainers	L/(S)	5000–50,000	2 to 50 g	Being developed
		(22,000–220,000)		
Experimental aircraft	L	3000–7000	0.5–4.0 g	X-1 and X-15
		(13,500–31,000)		research aircraft
Assisted takeoff	S	250–10,000		
		(1100–45,000)		

*1 g = 32.2 ft/s^2 = 9.8 m/s^2.
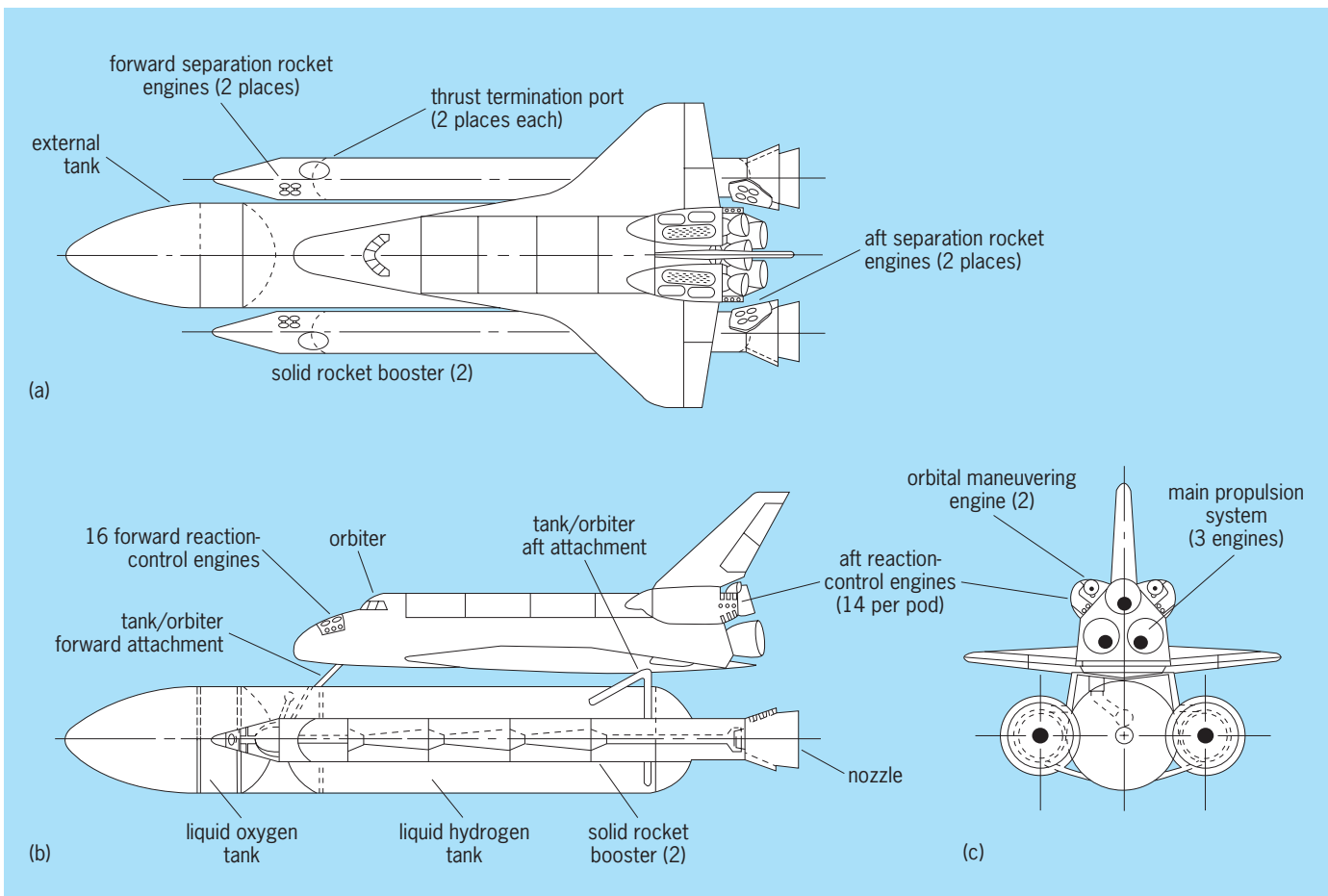
Fig. 2. Space shuttle and its propulsion systems. (*a*) Top view. (*b*) Side view. (*c*) Rear view. (*NASA*)

TABLE 4. Rocket propulsion systems used on the space shuttle

Type	Number of units	Propellants	Thrust, lbf (kN)	Specific impulse,* lbf·s/lbm	Gimbal angle, degrees	Absolute chamber pressure, lbf/in.² (MPa)	Nozzle area ratio
Main engine	3	Liquid oxygen and liquid hydrogen Mixture ratio = 6	At 100% thrust: 370,000 (1650) at sea level, 470,000 (2090) at altitude; throttled between 65 and 109%	454 at altitude	±10.5	~3000 (~21)	77.5
Orbital maneuvering system engine	2	Nitrogen tetroxide and monomethyl-hydrazine Mixture ratio = 1.65	6000 (26.7) in orbit	313 at altitude	±6° pitch, ±7° yaw	125 (0.86)	55
Reaction control system: Primary engine	38	Nitrogen tetroxide and monomethyl-hydrazine	870 (3.87) at altitude	~280	None	~152 (1.05)	20 to 30
Vernier engines	6	Mixture ratio = 1.6	24 (0.106)	~265	None	110 (0.76)	~22
Solid rocket booster motor	2	Composite solid: 16% aluminum, 69.8% ammonium perchlorate, 12% polymer, 1.96% epoxy curing agent, 0.17% burn-rate controller	3.3×10^6 (14.680) at sea level; 2.3×10^5 (10,230) at altitude	274	±8°	950 to 600 (6.5 to 4.1)	~7
Booster separation rocket motor	16	Composite solid: ammonium perchlorate, polybutadiene, 2% aluminum	~20,050 (~89) Lift-off thrust: $~7.7 \times 10^6$ (34,300)	225	None	1800 (12.4)	5.83
	Total number of units: 67						

*Specific impulse is usually expressed in the units given here, even where SI is used.

The three main engines are high-performance liquid-propellant rocket engines, which provide the major velocity increase to the space shuttle. During a typical shuttle flight they may operate from 8 to 9 min, which is rather long for any rocket propulsion system. Throttling of the thrust down to 65% allows the vehicle to limit its aerodynamic forces during ascent and also to limit the maximum vehicle acceleration for the last minute or two of its powered flight.
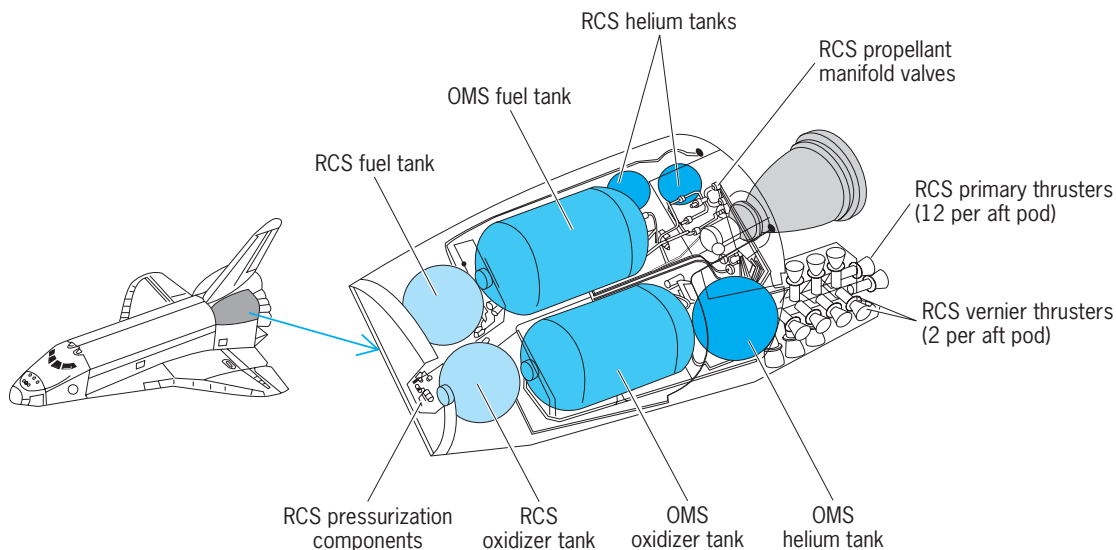
Fig. 3. One of two space shuttle aft pods containing the rocket engines of the orbital maneuvering system (OMS) and reaction control system (RCS). (*NASA*)

The orbital maneuvering system (OMS) has two liquid-propellant multipurpose engines that provide the thrust for orbit injection, orbit circularization, orbit transfer, rendezvous, deorbit, and several potential abort maneuvers. The reaction control system (RCS) has two sizes of liquid-propellant engines (38 primary engines and 6 smaller vernier engines) at three different locations. Any one engine may need to operate for thousands of pulses during a flight. The reaction control system provides attitude control of the orbiter vehicle in pitch, yaw, and roll; provides precise vehicle velocity corrections in three directions; and can assist in orbital maneuvers. Multiple primary engines provide for redundancy. There are two engine pods at the aft end of the space shuttle orbiter vehicle (**Fig. 3**), each of which contains one engine of the orbital maneuvering system with 6000 lb (26,700 newtons) vacuum thrust, and several engines of the reaction control system, together with their tanks of storable liquid propellants, valves, piping, and helium pressurization systems.

The two solid rocket boosters (SRBs) provide the principal thrust for lifting the space shuttle off the launch pad to about 150,000 ft (45,000 m). They are the largest solid-propellant rocket motors ever flown and the first designed for reuse. Each motor is made of four matched segments, which are assembled at the launch site. The pairs of boosters are matched by loading each of the four segments in pairs from the same batches of propellant ingredients so as to minimize any difference in burning rate between the two motors and thus also any thrust imbalance. The lift-off thrust is reduced to about two-thirds of its maximum value to prevent excessive stresses during ascent at maximum dynamic pressure. Each booster assembly has eight booster separation motors, four in the nose and four in the aft skirt, which burn for 0.8 s and push the spent boosters from the flying vehicle.

Liquid-propellant rocket engines. These use liquid propellants stored in the vehicle for their chemical combustion energy. The principal hardware subsystems are one or more thrust chambers, a propellant feed system, which includes the propellant tanks in the vehicle, and a control system.

Liquid propellants. **Table 5** lists several common liquid propellants. Bipropellants have a separate oxidizer liquid (such as liquefied oxygen or nitrogen tetroxide) and a separate fuel liquid (such as liquefied hydrogen or hydrazine). Those propellant combinations, which ignite spontaneously when the oxidizer liquid comes in contact with the fuel, are called hypergolic. Those that mix without self-ignition could result in a potentially explosive mixture, unless an ignition system (sparks, glowplugs, or a solid-propellant igniter) provides the initial energy to start the combustion. Cryogenic propellants (for example, liquid oxygen and liquid hydrogen) are those that are liquid only at very low temperature, evaporate rapidly at ambient temperature, and require

TABLE 5. Theoretical performance of liquid rocket propellant combinations

Propellant type (oxidizer/fuel)	Mass mixture ratio (oxidizer-to-fuel)	Bulk density, g/cm^3	Combustion temperature, °F(°C)	Molecular mass of chamber gases, lb/mole or kg/mole	Specific impulse,[*] lbf·s/lbm	Typical application	Vehicle where currently used	Hypergolic ignition
Cryogenic liquid propellant								
Oxygen/kerosine	2.56	1.02	6150 (3400)	23.2	294	Space launch vehicles	Atlas missile and first stage of Saturn	No
Oxygen/hydrogen	4.02	0.25	4935 (2724)	10.0	390	Space launch vehicles	Space shuttle main engine	No
Fluorine/hydrogen	7.60	0.35	6505 (3596)	11.8	405	Experimental		Yes
Storable liquid propellant								
Red fuming nitric acid/hydrazine mixture	1.47	1.23	5200 (2871)	22	280	Target drone	Sounding rocket	Yes
Red fuming nitric acid/kerosine	4.8	1.35	5590 (3088)	22.6	268	Surface-to-surface missile	Non-U.S. missiles	No
Nitrogen tetroxide/ 50% hydrazine with 50% dimethyl-hydrazine	2.0	1.21	5725 (3163)	22.6	283	Ballistic missiles, attitude control	Titan missile Agena vehicle	Yes
Nitrogen tetroxide/ monomethyl-hydrazine	1.6	1.23	5180 (2860)	22.5	282	Spacecraft attitude control	Space shuttle RCS engines	Yes
Liquid monopropellant								
90% hydrogen peroxide (at 300 lb/in.^{2} abs or 2.07 MPa abs)	—	1.44	1370 (743)	21.5	132	Gas generator for turbopump		Needs catalyst
Hydrazine (at 150 lb/in.^{2} abs or 1.03 MPa abs)	—	1.01	1115 (602)	10.7	155	Spacecraft attitude control	Deep space mission	Needs catalyst

[*]At a combustion chamber pressure of 1000 lb/in.^{2} abs (6.9 MPa abs) and full expansion at sea level to a nozzle exit pressure of 14.7 lb/in.^{2} abs (101 kPa abs). Special impulse is usually expressed in units given here, even where SI is used.
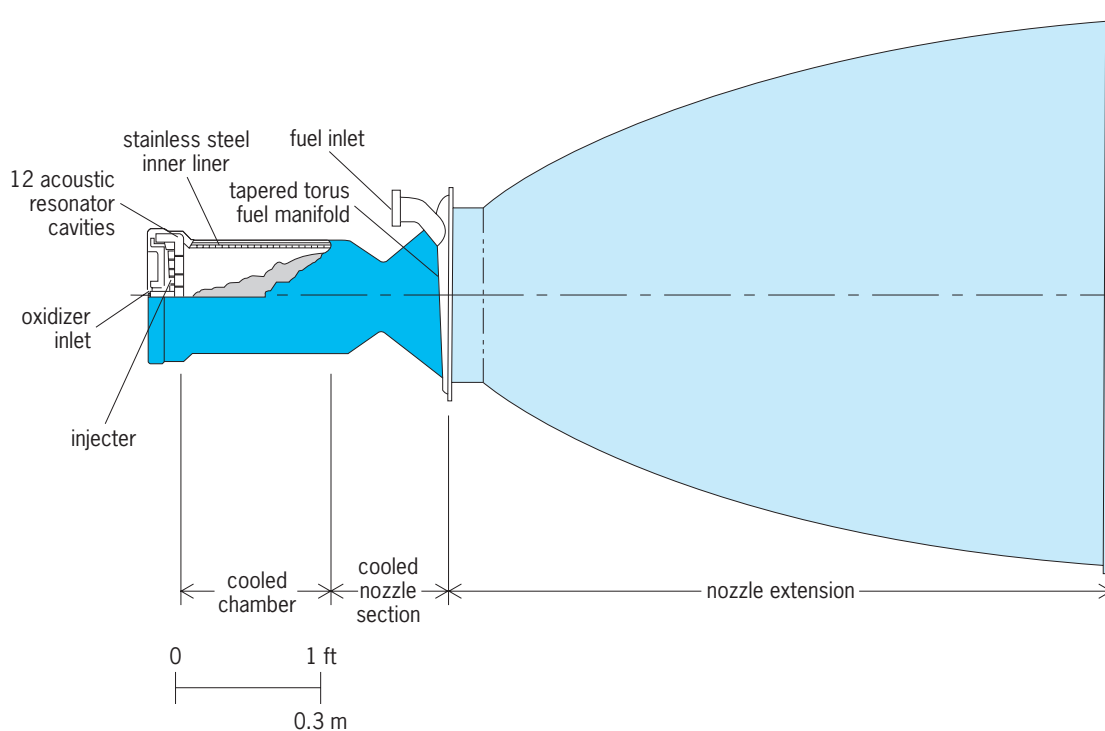
Fig. 4. Simplified half-section of one of the two thrust chambers of the orbital maneuvering engines used on the space shuttle. (*Aerojet Tech Systems*)

thermal insulation or refrigeration for prolonged storage. Storable propellants (for example, nitrogen tetroxide, hydrazine, and kerosine) are liquid at ambient temperatures.

Monopropellants consist of a single liquid that contains both oxidizer and fuel ingredients. A catalyst is required to decompose the monopropellant into gaseous combustion products. Bipropellant combinations allow higher preformance (higher specific impulse) than monopropellants. Hydrazine can be used as a monopropellant or, in combination with an oxidizer liquid, it can be a fuel for a bipropellant.

Liquid propellants have some undesirable properties and require careful handling. They each have their own peculiar types of hazard and all are either corrosive, toxic, flammable, or explosive; some propellants have all four of these problems. Special procedures, compatible containment materials, various safety precautions, and, in some cases, gas masks and protective clothing are required. *See* PROPELLANT.

Thrust chamber. The three principal components (**Fig. 4**) are the combustion chamber, where rapid, high-temperature combustion takes place; the converging-diverging nozzle, where the hot reaction-product gases are accelerated to supersonic velocities; and an injector (**Fig. 5**), which meters the flow of propellants in the desired mixture of fuel and oxidizer, introduces the propellants into the combustion chamber, and causes them to be atomized or broken up into small droplets. If the propellants are not hypergolic, then the thrust chamber also has an ignition system.

The burning of the liquid propellants takes place inside the combustion chamber at high chamber pressure, typically between 60 and 300 lb/in.^{2} (0.4

and 2 megapascals) for pressurized feed systems and between 300 and 3000 lb/in.^{2} (2 and 20 MPa) for pump-fed systems. The combustion is an intense, turbulent, unsteady, and noisy process with many strong, very rapid chemical reactions occurring simultaneously. Special design provisions, such as cavities or injector baffles, are often included to prevent combustion vibration from building up excessive pressure amplitudes that could damage or destroy the engine.

Some thrust chambers (such as the space shuttle's main engines and orbital maneuvering engines) are gimbaled or swiveled to allow a change in the direction of the thrust vector for vehicle flight motion control (**Fig. 6**). The angular movement of the nozzle is controlled by actuators; this enables controlling the vehicle in pitch, yaw, and roll. However, roll control requires at least two nozzles (or jet vanes).

Since the combustion temperature is roughly twice as high as the melting point of steel, bipropellant rocket engines must be cooled if they operate for longer than a few seconds. Cooling can be accomplished by (1) circulating one of the propellants in a cooling jacket around the chamber and nozzle and absorbing heat in the propellant prior to its introduction into the chamber (best for long duration); or (2) by using ablative liners or layers bonded to the inside of the chamber and nozzle; these liners absorb heat by melting, charring, or vaporizing some of the composite fiber-reinforced liner material, which then forms a cool-gas boundary layer near the wall surfaces. For monopropellant thrust chambers and for areas near the nozzle exit of a bipropellant engine (where the local propellant reaction-gas temperature is relatively low, 1600–2800°F or 900–1500°C), a radiation-cooled, simple, formed,
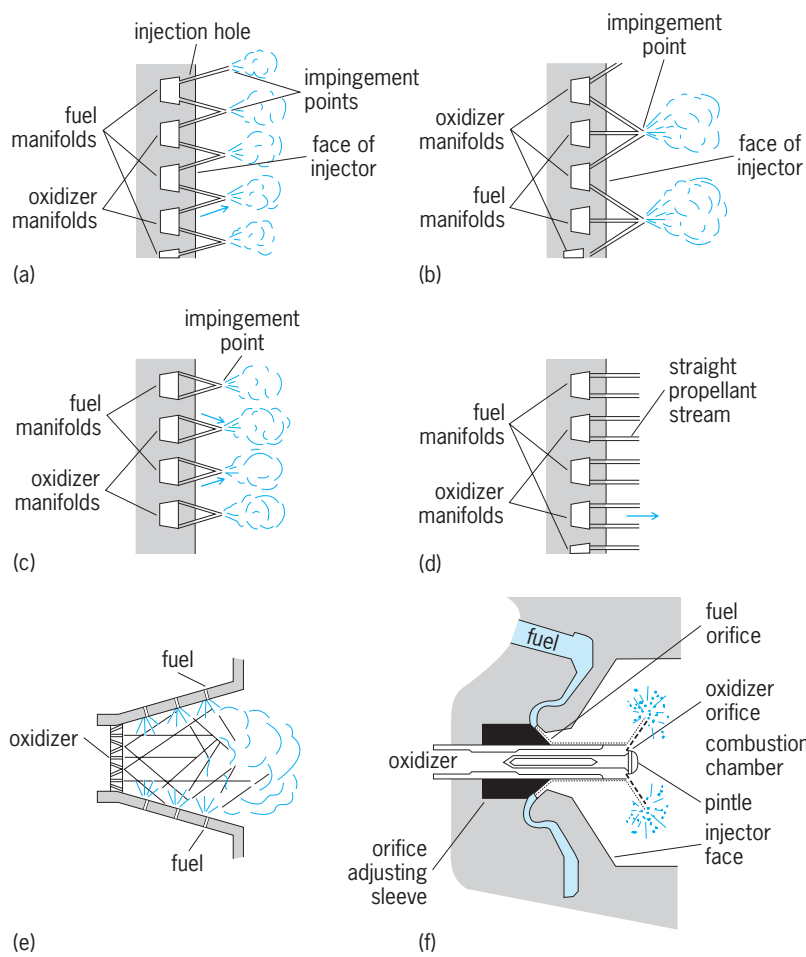
Fig. 5. Several types of injectors. (*a*) Doublet-impinging stream pattern. (*b*) Triplet-impinging stream pattern. (*c*) Self-impinging stream pattern. (*d*) Shower-head stream pattern. (*e*) Nonimpinging stream pattern. (*f*) Variable-injection-area concentric-tube injector. (*After G. P. Sutton*, Rocket Propulsion Elements, *5th ed.*, *John Wiley and Sons*, *1986*)

heat-resistant metal such as coated niobium is usually satisfactory.

The chamber and nozzle throat region of each engine of the space shuttle's orbital maneuvering system (Fig. 4) are fuel-cooled. There is a stainless steel inner liner with 120 grooves for the coolant passage and an electroformed nickel closure. The nozzle exit skirt is a thin niobium alloy shell. The injector is of the platelet type and has 12 acoustic resonator cavities at its periphery to suppress combustion vibration.

Feed systems. Liquid-propellant rocket engine transfers the liquid propellants from the vehicle storage tanks to one or more thrust chambers. There are two basic types: gas-pressure feed systems and turbopump feed systems.

In a gas-pressure feed system, propellants are displaced with a high-pressure gas which is fed into the tanks under a regulated pressure (**Fig. 7**). The stored high-pressure gas furnishes the pressurization energy. The thrust of a pressurized-gas rocket propulsion system is determined by the magnitude of the propellant flow as controlled by the gas pressure regulator setting. For low thrust and short duration this feed system is generally lighter and superior to other more complicated ones.

For spacecraft or satellite attitude control and maneuvering, the use of a pressurized feed system, together with storable liquid propellants, permits rapid pulsing of several thrust chambers for spacecraft stabilization, station keeping, or rendezvous and position control. Small thrust chambers of perhaps 1–1000-lb (4.5–4500-N) thrust are pulsed for short duration (as little as 0.01 s of operation) and many thousands of pulses during a given mission. Several thrust chambers (2–16) are fed propellants from a single pressurized feed system.

Fig. 6. Several thrust-vector control methods. L indicates use with liquid propllant engines, and S indicates use with solid-propellant motors. Actuators are not shown.
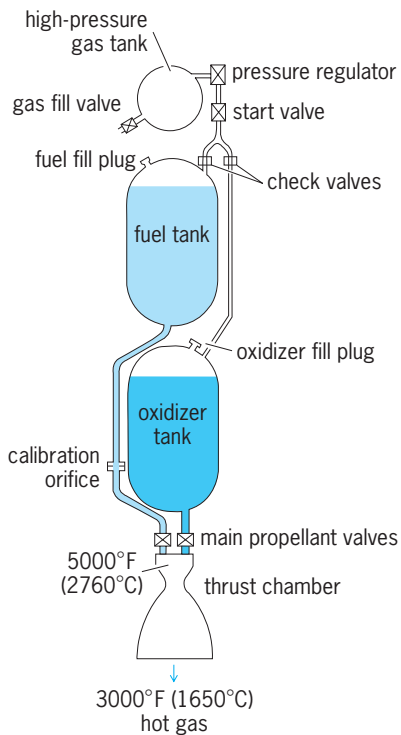
Fig. 7. Liquid-propellant rocket engine with gas-pressurized feed system.

In a field of zero gravity, as exists in outer space, the propellant liquids float around in a partially emptied tank; it is therefore necessary to include provisions for assuring that the outlet pipe opening in the propellant tank is covered with liquid (and not with the pressurized gas, which is also inside the tank) to permit proper starting and operation of the rocket. This has been accomplished by surface-tension devices, such as screens, or by including a positive expulsion device, such as a flexible bladder (made usually of thin metal or plastic), which separates the propellant liquid at all times from the pressurized gas. Pressure-fed systems are used on both the orbital maneuvering system and the reaction control system of the space shuttle.

In a turbopump feed system the propellant is pressurized by means of pumps driven by one or more turbines (**Fig. 8**) which derive their power from the expansion of hot gases. A separate gas generator ordinarily produces these gases in the required quantities and at a temperature which will not damage or partially melt the turbine buckets (1200–1800°F or 650–980°C). This is achieved in a chemical combustion chamber similar to the main thrust chamber, but the propellants are burned at a different and cooler mixture ratio. The assembly of pumps and turbine is termed a turbopump. Like other rotating machinery, it has bearings, seals, high-speed shafts, and lubrication devices. Turbopump feeds are generally superior to and lighter than other feed systems for high-thrust and long-duration applications. *See* GAS TURBINE.

For the large propellant flow and the high pressures involved in a rocket propulsion system, a centrifugal pump is generally considered superior to other types of pumps and is economical in weight

and space. A multiblade centrifugal impeller is used with a volute casing. For fluids of low density, like liquid hydrogen, it is necessary to use multiple impeller stages of axial-flow pumps to develop the required inlet pressures at the thrust chamber. It is important that the pressure at the inlet of the pump always be sufficiently high to avoid low-pressure boiling, which induces cavitation. Special impellers are often used to prevent cavitation because of its detrimental effects on the operation and combustion stability of the rocket. *See* CAVITATION; CENTRIFUGAL PUMP.

The hot exhaust gases from the turbine are often used for heating pressurizing gas or evaporating cryogenic propellant in order to pressurize a propellant tank or for producing a small additional thrust by ejecting these gases through a nozzle. A turbopump system can be started in several ways. In one method, a low-temperature solid-propellant cartridge is ignited and starts the turbine rotation, which then allows the pumps to raise the propellant pressures and to feed propellant to the combustion chambers.

The space shuttle main engine (**Fig. 9**) uses a turbopump feed. The cryogenic propellants, liquid hydrogen and liquid oxygen, are carried in separate slightly pressurized tanks in the external tank (Fig. 2), and are transferred to the shuttle orbiter, where the engine's four turbopumps feed it to the combustion chambers. The gases in the preburners that drive the turbopumps flow through the injector into the main combustion chamber and are mixed and burned with the liquid oxygen and gaseous hydrogen supplied to the thrust chamber. This unique
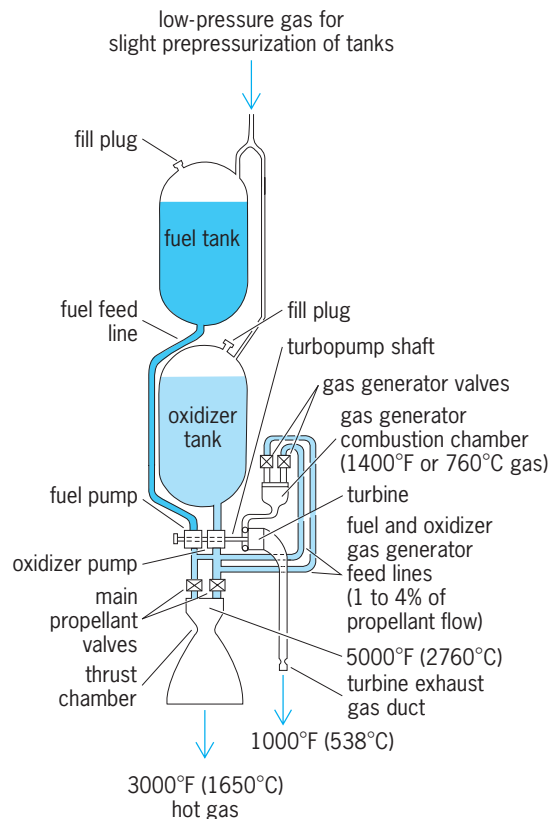

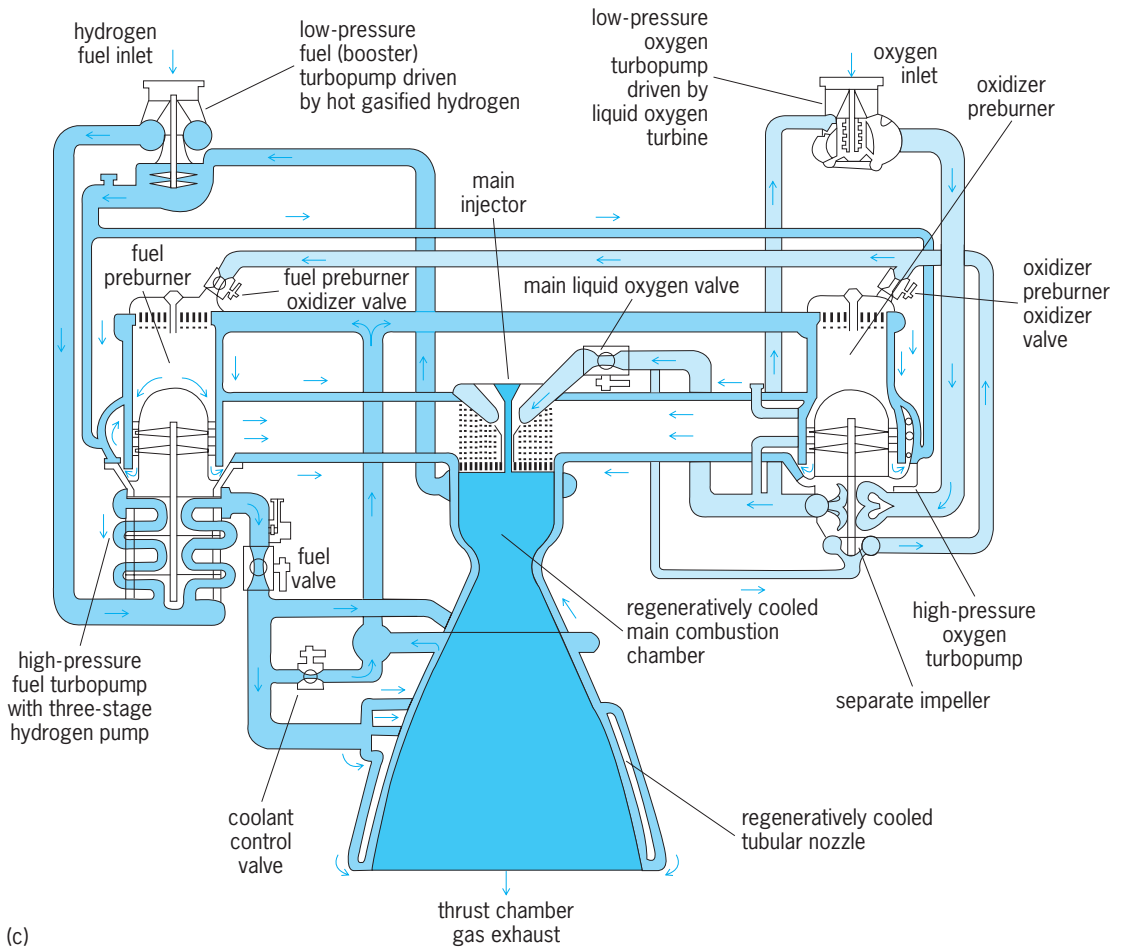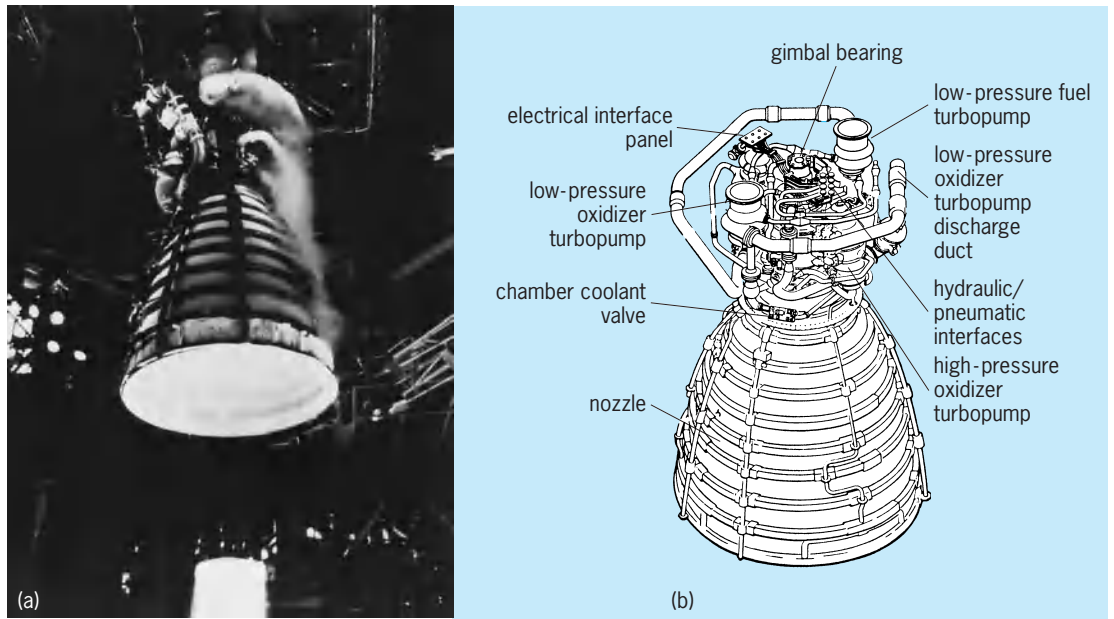**Fig. 8. Liquid-propellant rocket with turbopump feed system.**

Fig. 9. Space shuttle main engine. (*a*) Static firing test. Hydrogen/oxygen flame is invisible except for shock wave at figure bottom (*Rocketdyne Division, Rockwell International*). (*b*) Oblique view. Main injector, chamber, and high-pressure fuel turbopump are behind other hardware (*NASA*). (*c*) Simplified flow diagram. Part of oxygen flow is pressurized with a separate impeller.

flow cycle contributes to the engine's high performance.

Controls and operation. All liquid-propellant rocket engines have controls that accomplish some or all of the following functions: start and shut down rocket operation, restart, maintain a predetermined constant or variable thrust, make emergency shutdowns when safety devices sense a malfunction, fill and drain propellants, and permit functional checkout of critical components without actual rocket engine operation. For example, each space shuttle main engine is controlled by an electronic computer/controller, together with several sensors, valves, and actuators, which provide flight readiness verification, start and shutdown sequencing, closed-loop thrust control and mixture-ratio control, on-board engine monitoring and checkout, response to vehicle commands, data storage, and malfunction detection with triggering of remedial actions.

Starting, stopping, and restarting are generally accomplished by valves in the main propellant lines. Because of the high pressures and the large flows involved, these valves are often operated from separate pilot valves that control the energy going to these main valves. A hydraulic, pneumatic, or electrical system is often used to actuate the principal valves to the thrust chamber and the gas generator and to furnish power to the thrust vector control actuators.

A propellant utilization system ensures that both the fuel and the oxidizer tank are emptied simultaneously. This is achieved by continuously measuring the amount of propellant that remains in both of these tanks and by adjusting throttling valves in both of the main propellant lines leading to the thrust chamber. This minimizes the amount of residual propellant remaining in the tanks, and minimizes the inert mass that remains to be accelerated.

In the operation of liquid-propellant rocket engines it is often desirable to vary the thrust or to throttle the rocket engine. In multiple rocket engines this can be achieved by shutting down some of the engines while letting others continue to operate. In a single-thrust-chamber rocket engine the throttling can be achieved by reducing the chamber pressure and the flow. Throttling is necessary for an accurately reproducible flight trajectory, limiting maximum acceleration, and for lunar descent and landing maneuvers. For example, the lunar landing engine of the Apollo space program has a thrust variation of 10:1.

Solid-propellant rocket motors. In rocket motors the propellant is a solid material that feels like a soft plastic or soap. The solid propellant cake or body is known as the grain. It can have a complex internal geometry and is fully contained inside the solid motor case, to which a supersonic nozzle is attached (**Figs. 10** and **11**).

The propellant contains all the chemicals necessary to maintain combustion. Once ignited, a grain will burn on all exposed surfaces until all the usable propellant is consumed; small unburned residual propellant slivers often remain in the chamber. As
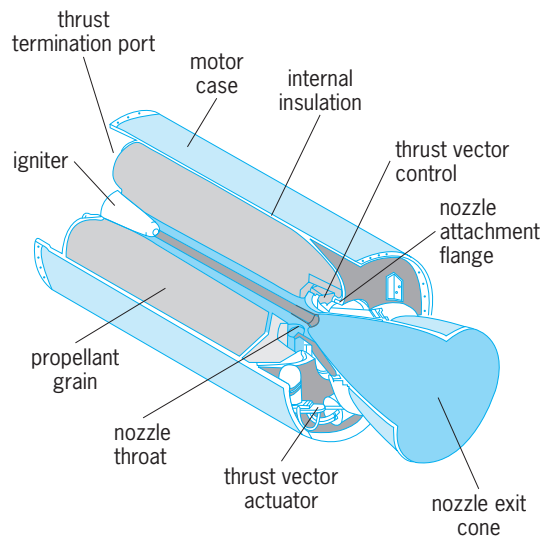

Fig. 10. Solid-propellant rocket motor. (*Hercules, Inc.*)

the grain surface recedes, a chemical reaction converts the solid propellant into hot gas. The hot gas then flows through internal passages within the grain to the nozzle, where it is accelerated to supersonic velocities. A pyrotechnic igniter provides the energy for starting the combustion.

Solid propellants. Of the many different solid propellants, the three most common types are composite propellants, double-base propellants, and composite double-base propellants (**Table 6**).

The largest number of motors use composite propellants. They contain crystalline, finely ground oxidizer particles, such as ammonium perchlorate or ammonium nitrate, dispersed in a fuel matrix of polymer. Several different polymers have been used successfully. Aluminum powder may be added as fuel; in the combustion it is burned into aluminum oxide, a white powdery solid material visible in the flame. There are minor quantities of other additives to improve the fabrication processing and the physical properties, control the burning rate, or limit chemical deterioration. Most composite solid-propellant grains are cast (as a viscous liquid) into the cavity of the case, where they polymerize and solidify. An internal rubberlike thin liner allows good bonding of the solid propellant to the motor case walls.

The next most common class are double-base propellants, which are based on two major ingredients: solid nitrocellulose and a liquid solvent, namely nitroglycerine. Sometimes other organic compounds that have oxidizing radicals attached to them are added, including compounds that are normally explosives. Double-base propellants can be extruded or cast by a solvent process. There are other minor additives, whose purposes are similar to those mentioned above. Aluminum powder may or may not be added as an extra fuel or as a combustion vibration suppressant.

Composite double-base propellants combine ingredients from both the above types. This gives a slightly higher performance, but the propellant is more difficult and hazardous to fabricate or handle.
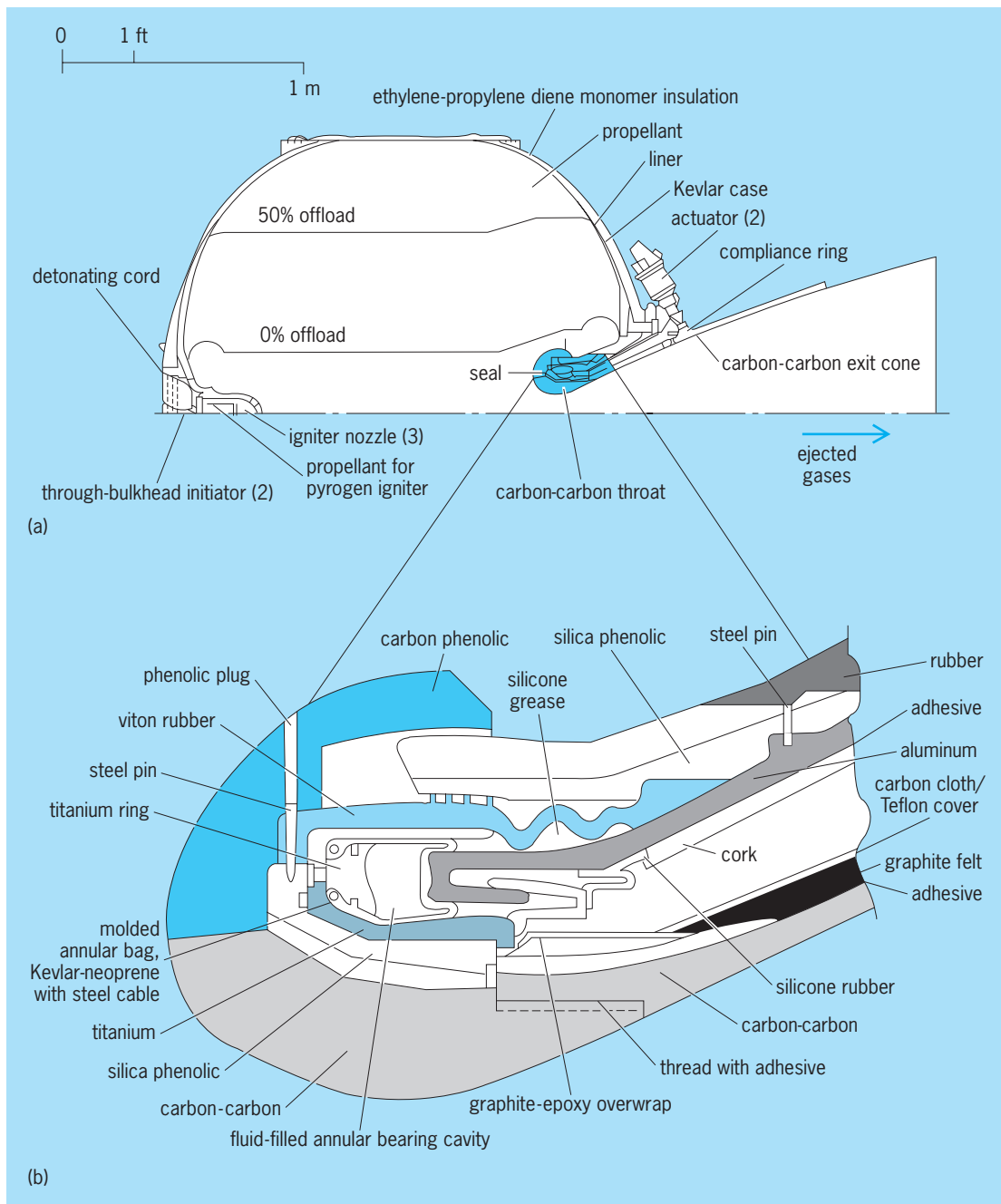
Fig. 11. Simplified cross sections of an upper-stage solid-propellant rocket motor (inertial upper stage; IUS) using an insulated carbon-fiber/carbon-matrix nozzle, an insulated carbon-filament-wound case, a pyrogen igniter, hydroxyl-terminated polybutadiene propellant, forward and aft stress-relieving boots, and a fluid-filled bearing and elastomeric seal assembly in the nozzle to allow 4.5° of thrust-vector deflection. (*a*) Half cross section along length of motor. (*b*) Enlarged cross section of nozzle package assembly. (*United Technologies Chemical Systems*)

All solid propellants, and many of the ingredients that are used in formulating them, are hazardous and have to be processed and handled using rigorous safety procedures. All solid propellants can be inadvertently ignited and thus provide thrust at an inopportune time and make the vehicle mobile, discharge hot flames that will start local fires, or cause areas to be exposed to toxic exhaust products. Under certain unfavorable circumstances, many solid rocket motors can explode.

Many solid propellants undergo an aging process consisting of a series of low-order chemical reactions between the fuel and the oxidizer ingredients (or sometimes between the propellant and the liner) during storage. Deterioration can often be minimized by the addition of stabilizers or inhibitors to the propellant. Nevertheless, the useful life of some solid rocket motors is limited to several years because aging and stress cycling degrade the physical properties or performance of the propellant, and may even make the rocket motor unsafe to operate.

Typical strengths of cast solid propellants vary between 25 and 500 lb/in.^{2} (0.17 and 3.4 MPa). This strength usually decreases as the ambient

TABLE 6. Characteristics of some operational solid propellants

Propellant type*	Specific impulse range, lbf·s/lbm (kN·s/kg)	Flame temperature, °F (°C)	Specific gravity	Metal content, weight %	Stress/strain, (lbf/in.²)/% (MPa/%)		Processing method
					−60 °F (−51 °C)	+150 °F (66 °C)	
DB	220–230 (2.15–2.25)	4100 (2250)	1.61	0	4600/2 (31.7/2)	490/60 (3.38/60)	Extruded
DB/AP-HMX/Al	265–270 (2.60–2.65)	6700 (3700)	1.80	20	2375/3 (16.4/3)	50/33 (0.34/33)	Solvent cast
PU/AP/Al	260–265 (2.55–2.60)	5400–6000 (3000–3300)	1.77	16–20	1170/6 (8.1/6)	75/33 (0.52/33)	Cast
Polymer/AN	180–196 (1.77–1.92)	2600 (1400)	1.55	0	400/2 (2.75/2)	50/15 (0.34/15)	Extruded
CTPB/AP/Al	260–265 (2.55–2.60)	5600–5800 (3100–3200)	1.77	15–17	325/26 (2.25/26)	88/75 (0.61/75)	Cast
HTPB/AP/Al	260–265 (2.55–2.60)	5600–5800 (3100–3200)	1.86	4–17	910/50 (6.3/50)	90/33 (0.62/33)	Cast
PBAA/AP/Al	260–265 (2.55–2.60)	5400–6000 (3000–3300)	1.77	14	500/13 (3.45/13)	41/31 (0.28/31)	Cast

*Al = aluminum; AN = ammonium nitrate; AP = ammonium perchlorate; CTPB = carboxy-terminated polybutadiene; DB = double base; HMX = cyclotetramethylene tetranitramine; HTPB = hydroxy-terminated polybutadiene; PBAA = polybutadiene-acrylic acid polymer; PU = polyurethane. The first propellant is typical of a double-base type, the second of a composite modified double base, and the other five are composite propellants.

temperature rises, and therefore most solid-propellant grains have a tendency to become soft on hot days or brittle on cold days. It is therefore necessary to restrict the temperature range over which certain propellants and grain designs may operate.

The finished, cast-in-place grain must absorb a variety of loads caused by thermal expansion and contraction due to ambient temperature variations, shocks from the sudden pressurization at start, rough handling, vibrations during transport, and flight accelerations. Solid propellant is a viscoelastic material that can suffer damage (cracks) and can become weaker by repeated applications of load at low stress such as in temperature cycling or transport vibrations. If the propellant should crack from an excessive temperature variation or rough handling, additional burning surfaces would be created in the cracks and an unregulated increase in the operating chamber pressure would cause failure of the chamber. It is necessary to handle propellants and solid rocket motors gently, so as not to subject them to sudden shocks by dropping or jarring.

As mentioned above, composite solid-propellant grains are frequently cast and case-bonded to the combustion chamber. Casting requires a fuel-binder mixture that can be safely handled as a viscous liquid at elevated temperatures (150–300°F or 65–150°C) and that cures and solidifies after it is poured into the case. Case bonding places stringent requirements on the mechanical properties of the propellant. About 65–85% of the propellant is dispersed small solids (namely, oxidizer crystals and aluminum powder), the mechanical properties are controlled by the 10–20% of binder and minor additives. This binder must absorb a wide range of stresses and differences in thermal expansion.

Special categories of solid propellants are smokeless propellants (which make detection more difficult) and propellants with nontoxic exhaust gas. Smokeless propellants contain little or no powered aluminum, which ordinarily would form solid particles of aluminum oxide in the exhaust gas. Most composite propellants contain a perchlorate oxidizer, which forms gaseous hydrochloric acid, a very toxic gas. Double-base and ammonium nitrate–based propellants can be both smokeless and nontoxic.

Components. These include the case, nozzle, igniter hardware, air seal, and safety devices. Sometimes there are also instruments, thrust vector control devices, and provisions for emergency shutdown, such as blow-out diaphragms and other devices.

The case confines the propellant grain and the reaction products of combustion. It is generally made of high-strength material, such as alloy steel, or high-strength fiber-reinforced plastics applied by a computer-controlled filament-winding machine. The case is usually protected on its exposed surfaces from excessive heating by internal insulating layers. Fiber-reinforced plastic case structures are usually lighter than comparable steel cases. The case has attachments to the vehicle structure for transmitting the thrust forces and a seal to prevent leakage between the case and the nozzle, which is usually fabricated separately. *See* COMPOSITE MATERIAL.

The nozzle must be protected from excessive heat transfer, from high-velocity hot gases, from erosion by small solid or liquid particles in the gas (such as aluminum oxide), and from chemical reactions with aggressive rocket exhaust products. The highest heat transfer and the most severe erosion occur at the nozzle throat and immediately upstream from there. Special composite materials, called ablative materials, are used for heat protection, such as various types of graphite or reinforced plastics with fibers made of carbon or silica. The development of a new composite material, namely, woven carbon fibers in a carbon matrix, has allowed higher wall temperatures and higher strength at elevated temperatures; it is now used in nozzle throats, nozzle inlets, and exit cones. It is made by carbonizing (heating in a nonoxidizing atmosphere) organic materials, such as rayon or phenolics. The fiber orientation influences the erosion resistance and strength. Multiple layers of different heat-resistant and heat-insulating materials are often

particularly effective. A three-dimensional pattern of fibers created by a process similar to weaving gives the nozzle extra strength. The divergent nozzle section does not experience such severe heat loads or erosion and often can be manufactured with thinner and less complex or less costly ablative materials.

The ablative materials are partly burned, charred, eroded, and progressively vaporized, thus absorbing some heat and forming a relatively cooler gas boundary layer. During this process the throat area can actually increase slightly as wall material is consumed. If the erosion is unsymmetrical, it may cause a slight misalignment of the thrust direction, which can result in vehicle flight control problems.

The nozzles shown in Figs. 10 and 11 have sophisticated thrust-vector control mechanisms. In Fig. 11 the nozzle forces are absorbed by a doughnut-shaped, confined, liquid-filled bag, in which the liquid moves as the nozzle is canted. This is a more detailed version of the flexible nozzle joint shown in Fig. 6. The space shuttle solid rocket boosters have gimbaled nozzles for thrust-vector control, with actuators driven by auxiliary power units and hydraulic pumps.

Most igniters use a pyrotechnic powder charge and an electric initiator to cause ignition of a separate propellant inside the igniter. In some motors there are also thrust termination devices, such as protechnically activated blow-out ports at the forward end of the case, that would allow gases to be ejected in a direction opposite to the nozzle.

There is usually a seal or diaphragm at the nozzle that is burst when the rocket motor starts. It prevents access of moisture and atmospheric oxygen which can cause a deterioriation of the propellant in the case or the igniter.

In many rocket motor cases the opening of the nozzle attachment or the igniter attachment is much smaller than the outside case diameter. During fabrication a mandrel is inserted through this relatively small hole prior to casting to ensure the proper geometry of the perforations and port area. This mandrel or fixture, which can have a complex geometry, is coated with an inert material [such as poly(tetrafluoroethylene)] to allow withdrawal after propellant casting. In some units the mandrel is made of porous, foamlike propellant that stays permanently in the motor and has a very high burning rate.

Grain configuration and burning. The thrust of the rocket is about equal to the product of mass flow and effective exhaust velocity, as shown in Eq. (2). The mass flow rate is equal to the product of the exposed burning area, the rate of burning, and the density of the propellant. By designing the grain so that more or less surface is exposed, the designer can increase or decrease the thrust during rocket motor operation (**Fig. 12**). It is also possible to limit the exposed burning surfaces by applying inhibitors (special inert or slow-burning chemicals) to parts of the surface. Grains that have large quantities of inhibitor applied to their exposed surfaces, called restricted-burning grains, are differentiated from grains that burn on all

exposed surfaces. Holes put into the grain to increase the burning surface are known as perforations.

If the grain design is such that the thrust increases during the operation of the rocket, it is called progressive burning; if thrust decreases during operation, it is regressive burning; if the thrust or the chamber pressure stays approximately constant, it is neutral burning.

In restartable solid rocket units, the combustion chamber is divided into segments or wafers separated by a flame barrier. When all the propellant of the first segment is burned, the flame is stopped at the barrier. The barrier is destroyed upon ignition of the next segment.

The burning rate of solid propellant is measured in a direction normal to the burning surface and is expressed as a linear distance per unit time. Different propellants have burning rates between 0.2 and 0.8 in./s (5 and 20 mm/s), but some burn as slowly as 0.02 in./s (0.5 mm/s) or as fast as 3.5 in./s (90 mm/s). The burning rate usually varies as the chamber pressure to a power between 0.4 and 0.9, and is also a function of the initial ambient temperature. The same grain gives more thrust but a proportionally
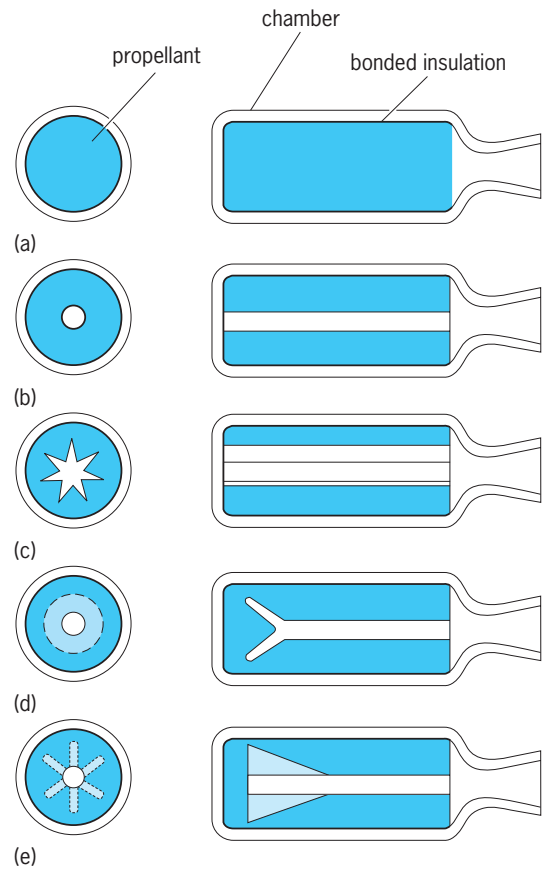

Fig. 12. Several solid-propellant grain configurations. In each part of the figure the diagram at the left is the cross section perpendicular to the length of the motor, and the diagram at the right is the cross section along the length of the motor. All configurations shown are case-bonded. (*a*) End burner, neutral burn. (*b*) Internal tube (end restricted), progressive burn. (*c*) Internal star, neutral burn. (*d*) Cone and cylinder, neutral burn. (*e*) Fins and cylinder, neutral burn.
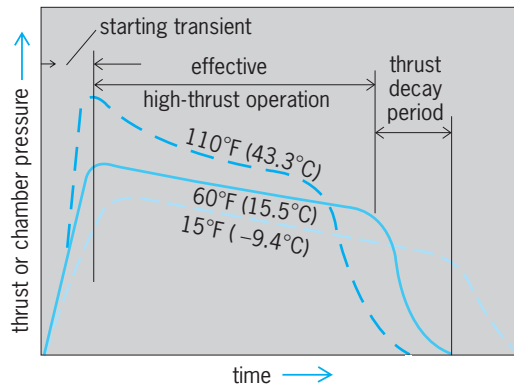
Fig. 13. Thrust-time variation of a typical solid-propellant rocket with slightly regressive burning characteristics. Broken alternative curves show effect of ambient grain temperature. Design of grain and varieties of propellants affect thrust and chamber pressure.

shorter firing duration at higher temperatures (**Fig. 13**). The burning rate also depends on the chemical and physical composition of the propellant; it can be changed by adding special burning-rate accelerators (such as iron compounds) or inhibitors, and is influenced by the erosive action of high-speed hot gases flowing past the internal grain surfaces. There is a maximum pressure above which smooth combustion can no longer be sustained and detonations may occur, and a minimum pressure below which stable, consistently smooth burning does not seem possible.

Comparison of propulsion systems. Both solid- and liquid-propellant systems have several important advantages that must be considered in the selection of a propulsion system. Solid-propellant motors have a simpler design, with few moving parts; are easier to operate, requiring little preflight checkout; have propellants that usually do not leak or spill; often require less overall weight for low-impulse applications; can be throttled or stopped and restarted a few times if preprogrammed (not at random); have propellants that can be stored for 5–20 years; and can be designed for recovery and reuse. Liquid-propellant engines have higher specific impulse; can be randomly throttled; have an accurately controllable thrust profile; can be randomly stopped and restarted, and pulsed many times; have a more controllable cutoff impulse; can be largely checked out just prior to operation; have easier thrust-vector control; can be designed for reuse after field services and checking; have storable propellants that can be kept in the vehicle for a very long time; and are only rarely subject to catastrophic failure.

Both types of propulsion system also have several important disadvantages that must be considered. In the case of solid-propellant motors, propellants can deteriorate in storage; the explosion and fire potential is larger; failure is often catastrophic; the grain displays cumulative damage with temperature cycling or rough handling, and therefore must be handled carefully; factory rework is required before reuse; each restart requires a separate ignition system; the thrust-vector control system may include liquid pro-

pellant; exhaust gases are usually toxic; and the system can be stopped at random only by disabling the rocket, and, once ignited, cannot change thrust or duration. In the case of liquid-propellant engines, the design is relatively complex, with more components; cryogenic propellants cannot be stored; propellant spills can be hazardous, corrosive, and toxic, and can cause fires; nonhypergolic propellants require an ignition system; combustion instability is more difficult to control; more overall weight is required for short-duration low-total-impulse applications; sloshing of liquids in the tank can cause flight stability problems; and, if the tank outlet is uncovered during sloshing, gas in the propellant lines can cause combustion interruption or instabilities.

Hybrid rocket propulsion. A hybrid uses a liquid propellant together with a solid propellant in the same rocket engine (**Fig. 14**). The arrangement of the solid fuel is similar to that of the grain of a solid-propellant rocket; however, no burning takes place directly on the surface of the grain because it contains little or no oxidizer. Instead, the fuel on the grain surface is heated, decomposed, and vaporized, and the vapors burn with the oxidizer some distance away from the surface. The combustion is therefore inefficient. Hybrids usually require a separate ignition device. Hybrids can be restarted and throttled, although at a sacrifice in performance, and they usually lend themselves to safe and rugged construction. Successful experimental target drones have been powered by liquid nitric oxides and a solid fuel consisting of an acrylic plastic with a magnesium compound as additive.

Testing. Because flights of rocket-propelled vehicles are usually fairly expensive and because it is sometimes difficult to obtain sufficient and accurate data from fast-moving flight vehicles, it is accepted practice to test rocket propulsion systems and components extensively on the ground under simulated flight conditions. Components such as an igniter or a turbine are tested separately. Complete engines are tested in static engine test stands; the complete vehicle stage is also tested statically. In the latter two tests the engine and vehicle are adequately secured by suitable structures. Only in flight tests are they allowed to leave the ground.

The static test facilities must be capable of handling and disposing of the hot gases (5000°F or 2760°C) that are expelled from the rocket. For those rockets that must be tested in a vertical position, water-cooled metal flame buckets are used to prevent
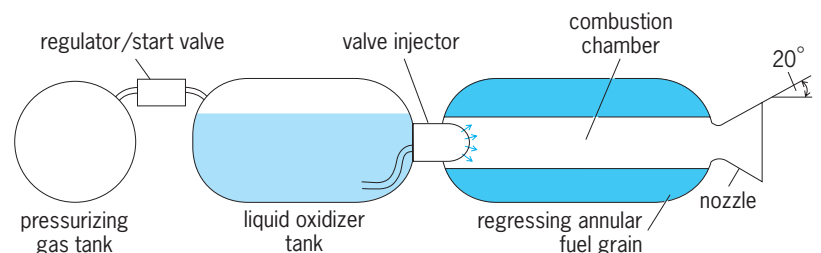

Fig. 14. Hybrid rocket propulsion system.

ground erosion. Propulsion systems which must work in a vacuum, such as space-vehicle engines, are usually also tested in a chamber with an artificial vacuum.

Since rocket propellants are dangerous to handle (they can explode, releasing an energy higher than that of an equivalent weight of high explosive), special precautions and safety provisions must be observed in fabrication, shipping, and testing. These include reinforced buildings of heavy construction that can withstand a blast, remote operation and control, an extensive series of personnel warning devices, special fire-fighting equipment, limitations on the amount of propellant in any one area and on the distance between buildings, the use of sparkproof tools, and rigorous training of personnel.

The instrumentation is usually of the remote indicating or remote recording type. Pressures, temperatures, forces, operating sequence, flows, vibrations, and strains are some of the more common parameters measured. Special instruments are needed for many of these, for example, a pressure recorder with unusually high-frequency response for combustion gas vibrations and optical temperature indicators for high-temperature gas.

Many tests are computer controlled, in part because a computer can monitor parameters and cause a parameter variation or an emergency shutdown faster than a human operator. Most test facilities use computers for data recording, data handling, and instrument calibration corrections. *See* LAUNCH COMPLEX.   George P. Sutton

Bibliography. J. W. Cornelisse et al., *Rocket Propulsion and Spaceflight Dynamics*, Pitman, 1979; A. Davenas (ed.), *Solid Rocket Propulsion Technology*, Pergamon, 1993; National Aeronautics and Space Administration, *National Space Transportation Reference*, vol. 1: *Systems and Facilities*, 1989; G. P. Sutton and O. Biblarz, *Rocket Propulsion Elements*, 7th ed., Wiley, 2001; M. J. L. Turner, *Rocket and Spacecraft Propulsion*, 2d ed., Springer, 2005.

Rocket staging

The use of successive rocket vehicle sections, each having its own propulsion system, such as one or more engines. One way to minimize the mass of large missiles, or space vehicles, is to use multiple stages. The first or initial stage is usually the heaviest and biggest and is often called the booster. It requires the highest thrust and largest rocket propulsion system. The next few stages are successively smaller, have much lower thrusts, and are generally called sustainers. Each stage is a complete vehicle in itself and carries its own propellant (either solid or liquid; both fuel and oxidizer) and its own propulsion system, and has its own tankage and control system.

Once the propellant of a given stage is expended, the dead or inert mass of that stage, including the hardware of a liquid-propellant rocket engine or a solid-propellant rocket motor with flight controls, is no longer useful in contributing additional kinetic en-
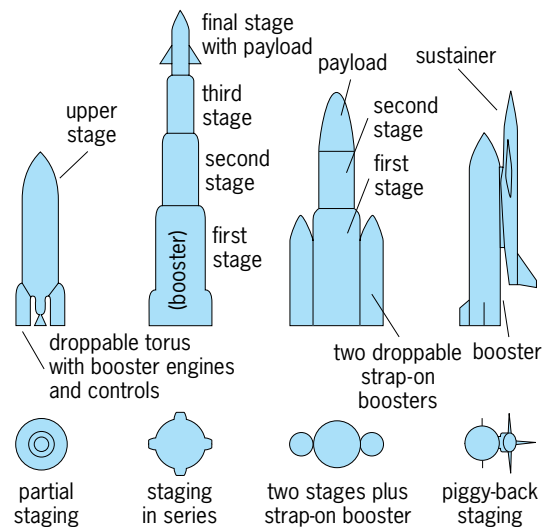
Typical schemes for staging missiles.

ergy to the succeeding stages. By dropping off this useless inert mass, the mass that remains to be accelerated is made smaller; therefore it is possible to accelerate the payload to higher velocity than would be attainable if multiple staging were not used.

Use. Various staging configurations can be used. Multiple-stage arrangements are used in large space-launch vehicles, in some long-range ballistic missiles, various research rockets, and certain antiaircraft and antimissile military weapon systems. It is quite possible to employ different types of power plants, different types of propellants, and entirely different configurations in successive stages of any one multistage vehicle (see **illus.**). Because staging adds separation mechanisms and extra structure and reduces overall reliability, it is impractical to have more than four to six stages in any one vehicle. The Saturn V/Apollo Moon landing and return flight had six stages.

Advantages. By staging a missile or space vehicle, it is possible to improve its performance either by providing more range, more altitude, faster flight, or more payload, or by reducing the gross weight and the size of the vehicle for a given specific mission without diminishing the payload.

For a given initial launch vehicle configuration and mass, the payload can be increased, but at a decrease in flight performance, such as ballistic missile range or spacecraft orbit altitude. Furthermore, the payload also depends on the specific flight trajectory and the vehicle design. The maximum payload that can be carried by any particular rocket-propelled vehicle therefore has to be evaluated individually by using trajectory analyses and design variations. The use of additional stages makes possible an increase in the mass of the payload (such as instruments, passengers, and equipment) or warheads (see **table**). For a fixed payload the range of a ballistic missile can be increased by adding a stage.

By redesigning existing space-launch vehicles and incorporating an additional stage, it is possible to achieve more ambitious space missions. For example, the Titan 34D, a particular two-stage version of

Total payload as a function of the number of stages for specific missions (payload values relative to each other)					
Mission	Missile takeoff weight, lb*	Payload, lb, for number of stages			
		1	2	3	4
1500-mi (2400-km) ballistic missile	50,000	2000	2700	†	†
6000-mi (9700-km) ballistic missile	500,000	‡	6000	9000	†
Moon impact	500,000	‡	100	1000	3000

* 1 lb = 0.45 kg.
† Gain from further staging is small.
‡ This mission cannot be achieved with a single-stage missile and chemical rocket propulsion. The dry weight is so large a percentage of the total weight that the missile will fall short of its target even with zero payload.

the Titan family of space-launch vehicles with two strap-on solid-propellant rocket motor boosters, can carry a payload of about 31,300 lb (14,200 kg) into a 100-nautical-mile-high (185.3-km) circular orbit, and a 15,000-lb (6818-kg) payload into a 5000-nmi (9265-km) circular orbit. This configuration cannot put payloads into high earth orbits, such as a synchronous orbit at 22,700 nmi (42,000 km) or a Moon or planetary mission. However, adding a third stage (in this case powered by an IUS solid-rocket motor) allows the three-stage Titan 34D to carry about 4000 lb (1820 kg) into a synchronous orbit.

Stage separation. Ideally it would be desirable to stop the propulsion unit of the operating stage at the same time that the operation of the propulsion unit of the next stage is initiated. This close timing is not possible. There are inherent difficulties in the separation of the stages caused by drag forces, the lack of gravity in space, the noninstantaneous starting and stopping of rocket engines, and the possible interference of various stages. Often a positive mechanical separation mechanism is included in the design that forces the two stages to separate. *See* ROCKET; ROCKET PROPULSION; SPACECRAFT PROPULSION.   George P. Sutton

Bibliography. M. D. Griffin and J. R. French, *Space Vehicle Design*, 1991.

Rodentia

The mammalian order consisting of the rodents, often known as the gnawing mammals. This is the most diverse group of mammals in the world, consisting of over 2000 species, more than 40% of the known species of mammals on Earth today. Rodents range in size from mice, weighing only a few grams, to the Central American capybara, which is up to 130 cm (4 ft) in length and weighs up to 79 kg (170 lb) [**Fig. 1**]. Rodents have been found in virtually every habitat, from arctic tundra to tropical rainforests, and on every continent except Antarctica. Rodents have adapted to nearly every mode of life, including semiaquatic swimming (beavers and muskrats), gliding ("flying" squirrels), burrowing (gophers and African mole rats), arboreal (dormice and tree squirrels), and hopping (kangaroo rats and jerboas). Nearly all rodents are herbivorous, with a few exceptions that are partially insectivorous to totally omnivorous, such as the domestic rat. Rodents

usually live in small family groups, but the naked mole rats of Africa live in large colonies with a distinct social structure that is similar to that of insect colonies with different classes of members. The great adaptability and rapid evolution and diversity of rodents are mainly due to their short gestation periods (only 3 weeks in some mice) and rapid turnover of generations.

Anatomy. The most diagnostic feature of the Rodentia is the presence of two pair of ever-growing incisors (one pair above and one below) at the front of the jaws. These teeth have enamel only on the front surface, which allows them to wear into a chisellike shape, giving rodents the ability to gnaw. Associated with these unique teeth are a number of other anatomical features that enhance this ability. Behind the incisors is a gap in the jaws where no teeth grow,
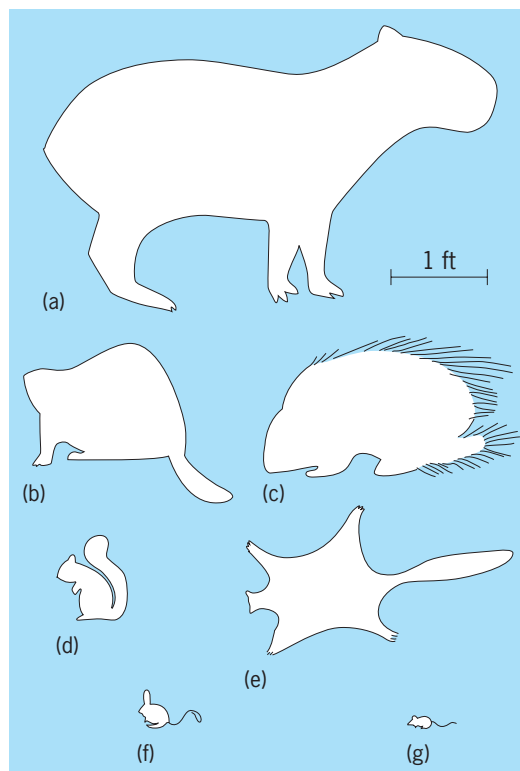

Fig. 1. Silhouettes of modern rodents, showing their wide range in size and shape. (*a*) Central American capybara (*Hydrochoerus*). (*b*) Beaver (*Castor*). (*c*) Old World porcupine (*Hystrix*). (*d*) Tree squirrel (*Sciurus*). (*e*) Asian giant flying squirrel (*Petaurista*). (*f*) Asian jerboa (*Allactaga*). (*g*) Common house mouse (*Mus*).
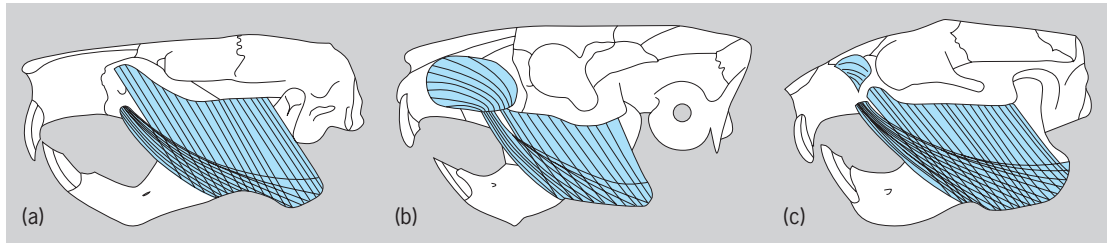
Fig. 2. Skull musculature of three fossil rodents exemplifying the three major groupings of rodents used in the most common classification scheme. (*a*) Sciuromorpha: *Eutypomys thompsoni*, a primitive beaverlike rodent from the Oligocene of North America. (*b*) Hystricomorpha: *Neoreomys australis*, a Miocene South American rodent. (*c*) Myomorpha: *Eumys elegans*, an early "field mouse" from the Oligocene of North America. Skulls are not to scale. (*Redrawn from A. E. Wood, Symposium of the Zoological Society of London, 34:21–60, 1974*)

called a diastema. The diastema of the upper jaw is longer than that of the lower jaw, which allows rodents to engage their gnawing incisors while their chewing teeth (molars and premolars) are not being used. The reverse is also true; rodents can use their chewing teeth (also called cheek teeth) while their incisors are disengaged.

The entire skull structure of rodents is designed to accommodate this task of separating the use of the different types of teeth. Rodent skulls have long snouts; the articulation of the lower jaw with the skull is oriented front to back rather than sideways as in other mammals; the jaw muscles (masseter complex) are extended well forward into the snout; and the number of cheek teeth is less than in most other mammals—all features unique to rodents.

The cheek teeth of rodents vary greatly, from very low crowned teeth only a few millimeters high (squirrels or hamsters) to cheek teeth that (like the incisors) continue to grow throughout the life of the animal (voles, beavers, and chinchillas). In the species with low-crowned cheek teeth, enamel covers the entire tooth surface until it is worn away in extreme old age. Those species with very high crowned (or ever-growing) cheek teeth usually have teeth that are surrounded by enamel on the sides only and have a chewing surface made mostly of dentine. The maximum number of cheek teeth in rodents is two premolars and three molars (on one side) on the upper jaw and one premolar and three molars in the lower jaw. The maximum dental formula for rodents is 1-0-2-3/1-0-1-3, where first is the number of incisors, second the number of canines, third the number of premolars, and fourth the number of molars. This dental formula refers to only one side of the jaw, so the maximum number of teeth in any rodent is 22. Some species of African naked mole rats have only two molars and no premolars, reducing their total number of teeth to 12. *See* DENTITION.

The bodies and skeletons of rodents vary greatly, as do their habitats and modes of life. Those that swim have webbed feet and even paddlelike tails. Those that glide have a special flap of skin that attaches the front and hind limbs together for a sort of wing. Those that are burrowers develop enlarged forearms and hands for digging, and those that hop have elongated hindlimbs and smaller forelimbs. Other features are unique to certain families of rodents, such

as the fur-lined cheek pouches for storing food in the pocket mice and gophers. The naked mole rats of Africa have virtually no fur at all.

Classification. The classification of rodents has always been difficult because of the great diversity of both Recent and fossil species. Traditionally, there are two ways that rodents have been divided: into three major groups based on the structure of the attachment of the jaw muscle (masseter) on the skull (Sciuromorpha, Hystricomorpha, Myomorpha) [**Fig. 2**]; or into two groups based on the structure of the lower jaw (Sciurognathi, Hystricognathi). The difficulty in using these groups (usually considered suborders or infraorders) is that the distinctive adaptations of one group of rodents are also present in others, derived in completely separate ways. Thus, any classification of rodents has always had a number of families that cannot be designated to any one of the suborders. In some cases, new suborders of rodents are made to accommodate the families that do not fall into the traditional categories, but these groups usually consist of a number of unrelated families and are referred to as "waste basket" groups. Still, the most commonly used classification is that of the three major groups based on the musculature of the skull.

Fossil history. The earliest record of rodents is from the late Paleocene of both North America and Inner Mongolia. These earliest examples already have all of the distinguishing features of rodents but belong to primitive extinct families. By the early Eocene, rodents are also known from Europe, India, and Africa. In the early Oligocene, the first rodents appear in South America. Rodents do not occur in Australia until the Pleistocene (**Fig. 3**).

The pattern of rodent evolution is characterized by the development of endemic lineages on each of the continents, with numerous intercontinent emigration events involving the more successful groups. There are more than a dozen extinct families of rodents along with the 30 or more recognized extant families. The Eocene is dominated by species of extinct families in North America and Europe; the still-extant family of ctenodactyloid rodents (gundis) in eastern Asia and India; and an extinct family, the Phyomyidae, in Africa (believed to be ancestral to later families that originated in Africa). By the end of the Eocene, the earliest a members of many modern
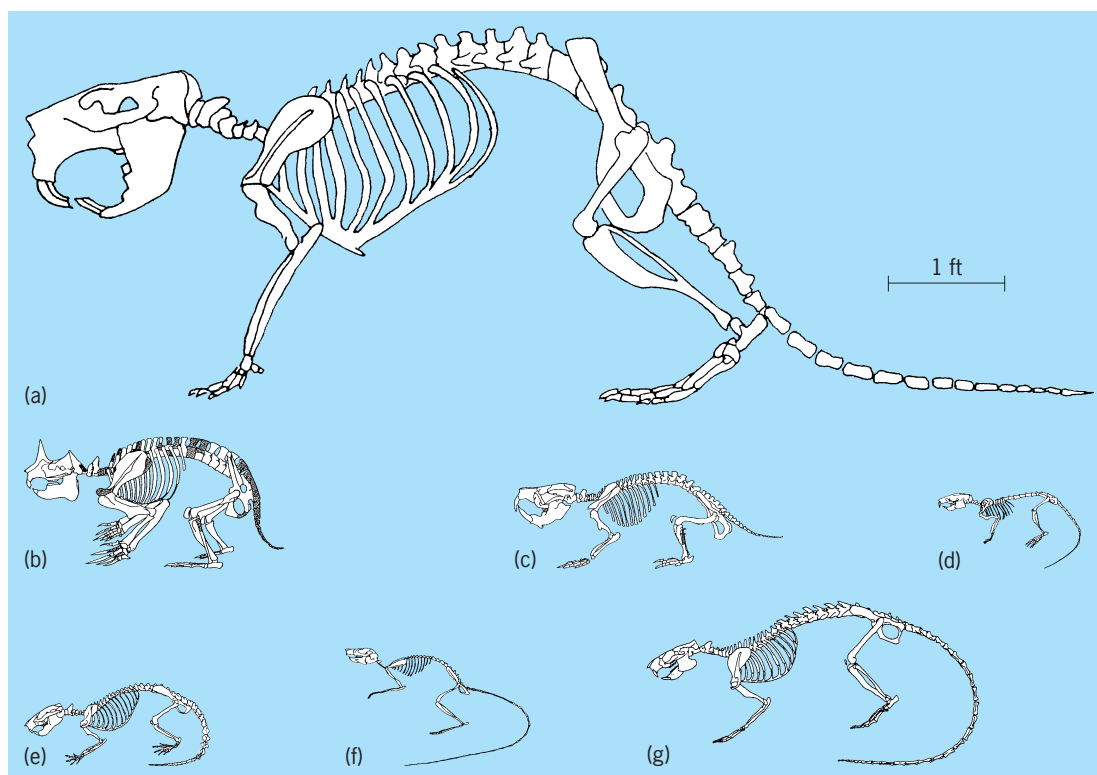
Fig. 3. Reconstructions of the skeletons of fossil rodents from North America. (*a*) Pleistocene giant beaver *Castoroides ohioensis*. (*b*) Late Miocene "horned" rodent *Epigaulus hatcheri*. (*c*) Late Oligocene to early Miocene burrowing beaver *Palaeocastor fossor*. (*d*) Late Oligocene to early Miocene "field mouse" *Geringia mcgregori*. (*e*) Late Miocene to early Oligocene primitive rodent *Ischyromys typus*. (*f*) Late Eocene "jumping" rodent *Protoptychus hatcheri*. (*g*) Early Eocene primitive rodent *Paramys delicatus*. (*Part a redrawn from A. S. Romer, Vertebrate Paleontology, University of Chicago Press, 1966; b from J. W. Gidley, A new horned rodent from the Miocene of Kansas, Proc. U.S. Nat. Mus. 32:627–636, 1907; c–g from W. W. Korth, The Tertiary Record of Rodents in North America, Plenum Press, New York, 1994*)

families had appeared: in North America, the ancestors of pocket mice, gophers, and jumping mice; in Europe, the earliest dormice; in Asia, the cricetid rodents (field mice, hamsters); and in Africa, still geographically isolated, the endemic species of phiomyids and the extant family Anomaluridae (scaly tailed flying squirrels).

At the beginning of the Oligocene, there was a great turnover in the rodent faunas of the world, due to both migrations and evolution of new families. The cricetids from Asia spread to both North America and Europe. The first appearance of squirrels was nearly simultaneous in North America and Europe. Similarly, the earliest beavers occurred in both North America and Asia. In Asia, an extinct family of very specialized burrowing rodents that is restricted to that continent, the Tsagonomyidae, appeared. The earliest jerboas also appeared in Asia at this time. By the end of the Oligocene, most of the Eocene families of primitive rodents had become extinct. The Oligocene is significant in South America because the first record of rodents on that continent dates to that period; they are most similar to those from the Oligocene of Africa. This group of unique rodents in South America diversified without interaction with rodents from any other continent until the very end of the Miocene, when there was a great faunal interchange with North America due to the form-

ing of the isthmus of Panama joining the two continents.

During the Miocene, all extant families of rodents appeared. A notable event was the first occurrence of the Muridae (true rats and mice) in central Asia. This group rapidly diversified and invaded Africa and Europe. Specialized members of the Muridae (voles, lemmings, and muskrats) developed near the end of the Miocene and appeared in both Eurasia and North America. The connection of North and South America allowed the spread of squirrels and cricetid rodents into South America, and the northward expansion of the endemic South American rodents (porcupines and nutria). In Africa, the endemic modern families appeared (springhares, naked mole rats) at the time of the extinction of the primitive African rodents (phiomyids) and the introduction of squirrels, Old World porcupines, and cricetids. During the Miocene, a family of unusual rodents, the Mylagaulidae, appeared and diversified in North America. These rodents are well adapted for burrowing, and some species developed horns on their snouts much like that of a rhinoceros. In Asia, the ctenodactyloid rodents, which had been dominant since the early Eocene, migrated to Africa and became extinct in Asia.

By the Pleistocene, nearly all the extant genera of rodents around the world had appeared in the

fossil record, producing a rodent fauna much like that of today. A notable exception is the gigantic beaver *Castoroides* from North America, which was up to 2.3 m (7 ft) long and weighed hundreds of pounds. These large creatures disappeared by the end of the Pleistocene, along with the other large mammals that were part of the late Pleistocene extinctions (mammoths, giant rhinos, giant bears). *See* EOCENE; MIOCENE; OLIGOCENE; PALEOCENE; PLEISTOCENE.

Economic and historical importance. A few rodent species have been trapped or farmed for their fur. Most notably, the beaver (*Castor*) was trapped in North America, which led to European exploration of the continent. Currently, muskrats (*Ondatra*), the South American chinchillas (*Chinchilla*), and nutria (*Myocastor*) are the main sources of rodent furs. Domestic rats and mice (*Rattus* and *Mus*) are almost universally used as laboratory subjects for medical research. Guinea pigs (*Cavia*) were domesticated in South America as a food source a few thousand years ago.

In the twentieth century a number of rodents were domesticated as pets: gerbils, hamsters, rats, mice (all Old World members of the mouse family), guinea pigs (from South America), and even prairie dogs (a member of the squirrel family from western North America).

Rodent damage to harvested crops is high around the world. Domestic rats consume large amounts of stored food. The amount of stored grains destroyed by these rodents has been estimated at as much as $1 billion per year in the United States alone. Other rodents have been responsible for the destruction of unharvested crops and grazing lands around the world.

Rodents have also been the source of the spread of disease. In the fourteenth century, tens of millions of people in Europe and Asia died from the bubonic plague, which was transmitted by the fleas that infested rats. Other diseases, such as Lyme disease (Lyme borreliosis) and Rocky Mountain spotted fever, are spread similarly, not by rodents directly but by the parasites that they carry. Hemorrhagic fever viruses (such as hantaviruses) are also spread by rodents of many species worldwide. The diseases caused by these viruses, spread in air that carries microscopic particles from rodent secretions or excretions, are severe and often result in death.   William W. Korth

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, H. Freeman, New York, 1988; J. E. Childs, Special feature: Zoonoses, *J. Mammal.*, 76:663, 1995; W. W. Korth, *The Tertiary Record of Rodents in North America*, Plenum Press, New York, 1994; M. C. McKenna and S. K. Bell, *Classification of Mammals above the Species Level*, Columbia University Press, New York, 1997; R. M. Nowak, *Walker's Mammal Species of the World*, Johns Hopkins University Press, Baltimore, 1991; D. E. Savage and D. E. Russell, *Mammalian Paleofaunas of the World*, Addison-Wesley, London, 1983; D. E. Wilson and D. M. Reeder, *Mammal Species of the World*, Smithsonian Institution Press, Washington, DC, 1993.

Rodenticide

A toxic chemical that is used to kill pest rodents and sometimes other pest mammals, including moles, rabbits, and hares. Rodenticides are the least widely used of the pesticides, which also include insecticides, fungicides, and herbicides.

Most rodenticides are used to control rats and house mice, the common rodents that infest urban, suburban, and rural buildings and feed on a wide variety of edible commodities. Their droppings and urine contaminate several times more food than is actually consumed. Rats and mice cause destruction by gnawing on buildings and valuable items such as books, paintings and other art works, and rare manuscripts. They also can transmit a number of serious diseases to people and domestic animals. It is a common practice to make buildings rodent-proof in an effort to reduce the need for rodenticides.

Certain field and forest rodents, such as pocket gophers, ground squirrels, prairie dogs, and voles, become pests when they conflict with crop production, reforestation, and environmental quality. When nonlethal methods either fail to resolve the problem or are impractical, these animals must be controlled with rodenticides.

Use formulations. Rodenticides are generally combined with some rodent-preferred food item such as grain (corn, wheat, oats) or a combination of grains in low yet effective amounts. Bait formulations may be in pellet forms or incorporated in paraffin blocks of varying sizes. Loose-grain kernels and ground-meal baits are also available. As a safeguard against accidental ingestion by nontarget species, baits are placed either where they are inaccessible to children, domestic animals, or wildlife, or within tamper-resistant bait boxes designed to exclude all but rodent-size animals. Some exceptions exist in agricultural situations. To prevent ingestion by children, a bitter substance, such as denatonium benzoate, is sometimes added at concentrations that do not adversely affect rodent consumption.

Types. As a group, anticoagulant rodenticides dominate the market and are sold under a wide variety of trade names. In order of their development, they are warfarin, pindone (Pival), diphacinone, and chlorophacinone. When small amounts of these anticoagulants are consumed over several days, death results from internal bleeding. The newer, second-generation anticoagulant rodenticides, such as brodifacoum, bromadiolone, and difethialone, were developed to counteract the growing genetic resistance in rats and house mice to the earlier anticoagulants, especially warfarin. The second-generation compounds are more potent and capable of being lethal following a single night's feeding, although death is generally delayed by several days. Vitamin K_1 is the antidote for all anticoagulant rodenticides, if administered soon enough after ingestion.

Rodenticides that do not belong to the anticoagulant group include zinc phosphide, bromethalin, and cholecalciferol (vitamin D_3). The feeding and lethal characteristics differ among them. Acutely

toxic strychnine baits are also available but are restricted to underground application, primarily for pocket gophers and moles.

A number of toxic agents have been used in the past as rodenticides but are not currently registered for use in the United States. These include arsenic, thallium sulfate, antu, red squill, sodium fluoroacetate (compound 1080), fluoroacetamide (1081), and cyanide. The reasons for their discontinuation vary, but environmental concerns and the availability of more efficacious rodenticides are the most significant.

Several lethal fumigants or materials that produce poisonous gases are used to kill rodents in burrows and within other confined areas such as unoccupied railway cars or buildings. Lethal fumigants include aluminum phosphide, carbon dioxide, chloropicrin, and smoke or gas cartridges, which are ignited to produce carbon monoxide and other asphyxiating gases.

Regulations. Because of their high toxicity, rodenticides are inherently hazardous to people, domestic animals, and wildlife. They are highly regulated, as are certain other types of pesticides. Rodenticide products must be proven effective and most demonstrate that they can be used without presenting unacceptable hazards to nontarget species or to the environment. Some rodenticides can be purchased and used only by trained certified or licensed pest control operators, while others with a greater safety margin can be used by the general public. Rodenticide products, prior to marketing, are registered under the comprehensive rigid criteria of the U.S. Environmental Protection Agency (EPA). States may also require pesticide registration and may demand safety and other additional pesticide data beyond that required by the EPA. Certain rodenticides may be registered with EPA approval for specific uses only in a particular state. All rodenticides must be used in accordance with the label directions and may be prohibited where they may jeopardize certain endangered species. *See* PESTICIDE; RODENTIA.

Rex E. Marsh

Bibliography. L. L. Bergeson (ed.), *FIFRA: Federal Insecticide, Fungicide and Rodenticide Act* (*Basic Practice Series*), 2000.

Roentgenium

The nineteenth of the synthetic transuranium elements, symbol Rg, atomic number 111. Discovered in late 1994, roentgenium should be a homolog of the elements gold, silver, and copper. It is the ninth element in the 6*d* shell. *See* COPPER; GOLD; HALF-LIFE; RADIOISOTOPE; SILVER.

The element was discovered on December 17, 1994, at GSI (Gesellschaft für Schwerionenforschung), Darmstadt, Germany, by detection of the isotope ^{272}Rg (the isotope of roentgenium with mass number 272), 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 meitnerium-268, bohrium-264, dubnium-260,


and lawrencium-256 allowed identification from the known decay properties of ^{260}Db and ^{256}Lr. In the decay chain (see **illustration**), 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 new isotopes meitnerium-268 and bohrium-264 with their half-lives of 70 ms and 0.4 s, respectively, show a trend toward longer half-lives than those of the previously known isotopes of these elements. *See* ALPHA PARTICLES; NEUTRON; NUCLEAR REACTION.

The discovery of roentgenium followed the discovery of ^{269}Ds within 5 weeks. Both elements were produced by the same techniques. The production
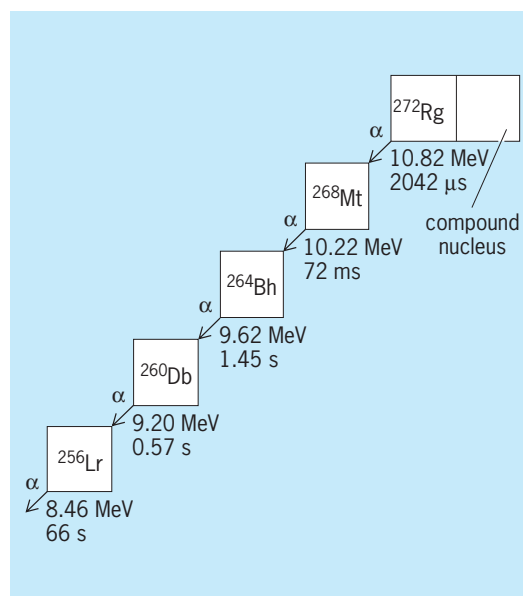
Sequence of decay chains that document the discovery of roentgenium. Numbers below boxes are alpha energies and correlation times. Roentgenium is produced in the reaction ^{64}Ni + ^{209}Bi → ^{272}Rg + 1n.

cross section decreases by a factor of 4 compared to the even darmstadtium. This factor is smaller than the corresponding factor of 9 between the cross sections to produce elements meitnerium and hassium. *See* DARMSTADTIUM; HASSIUM; NUCLEAR STRUCTURE.

The group of RIKEN, Japan, in 2004 observed a single decay chain of $^{278}113$, which shows as first descendant the new isotope ^{274}Rg.   Peter Armbruster

Bibliography.   S. Hofmann et al., The new element 111, *Z. Phys. A*, 350:281–282, 1995.

Roll mill

A series of rolls operating at different speeds. Roll mills are used to grind paint or to mill flour. In paint grinding, a paste is fed between two low-speed rolls running toward each other at different speeds. Because the next roll in the mill is turning faster, it develops shear in the paste and draws the paste through the mill. The film is scraped from the last high-speed roll. For grinding flour, rolls are operated in pairs, rolls in each pair running toward each other at different speeds. Grooved rolls crush the grain; smooth rolls mill the flour to the desired fineness. The term roller mill is applied to a ring-roll mill. *See* GRINDING MILL.   Ralph M. Hardgrove

Rolling contact

Contact between bodies such that the relative velocity of the two contacting surfaces at the point of contact is zero. Common applications of rolling contact are the friction gearing of phonograph turntables, speed changers, and wheels on roadways. Rolling contact mechanisms are, generally speaking, a special variety of cam mechanisms. An understanding of rolling contact is essential in the study of antifriction bearings. The concepts are also useful in the study of the behavior of toothed gearing. *See* ANTIFRICTION BEARING; GEAR; MECHANISM.

Pure rolling contact can exist between two cylinders rotating about their centers, with either external or internal contact. Two friction disks (**Fig. 1**) have external rolling contact if no slipping occurs between them. The rotational speeds of the disks are then inversely proportional to their radii.
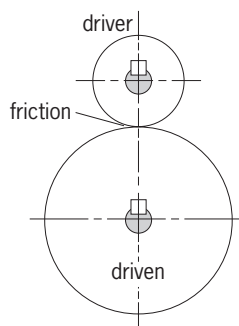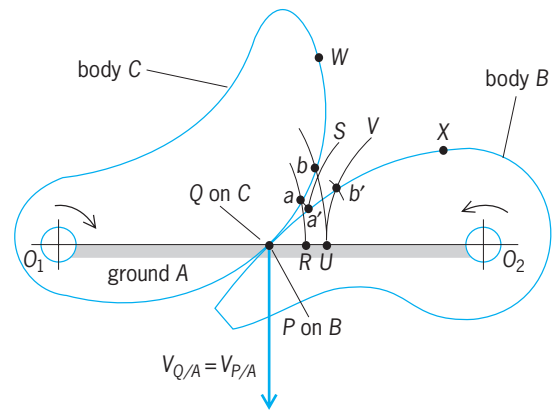

Fig. 1. Rolling friction disks.

Fig. 2. Pure rolling between bodies in contact. (*After H. A. Rothbart, Cams, John Wiley and Sons, 1956*)

The general case of pure rolling between contacting bodies is illustrated in **Fig. 2**. Particle Q on body C and particle P on body B coincide at the moment. The velocities of these particles relative to ground frame A must be the same if there is to be no slip. It can be shown, using the theory of kinematics of mechanisms, that for two bodies to have rolling contact at an instant the point of contact must lie on the line joining the two centers of rotation. *See* MECHANISM; VELOCITY ANALYSIS.

Figure 2 shows that if the contacting particles Q and P were anywhere except on the line of centers O_1O_2 their velocities would necessarily be different, for they would have different directions. Because the centers of rotation are fixed, the sum of the distances from the centers of rotation to the point of contact must equal the center distance. If, in Fig. 2, W and X are two points that ultimately will come into contact, then the bodies must be designed to satisfy Eq. (1).

$$O_1W + O_2X = O_1O_2 \qquad (1)$$

Finally, because there is to be no slip between the contacting surfaces, the length of the arc of action on each body must be the same. In Fig. 2, for example, arc length QW must equal arc length PX.

At an instant the angular velocities of the two links are inversely proportional to the distances from the centers of rotation to the point of contact, stated mathematically as Eq. (2).

$$\frac{\omega_C}{\omega_B} = \frac{O_2P}{O_1Q} \qquad (2)$$

A graphical technique for designing the shape of body B so that it will have pure rolling contact with given body C is illustrated in Fig. 2. Select a sequence of points along the profile of body C such as a, b, and so on. Then the corresponding points a', b', and so on of body B must be found. Revolve point a about O_1 to the line of centers, giving point R. With O_2 as center and O_2R as radius, draw an arc RS of indefinite length. Because arc length Pa' must equal arc length Qa, the location of a' can be found approximately by using chord Qa as a compass distance and swinging an arc from R to intersect arc RS at a'. The same

construction is used for finding b'. By repeating this procedure for a number of closely spaced points, the profile of body B can be determined.

Pure rolling occurs between the bodies with profiles that are the basic plane curves, that is, logarithmic spirals, ellipses, parabolas, and hyperbolas. To achieve pure rolling between these curves, the centers of rotation must be at the foci of the curves; the curves must be properly proportioned and positioned with respect to one another (**Fig. 3**).

Logarithmic spirals of equal obliquity are shown pivoting about their foci in Fig. 3a and b. In the first instance the links turn in opposite directions; in the second they turn in the same direction. The latter arrangement is the more compact. A logarithmic spiral rolling against a straight-sided follower is shown in Fig. 3c. Here the follower has a reciprocating motion. For the pair of ellipses shown in Fig. 3c, pure rolling contact can occur for complete rotations of both members. Rolling parabolas, as in Fig. 3e, provide a reciprocating output motion. Where space is limited, the equal hyperbolas of Fig. 3f are an especially effective choice.

A difficulty with most of these arrangements is that only partial rotations are possible. The lobe wheels shown in Fig. 3g, on the other hand, make

(a)   (b)

(c)

Fig. 4. Friction gears. (a) Cones. (b) Variable speed. (c) Mechanical integrator. (International Textbook Co.)

it possible to have continuous action. The lobes are made up of sectors of logarithmic spirals or ellipses. Friction gearing consists mainly of rolling cylinders, cones, and disks. The limited areas of surface contact permit only small quantities of power to be transmitted. Slippage can be often reduced by the use of spring-loaded bearings. Friction devices include friction cones (**Fig. 4a**); Evan's friction cones, which permit variations in output speed by shifting the belt position (Fig. 4a); and the Bush wheel and plate arrangement (Fig. 4c) used as an integrator for the original Vannevar Bush analog computer. The Bush arrangement permits not only variations in output speed but also reversal of the direction of the output shaft. *See* ANALOG COMPUTER; INTEGRATION.

<div align="right">John R. Zimmerman</div>

Bibliography. F. T. Barwell, *Bearing Systems*: *Principles and Practice*, 1979; B. Bolt, *Mathematics Meets Technology*, 1990; C.-H. Chiang, *Kinematics of Spherical Mechanisms*, 1988; V. M. Faires and R. M. Keown, *Mechanism*, 5th ed., 1960, reprint 1980; H. B. Kepler, *Basic Graphical Kinematics*, 2d ed., 1973; Society of Automotive Engineers, *Bearing Technology: Analysis, Development and Testing*, 1985; C. H. Suh and C. W. Radcliffe, *Kinematics and Mechanisms Design*, 1978, reprint 1983.

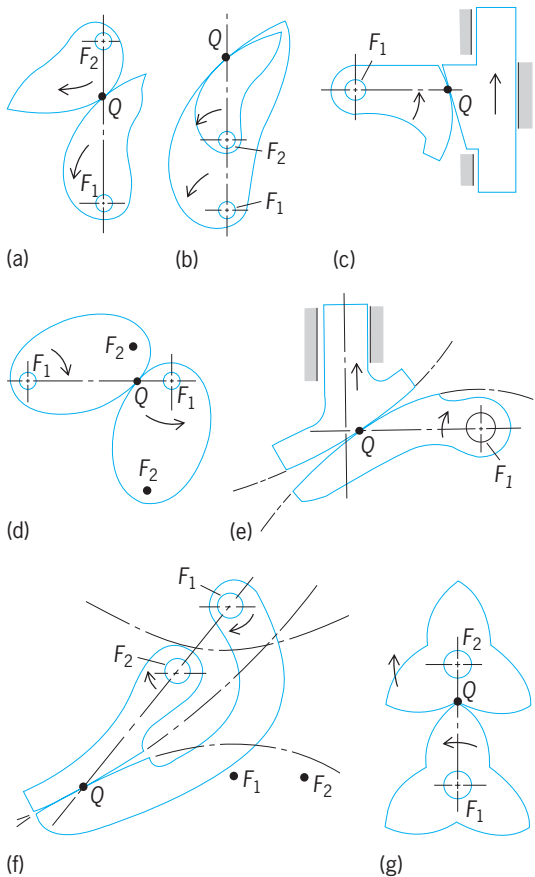(a)   (b)   (c)

(d)   (e)

(f)   (g)

Fig. 3. Pure rolling between bodies in contact. (a) Externally rolling logarithmic spirals. (b) Internally rolling logarithmic spirals. (c) Logarithmic spiral and straight-sided follower. (d) Ellipses. (e) Equal parabolas. (f) Equal hyperbolas. (g) Lobe wheels contoured either from sectors of ellipses or logarithmic spirals. (After H. A. Rothbart, Cams, John Wiley and Sons, 1956)
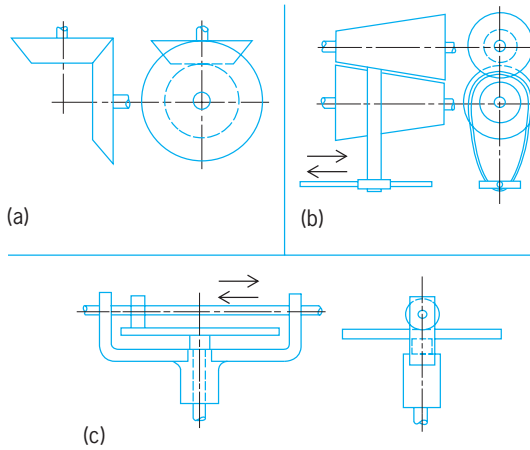
Roof construction

An assemblage to provide cover for homes, buildings, and commercial, industrial, and recreational areas. Roofs are constructed in different forms and shapes with various materials. A properly designed and constructed roof protects the structure beneath it from exterior weather conditions, provides structural support for superimposed loads, provides diaphragm strength to maintain the shape of the structure below, suppresses fire spread, and meets desired esthetic criteria. Roofs have been constructed in many shapes by using structural assemblies and materials including beams, arches, cut stone slabs, cast-iron segments, masonry or terra-cotta tiles, and thatched grasses.

Modern roof construction, evolving from older technologies, usually consists of an outer roofing assembly that is attached atop a deck or sheathing surface, which in turn is supported by a primary framework such as a series of beams, trusses, or arches. A hung ceiling may be placed below the roof structure to give a closed or finished look and provide a space for modern utilities such as heating ducts and electrical and communication lines. The shape of the roof and type of roof construction are usually determined by, and consistent with, the materials and deck of the primary structure underneath. *See* ARCH; BEAM; TRUSS.

Roof shapes include flat; hipped, where two sloping deck surfaces intersect in a line, the ridge or hip; pyramidal, which involves three of more sloping planes; domed, or other three-dimensional-surface, such as spherical, parabolic, or hyperbolic, shells; and tentlike, which are suspended fabric or membrane surfaces. Some famous old surviving roofs include the steep buttressed Gothic church roofs such as Notre Dame Cathedral in Paris; large span domes such as St. Peter's Church in Rome; the dome over the U.S Capitol in Washington, D.C.; and the mosque-like dome such as that over the Taj Mahal in Agra, India.

Roof assembly. A roof assembly (**Fig. 1**) is a series of layers of different materials placed on and attached to the roof deck. Each type of roof assembly—related to protection against water entry from rain, snow, or ice; and insulation for temperature change, fire propagation, wind uplift, and moisture migration—has its own design requirements and methods of construction and attachment.

Materials. Roofing materials vary, and are selected according to design requirements to keep water out, to provide adequate strength against wind uplift, to sustain light impact loads, and to reduce wear if the roof is to sustain human activity, and according to availability of local materials.

Shingles. Usually used for hipped or pyramidal roofs, these are small water shedding pieces made of wood, slate, asphaltic felts, and other materials. The shingles are overlapped downslope to shed water, and they are usually nailed or stapled onto the roof deck made of membrane-coated wood. A variation of the shingle used in warmer climates is either flat or half-round ceramic or concrete, known as Spanish tiles.

Built-up roofing. Built-up roofing is a traditional type of roof cover; it is usually used on flat or low-slope roofs. It is constructed of several layers of asphalt or tar-saturated fabric sheets, reinforced with plastic or fiberglass, with each sheet being called a ply. Each ply is bonded to the surface below by hot or cold applications of asphalt-based adhesives in order to form a multi-ply built-up roof. The top ply may have granules of sand or stone embedded into it, making it a mineral-surfaced ply that can provide some resistance to wear and tear from human activity. Flat roofing may be ballasted by being covered with loose stone or gravel of $1/2$–3 in. (15–75 mm) diameter to depths of 4 in. (100 mm) to resist wind uplift forces.

Single-ply roofing. Single-ply roofing utilizes large factory-fabricated sheets of flexible plastic or rubberlike materials, such as neoprene rubber, chlorosulfonated polyethylene elastomer, or poly(vinyl chloride), about 0.03–0.50 in. (0.76–12.7 mm) thick.
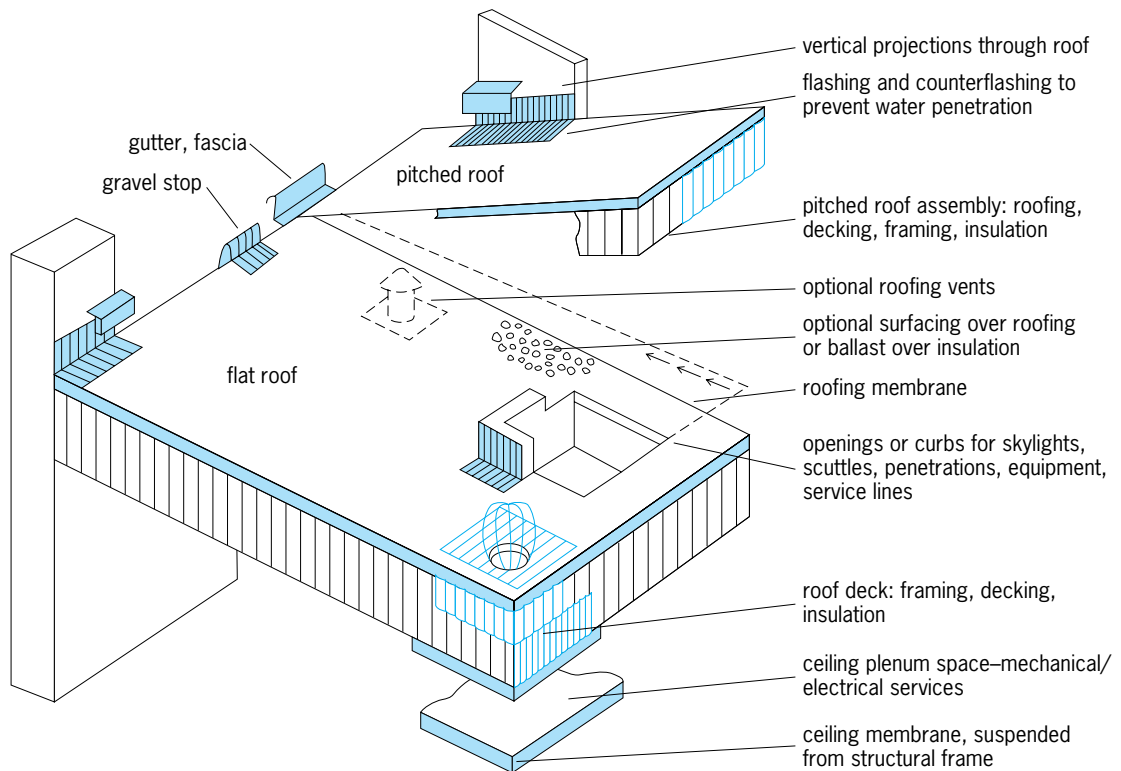

Fig. 1. Common components of roof assembly. (*After Sweets Catalog, The McGraw-Hill Companies, 1993*)

If one factory-produced sheet does not cover the whole roof, additional single-ply sheets are laid down, with adjacent edges or ends lapped, glued, and sealed to form watertight seams between adjoining sheets. Some single-ply systems are loose-laid and held down by gravel ballast or a waterproof batten system, where thin structural planks are spaced atop the single ply and fastened to the roof structural deck below, or are adhered to the insulation and roof deck below.

Other occasionally used roofing materials include corrugated, flat, or standing seam sheets of copper, terne plate metal, painted or galvanized steel, or aluminum. In a standing seam configuration, the two metal pieces are joined above the roof surface (**Fig. 2**).

Vapor barriers. Air, which contains moisture, generally moves from a warm area to a cooler one, and the moisture within condenses. Condensation can cause water corrosion or other damage to a roof structure, and reduce the effectiveness of insulation. Vapor barriers are a ply or membrane of plastic film, aluminum foil, foil-laminated paper, or other prepared materials usually placed on the warm side of insulation to prevent or reduce the passage of moisture, and subsequent condensation.

Venting sheets. Vapor barriers are not totally effective. In areas where large amounts of water vapor may be encountered, a special type of roofing ply is installed above the insulation to wick or expedite the drying out of any moisture that reaches the roofing assembly.

Insulation. Usually composed of materials such as fiberglass, urethane, or styrofoam, this is placed on or under modern roof decks in order to reduce energy losses for both winter and summer. Some insulation is fabricated with a vapor barrier on one side for rapid application.

Edges. The outer edges of roofing are embedded, attached, and sealed to the edge structure of the roof
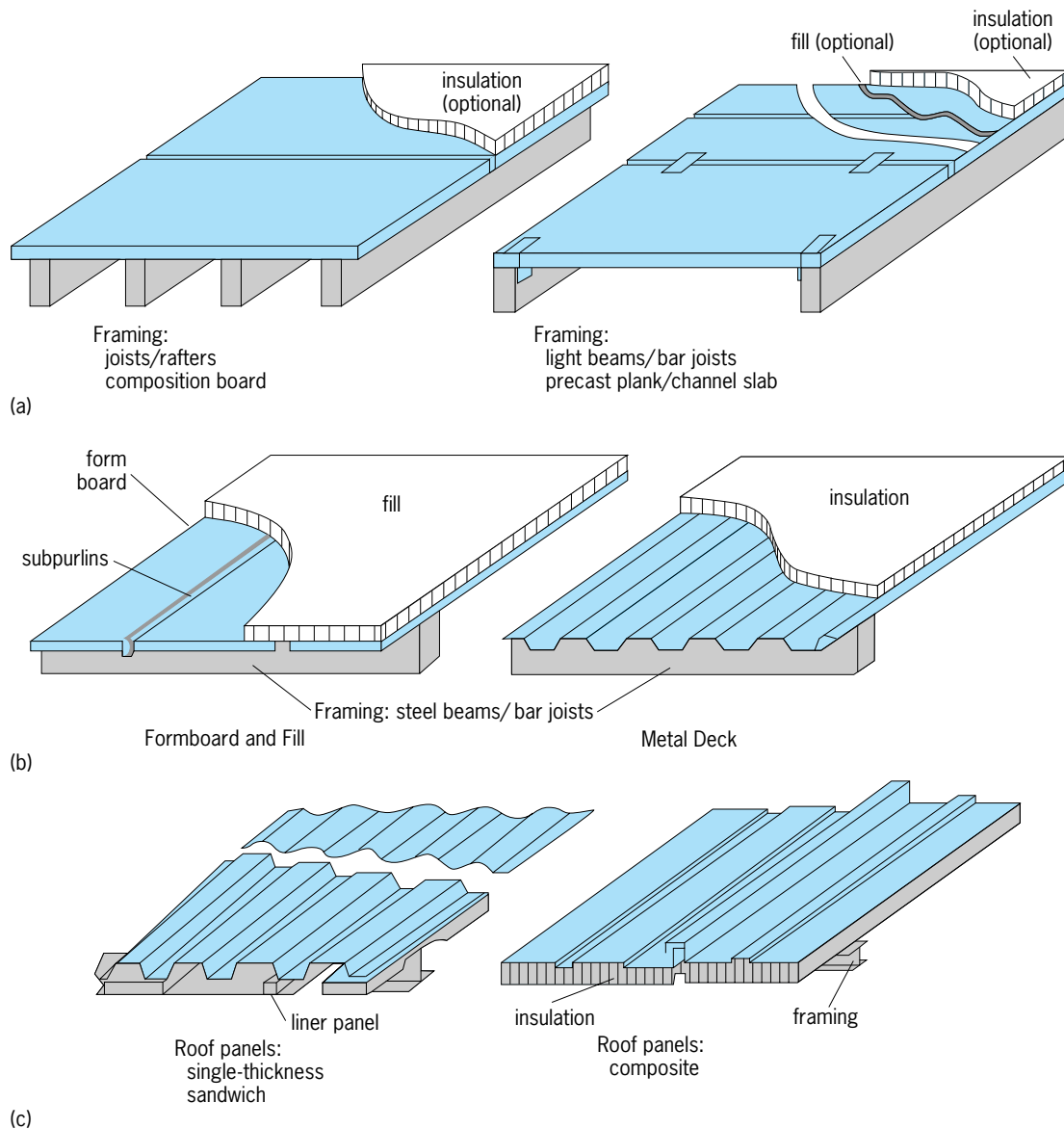

Fig. 2. Types of roof deck structure. (*a*) Site-assembled roof. (*b*) Flat roof. (*c*) Combined decking/roofing for a pitched roof. (*After Sweets Catalog, The McGraw-Hill Companies, 1993*)
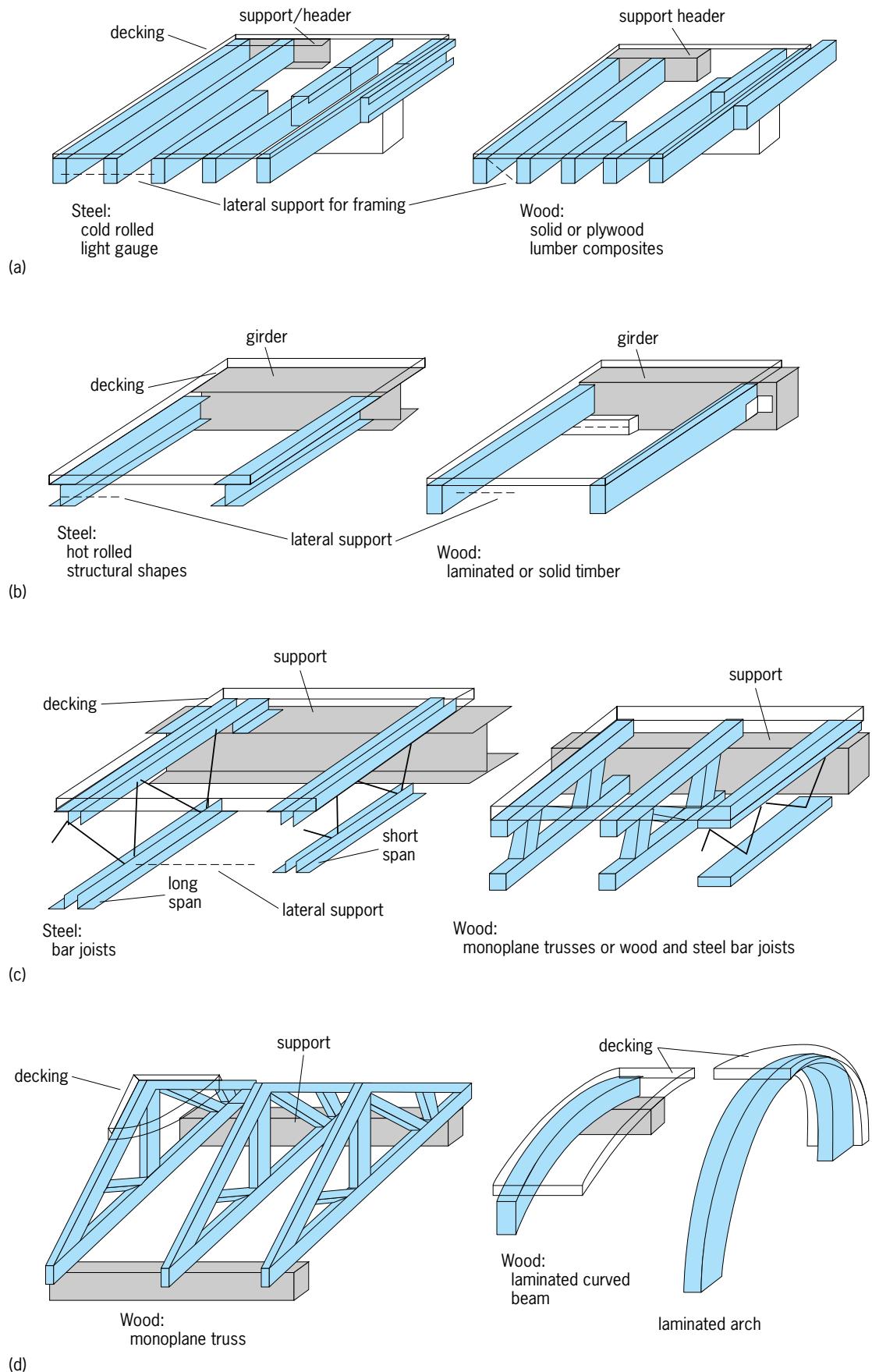
Fig. 3. Types of roof primary structural framing. (*a*) Joists/rafters, close spaced. (*b*) Beams. (*c*) Trusses, flat top chord. (*d*) Trusses/beams, pitched/curved. (*After Sweets Catalog, The McGraw-Hill Companies, 1993*)

as well as where there is a penetration through the roof. Eave constructions using gravel stops or metal fascia plates are designed to physically protect the roofing edge and prevent the entry of rainwater or snow melt under the roofing or into the top edge of the building wall construction. Where there are penetrations through the roof, cant strips, flashings, curbs, and other means are used to control water and see page (Fig. 1).

Structural components. A roof, which is the closure, is built upward from the structure below. The framework, or primary structural components, rest on the walls and columns of the structure, and these support the roof deck or the sheathing, which in turn carries the roofing assembly. The walls and columns may have girders framing into them. Beams rest on or are connected to girders. For flat roofs, these beams are called joists or purlins, while for hipped roofs they are termed rafters. Beams that are larger and made up of rigid geometric shapes, usually triangles, are termed trusses. Trusses can be made of wood or steel. Special types of trusses, formed from prefabricated light steel members, are called open web joists. The primary components may also be arches made of reinforced concrete, steel, wood, or laminated and glued wood sections. Under certain conditions a cast-in-place reinforced concrete deck, made integral with reinforced-concrete beams or arch sections, may be employed. *See* REINFORCED CONCRETE.

The roof deck or sheathing, the components that provide the basic support for the roofing assembly, span between and are anchored to the primary structural framing (**Fig. 3**). The various types of decking include wood planking, often tongue and groove, or plywood or special composition board; corrugated decking of steel or aluminum, sometimes covered with and structurally integral with the reinforced-concrete slab; precast planks of reinforced concrete or gypsum; cast-in-place reinforced concrete or gypsum, often placed on top of a sheet-rock-type ceiling material to provide a finished underside; and flat plates, sometimes with drop panels, which are reinforced-concrete slabs that are supported directly on columns without the use of separate beams.

Special roof designs. Where a long, clear space is required within a building, where a unique shape is desired for a structure, or where other constraints are demanding, special engineering designs are necessary for both the roof primary framing and the roof assembly. In many of these structures the deck and primary framework are effectively combined. Long-span roofs use space trusses, usually of steel; but reinforced-concrete domes or other shell shapes, including folded plates, may be employed. The TWA Terminal Building at Kennedy Airport in New York City, and the U.S. Air Force Academy Dining Hall in Colorado are examples of long-span roofs.

Cable supports. In cable-supported roofs, the primary framework is composed of cables in tension that are slung between separate posts or from the top of the surrounding building perimeter. The cables are restrained both vertically and horizontally by the building walls or other structural framing. The deck, which is supported by the tensioned cables, may be of any of the previously described assemblies or may be of a specially fabricated waterproof fabric. Tent-like or membrane roofs are a special application of a cable-supported roof. Here the membrane or the tentlike fabric is placed in tension directly, and it serves as both the support and the covering.

Air-supported roofs. Air-supported roofs utilize a waterproof coated fabric that is inflated to its rigid shape by developing and maintaining a positive air pressure inside the structure, which keeps the roof surface under tension. Tennis-court "bubbles" utilize this design to enable the fabric surfaces to resist the bending moments caused by external roof loadings.

Design loadings. These design parameters have evolved from much engineering study, field measurements, testing, experimentation, and construction practices; and they are formalized into codes. The basic design loadings include dead loads, live loads, wind loads, and seismic loads. The dead loads are essentially the constant loads due to the weights of the roofing assembly, the deck, the primary framework, and the hung ceiling loads, including any piping and ductwork in the space formed by the hung ceiling. The live loads are variable gravity loading, which include human occupancy or utilization of the roof surface for maintenance and repair; snow loads, which vary with the geographic location and the shape and slope of the roof; rainfall or snow melt water loads, which are the loads the roof must sustain if it is not properly drained; and special or concentrated loads due to roof-mounted equipment. Wind loads are generally complex and will vary with location, with variations due to roof height, slopes, and shape-caused turbulence. Wind loads may be applied both inward and outward on the roof surface, and they may oscillate. Seismic loading, which adds loading factors because of rapid earth movement during earthquakes, is a very specialized consideration, since it produces increases in both vertical and horizontal forces on the roof, requiring special design and construction considerations.

Roofs must be constructed under proper conditions, for if the deck is not dry, clean, or properly prepared, and the roof assembly is not properly put in place and anchored, with attention to flashing and eave details, the cover over the building will not last. *See* BUILDINGS; LOADS, DYNAMIC; SEISMIC RISK.

Milton Alpern

Bibliography. D. K. Ballast, *Architect's Handbook of Construction Detailing*, 1990; H. J. Cowan, *Design of Reinforced Concrete Structures*, 2d ed., 1988; K. F. Faherty, *Wood Engineering and Construction Handbook*, 3d ed., 1998; C. W. Griffin and R. Fricklas, *The Manual of Built-Up Roof Systems*, 1996; S. Hardy, *Time Saver Details for Roof Design*, 1997; F. S. Merritt, M. K. Loftin, and J. T. Ricketts, *Standard Handbook for Civil Engineers*, 4th ed., 1995; National Fire Protection Association Staff, *Roof Coverings and Roof Deck Constructions*, 1992.

Root (botany)

The absorbing and anchoring organ of vascular plants. Roots are simple axial organs that produce lateral roots, and sometimes buds, but bear neither leaves nor flowers. Elongation occurs in the root tip, which is usually 0.004–0.04 in. (0.1–1.0 mm) in diameter and 0.4–4 in. (1–10 cm) long. The older portion of the root, behind the root tip, may thicken through cambial activity. Some roots, grass for example, scarcely thicken, but tree roots can become 4 in. (10 cm) or more in diameter near the stem. Roots may be very long. The longest maple (*Acer*) roots are usually as long as the tree is tall, but the majority of roots are only a few centimeters long. The longest roots may live for many years, while small roots may live for only a few weeks or months.

Function. Root tips and the root hairs on their surface take up water and minerals from the soil. They also synthesize amino acids and growth regulators (gibberellins and cytokinins). These materials move up through the woody, basal portion of the root to the stem. The thickened, basal portion of the root anchors the plant in the soil. Thickened roots, such as carrots, can store food that is later used in stem growth. *See* CYTOKININS; GIBBERELLIN.

Growth environments. Roots usually grow in soil where it is not too dense to stop root tip elongation; there is enough water and oxygen for root growth; and temperatures are high enough (above 39°F or 4°C) to permit root growth, but not so high that the roots are killed (above 104°F or 40°C). In temperate zones, most roots are in the uppermost 4 in. (10 cm) of the soil; root numbers decrease so rapidly with increasing depth that few roots are found more than 6.5 ft (2 m) below the surface. Roots grow deeper in areas where the soil is hot and dry; roots from desert shrubs have been found in mines more than 230 ft (70 m) below the surface. In swamps with high water tables the lack of oxygen restricts roots to the uppermost soil layers.

Roots may also grow in the air. Poison ivy (*Toxicodendron*) vines form many small aerial roots that anchor them to bark or other surfaces. Fig (*Ficus*) and mangrove (*Rhizophora*) trees have aerial roots which may grow several meters through the air before reaching the soil. Aerial roots often have some characteristic that prevents desiccation; orchid roots have a special surface layer, the velamen, and screw pine (*Pandanus*) roots are up to 0. 4 in. (10 mm) in diameter with a thick cortex protecting the central vascular tissues.

Origin. The primary root originates in the seed as part of the embryo, normally being the first organ to grow. It grows downward into the soil and produces lateral second-order roots that emerge at right angles behind the root tip. Sometimes it persists and thickens to form a taproot. The second-order laterals produce third-order laterals and so on until there are millions of roots in a mature tree root system. In contrast to the primary root, most lateral roots grow horizontally or even upward. In many plants a few horizontal lateral roots thicken more than the primary, so no taproot is present in the mature root system.

Roots can also form laterals just behind the cut surface when they are severed. Unlike regular lateral roots, these roots bend to grow in the same direction as the parent, thus replacing the parent root. This type of root formation occurs after root pruning of woody root systems.

Adventitious roots originate from stems or leaves rather than the embryo or other roots. They may form at the base of cut stems, as seen in the horticultural practice of rooting cuttings. Adventitious roots may also form on branches which have been covered by fallen leaves or soil, a layering process common in shrubs like mountain laurel (*Kalmia*), and some trees like black spruce (*Picea*) where branches grow near the ground.

Bud formation. Roots of many plants can also form buds. When a root is cut into segments, buds can form at the end toward the stem and roots at the end toward the tip so that each segment develops into a new plant. Aspen (*Populus*), locust (*Robinia*), beech (*Fagus*), and other species produce root suckers, stems that grow out from buds on the roots when the parent stem is cut or injured. These new stems form large clones of genetically identical plants. *See* BUD.

Special forms. Woody roots can graft together. In grafts between individuals of the same species the roots grow next to each other, so that as they increase in diameter the two roots are pressed against each other and eventually grow together. Grafting may also occur between roots of the same individual, as in the case of the strangler fig (*Ficus*), which starts as a small plant growing in another tree. The roots grow down along the stem of the host tree, and after reaching the soil they thicken and graft together to form a woody cylinder that seems to be the stem of the fig tree when the host tree dies.

Some tree species that grow in flooded areas, for example, mangrove (*Avicennia*), form roots called pneumatophores that grow vertically upward from horizontal roots to reach above the water surface. Pneumatophores contain loosely packed cells that permit air movement which may help to aerate the underwater roots.

Fungal and bacterial relations. Hyphae of some fungi can invade root tips and form mycorrhizae. Almost all plants have mycorrhizae which help them grow on mineral-deficient soils. Apparently the hyphae provide minerals to the plant, and the plant provides carbohydrates to the fungus. Many form characteristic structures such as many-forked short roots in pines (*Pinus*) and a series of beads in maples (*Acer*). *See* FUNGI; MYCORRHIZAE.

Some plants can form root nodules that fix inert organic nitrogen into usable organic compounds. Leguminous plants, such as peas (*Pisum*) and black locust trees (*Robinia*), form nodules that contain *Rhizobium* bacteria. Nonleguminous nitrogen-fixing plants occur in many different plant groups, such as

gale (*Myrica*) and alder (*Alnus*), which form nodules with actinomycete fungi. *See* NITROGEN FIXATION.

B. F. Wilson

Root systems. The extent of the system or root mass varies in relation to inherited growth characteristics and factors such as soil porosity, aeration, and the availability of water (**Fig. 1**). In general, root systems are extensive, spreading widely or penetrating deeply, the numerous finer branches and their root hairs being in contact with a large volume of soil. Their physical attributes are evaluated in different ways (**Tables 1** and **2**). For example, as much as 30 ft (9 m) of root with an external surface up to 65 in.^{2} (420 cm^2) have been found in 1 in.^{3} (16 cm^3) of soil from under grass. In many species the root system constitutes the bulk of the plant body. Although most of the roots of many trees may be found in the top 4–5 ft (1–1.5 m) of the Earth's crust, their lateral spread may extend through the soil considerably beyond the tips of the longest branches of the crown. The form of root systems differs with the origin and manner of growth of their members. Nearly all are variations of taproots and fibrous roots.

Taproots. In taproot systems the primary root forms a dominant central axis. The main root penetrates rather deeply into the soil and is generally larger than its branches. Plants with taproots range from trees, in which the primary root and its main branches are thick and woody, to herbs, in which the taproot may be rather slender or develop into a fleshy food-storage organ as in a carrot plant.

Fibrous roots. Several to many main roots of equal dominance typify fibrous root systems. Most commonly these main roots arise adventitiously from the stem as in grasses, but sometimes fibrous roots are composed of branches of a primary root that ceased to be dominant. In some species, root systems organized like fibrous roots are composed of thick and fleshy units, for example, those of dahlia.

Root cuttings. Roots of many plants normally form buds from which shoots develop, or they may be induced to form buds by injury or severe pruning. Cuttings, or short lengths, of roots may be planted to propagate such plants. *See* REPRODUCTION (PLANT).

Primary tissues. Roots are characterized by a pattern of apical growth similar to that of stems. Through the persistent activities of meristematic (rapidly dividing) cells at the root apex, new cells are added continually to the root. A terminal root cap covers this meristematic region (**Fig. 2**). *See* STEM.
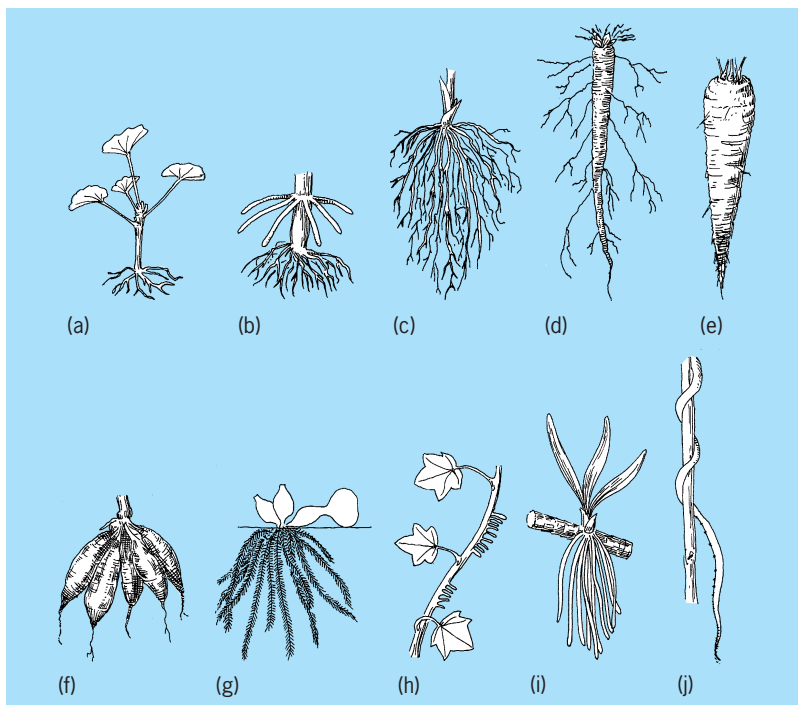

Fig. 1. Kinds of roots. (*a*) Adventitious, on a cutting or slip (geranium). (*b*) Prop and other adventitious (corn). (*c*) Fibrous adventitious (grass). (*d*) Primary, or taproot (dandelion). (*e*) Fleshy taproot (carrot). (*f*) Fleshy, fascicled adventitious (dahlia). (*g*) Aquatic adventitious (water hyacinth). (*h*) Aerial adventitious (English ivy). (*i*) Aerial adventitious (orchid). (*j*) Parasitic, or haustorial, adventitious (dodder).

Apical meristem. In a rapidly growing root, two distinct meristematic regions are evident. They are separated from each other by a group of nonmeristematic cells known collectively as the quiescent center (**Fig. 3**). On the basal or proximal side of the quiescent center is located the proximal meristem, which produces most of the new cells in a root. On the opposite or distal side of the quiescent center is the distal meristem. Both meristems may fluctuate in size, and often can be shown to have specific developmental relationships to the primary tissues or to the root cap. Depending upon the plant species, approximately 500–2000 cells can compose the quiescent center. Three-dimensionally these cells often assume the shape of an inverted cup, along the periphery of which are located cells of the proximal and distal meristems. The quiescent center is visualized most clearly by the techniques of autoradiography, which demonstrate that cells of the quiescent center divide rarely. A quiescent center is absent in embryonic roots, but is evident very soon after seedling

TABLE 1. Dimensions of root systems		
Plant	Lateral spread, ft (m)	Depth, ft (m)
Little bluestem (*Andropogon scoparius*)	1.0 (0.3)	4.0 (1.2)
Blazing star (*Liatris punctata*)	9.0 (3.0)	16.0 (5.6)
Comanche cactus (*Opuntia camanchica*)	9.0 (3.0)	2.8 (0.8)
Wheat (*Triticum aestivum*)	2.0 (0.6)	5.0 (1.5)
Corn (*Zea mays*)	8.0 (2.5)	7.0 (2.4)
Sugarbeet (*Beta vulgaris*)	3.0 (0.9)	6.0 (1.8)

TABLE 2. Quantitative characteristics of root systems

Characteristic	Crested wheat grass	Winter rye	Coffee
Age	2 years	4 months	3 years
Soil mass	56 ft^3 (1.59 m^3)	2 ft^3 (0.056 m^3)	270 ft^3 (7.65 m^3)
Total root length	315 mi (507 km)	387 mi (623 km)	14.6 mi (23.5 km)
Number of roots		13,000,000	
Total surface		2554 ft^2 (237 m^2)	

germination, and may fluctuate in size throughout the life of an individual root. Developmentally the function of the quiescent center is not known, though it has been implicated in the production of plant growth substances in the root.

In roots of ferns a quiescent center is absent, and new cells are thought to arise from the meristematic activities of a single enlarged apical cell and its immediate derivatives. *See* APICAL MERISTEM.

Root cap. The root cap is composed of cells at various stages of differentiation and, because of its terminal position, is considered to be a protective structure. It serves as the site of gravity perception in the root, and is believed to facilitate the movement of the root through the soil by the production of mucilaginous substances.

Primary meristems. Cell divisions are most frequent in the primary tissue regions of the root, which usually become distinct immediately behind the apical meristem. In recognition of their meristematic nature at this root level, these regions are often designated as primary meristems. These meristems are the protoderm, ground meristem, and procambium. Protoderm represents the surface layer which develops into the epidermis; procambium designates the central region which forms the vascular cylinder (stele); and ground meristem constitutes the remaining tissue which gives rise to the root cortex.
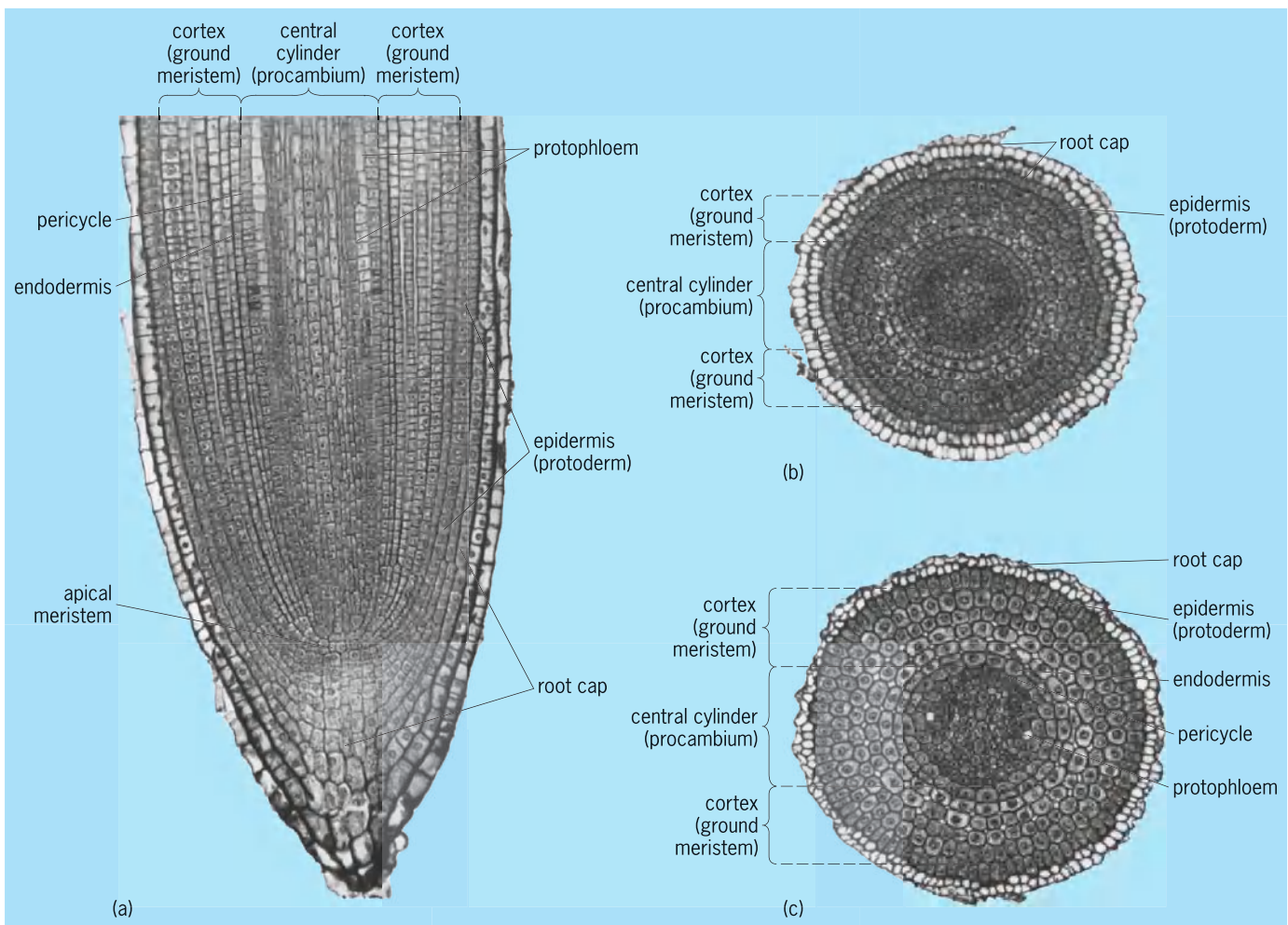

Fig. 2. Root-tip organization of a tomato plant. (*a*) Median longitudinal section, showing meristematic region, which is covered by a terminal root cap. (*b*, *c*) Transverse sections of same root at 160 and 480 μm behind root apex (apical meristem), respectively.
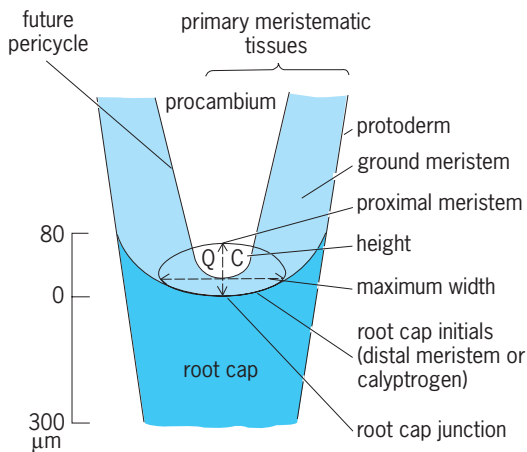
Fig. 3. Distribution of tissues in a typical root apex.

Epidermis. The root epidermis usually consists of a single layer of elongated, thin-walled cells in which the cuticle is difficult to demonstrate with certainty (**Fig.** 4*a* and *b*). In many older roots the epidermis is sloughed off. If the layer persists in mature roots, more or less conspicuous cutinization or other wall modification may develop. The production of root hairs is characteristic of the root epidermis. *See* EPIDERMIS (PLANT).

Root hairs. These structures develop as narrow, tubular outgrowths of epidermal cells, their walls and contents being continuous with those of the main body of the cell. They increase the absorbing surface of the root by contact with a much greater surface of soil than would be touched by the root surface without hairs. All epidermal cells may be capable of producing root hairs, or in some species they may be formed only from special short cells, called trichoblasts. Root hairs are usually confined to a short length of root behind the zone of elongation. Those toward the root tip are youngest, and toward the base they are progressively older and longer. As a root grows in length, new root hairs are initiated toward the tip while the oldest ones usually degenerate. The newly developing root hairs make contact with additional soil particles and maintain an extensive absorbing surface.

Cortex. Thin-walled cells among which intercellular spaces develop to varying degrees are common features of the root cortex. Cells may be arranged more or less irregularly, or they may show radial or concentric patterns, separate or in combination. Cell arrangement may reflect some aspects of cortical growth; for example, radial seriation arises as the result of centripetal growth involving successive periclinal (parallel with the circumference) divisions of the innermost cells. If these divisions are synchronized, rings of cells or a concentric pattern may also be apparent. Small intercellular spaces, schizogenous (splitting of adjacent cell walls) in origin, develop near the root apex, particularly in the inner cortex. Larger lysigenous (dissolution of cells) cavities may also develop in the older root. One or more layers of the outermost cortical cells may develop suber-

ized or lignified walls which serve protective functions. Those of the innermost layer usually become specialized as an endodermis. *See* CORTEX (PLANT).

Endodermis. Cells of the endodermis exhibit wall modifications which involve thickening and deposition of suberin or lignin. After the centripetal growth of the cortex is completed, suberin and lignin are deposited in the form of a band, the Casparian strip, on and within the transverse and radial walls of the endodermal cells. In many roots the endodermal cells undergo no further changes; in others an extensive thickening of the radial and inner tangential walls follow. *See* ENDODERMIS.

Pericycle. The outer portion of the central cylinder, or stele, is designated as the pericycle, which may be one or more cell layers in thickness. Although generally continuous, it may be interrupted by xylem elements as in roots of some grasses (Fig. 4*c*). Cell divisions in the pericycle initiate the growth of lateral roots and contribute to the formation of the vascular cambium and periderm in roots with secondary growth. *See* PERICYCLE.

Vascular system. The vascular system, consisting of primary xylem (water-conducting) and phloem (food-conducting) tissues, constitutes the bulk of the central cylinder in many roots. Phloem occurs in strands near the periphery the periphery; xylem alternates with phloem, either as strands or as ridges of a central xylem mass. This distribution, that is, the arrangement of xylem and phloem on separate radii, is characteristic of roots. Whereas a central pith occurs in the roots of some dicotyledons, it is common among monocotyledons, although in many of the latter the central core consists of lignified parenchyma, including one or more strands of water-conducting cells.

Maturation of the primary vascular tissues is a complex process and occurs at various levels in the root. The outer xylem in the ridges or strands is the first xylem to mature, and is designated as protoxylem; the inner xylem matures after the protoxylem, and is called the metaxylem. Thus the maturation of the primary xylem occurs in a centripetal direction, and the xylem having such a pattern of maturation is termed exarch. Protophloem is likewise external to metaphloem. Phloem usually matures closer to the root tip than the xylem. As seen in cross sections, the points of origin of the protophloem and protoxylem may be called phloem and xylem poles, respectively. Roots are described as monarch, diarch, and triarch, on the basis of whether there are one, two, or three poles; those with many poles are designated as polyarch. The number of poles may vary among roots of different plants, of different roots of the same plant, or at different levels of an individual root. This variation in pole number is now believed correlated with changes in the dimensions of the quiescent center. A large quiescent center is suggested to be associated with complex vascular tissue patterns. Dicotyledon roots generally have fewer poles than those of monocotyledons. *See* PHLOEM; XYLEM.

Lateral roots. Root branches or lateral roots are generally produced at some distance behind the apex,
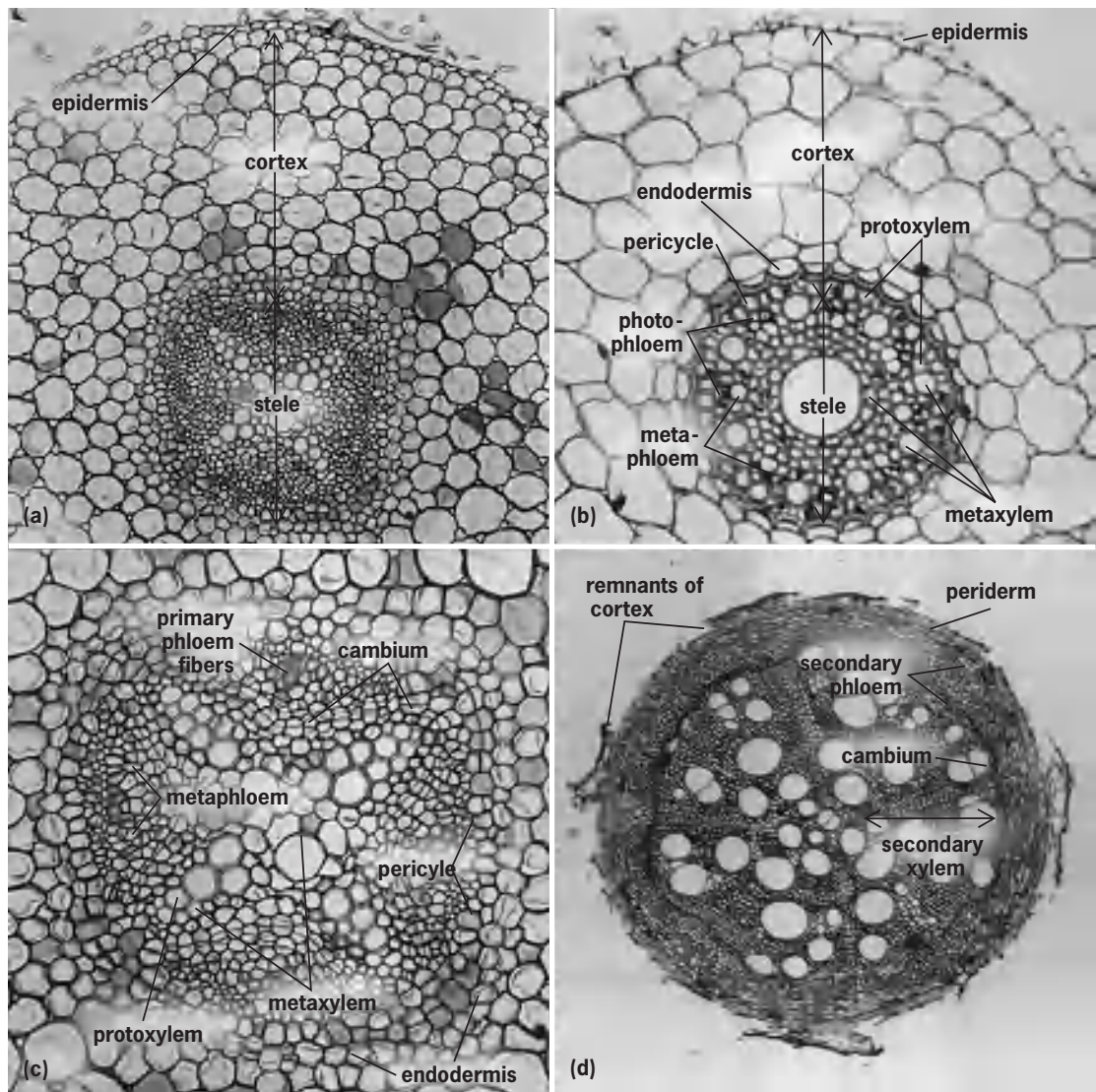
Fig. 4. Transverse sections of roots, showing primary and secondary tissues. (*a*) Soybean. (*b*) Barley. (*c*) Soybean; enlarged part of *a*, showing early formation of cambium. (*d*) Soybean with abundant secondary tissue.

and they originate endogenously, or deep within the tissues. Depending on the number of poles in the parent root, lateral roots arise in distinctive positions in relation to the xylem and phloem. A lateral root is commonly initiated by localized cell divisions in the pericycle; however, in lower vascular plants the initial divisions may involve endodermal cells. Continued divisions produce a primordium in which the root cap, apical meristem, and primary tissue regions are soon organized. The young root grows through the cortex, and by the time it emerges, vascular connections with the main axis are established.

Elongation. Some elongation characteristic of primary growth of the root results from the increase in number of cells in the meristematic tip, but most of it is produced by the elongation of cells behind the meristematic region. The location and extent of the zone of elongation varies, in part with root size and stage of growth. For example, in relatively thin, rapidly growing roots of timothy (*Phleum*) the zone of cell elongation occurs 450–1400 micrometers be-

hind the root apex, whereas in larger primary roots of corn (*Zea*) the zone lies 2000–9000 μm behind the root tip.

Differentiation. The primary tissues of the root begin to differentiate during the elongation but complete their differentiation, or become mature, after the elongation ceases. Differentiation consists in a series of changes through which the cells become more or less specialized with regard to the functions they carry out in the mature state. The differentiation follows characteristic patterns in both longitudinal and transverse directions. Thus, for example, the protophloem matures more closely to the apex than the protoxylem, and both differentiate in the direction from the base of the root toward the apex, that is, acropetally. The metaphloem and metaxylem follow in the same direction. Because the phenomena of growth, that is, cell division and cell enlargement, overlap with the earlier phenomena of differentiation, the sequence of maturation of a root is highly complex.

Secondary tissues. The formation of secondary xylem and secondary phloem by the vascular cambium is the chief characteristic of secondary organization in roots. Other structural modifications invariably accompany cambial activity and thus are also features of the secondary state. *See* LATERAL MERISTEM.

The extent of secondary tissue formation in roots generally parallels that in stems. Trees and shrubs with woody stems are known to develop woody roots. Herbaceous plants with a smaller volume of secondary tissues in the stem also show less cambial activity in the roots. However, roots of plants that have little or no secondary vascular tissues in the stems have not been investigated sufficiently to determine the presence or absence of a cambium.

Development of vascular cambium. Formation of the vascular cambium in the dicotyledons begins at root levels where primary development is approaching completion. When viewed in transverse section, the cambium first appears as short disconnected arcs of periclinally dividing cells internal to the phloem strands. As vascular cells are formed by the periclinal divisions in the arcs of cambium, pericycle cells opposite the xylem poles also divide periclinally. Inner cells formed by divisions in the pericycle differentiate as cambium and unite the arcs internal to the phloem. A circular or cylindrical distribution of the cambium results from the accumulation of secondary xylem internal to the phloem strands, whereby these strands and adjacent cambium are displaced outwardly. Subsequently, complete cylinders of secondary xylem and secondary phloem are deposited inside and outside the cambium, respectively (Fig. 4d). As cambial activity progresses, the outer tissues are subjected to tension and pressure. Primary phloem is crushed and individual cells degenerate, except for those which in some plants differentiate into primary phloem fibers. The cortex may be split and completely shed, as described in the subsection on periderm, which follows. The original cambium adjusts in circumference to diameter increase and normally continues to function throughout the life of the root.

Although monocotyledons generally lack cambial activity, in some, secondary growth results from the activity of a cambiumlike thickening meristem that originates in the pericycle or cortex and produces a secondary body of vascular bundles embedded in parenchyma tissue.

Secondary vascular tissues. The cambium and secondary tissues derived from it are similar in general organization to those of the stem, but certain quantitative differences between the root and stem tissues and their constituent cells are present. One of the most common characteristics of roots is a relatively large volume of parenchyma, a feature considered to be a result of high degree of specialization for storage. The abundant parenchyma results from normal cambial activity or from anomalous growth involving irregular cell proliferation in the secondary xylem or formation of additional layers of cambium outside the normal.

Periderm. Periderm, a secondary protective tissue, is characteristic of woody roots, but it may not develop in those with only a small degree of cambial activity. This tissue is derived from a lateral meristem called phellogen, or cork cambium, which most commonly has its origin in the pericycle. Its characteristic component is the cork tissue. *See* BARK; PERIDERM.

After the vascular cambium becomes complete outside the xylem poles, periclinal divisions in the pericycle continue around the vascular cylinder, and the outer cells thus formed function as a phellogen. It begins to form periderm as the vascular core enlarges and stretches the cortex, which is ultimately split and shed. If the original periderm is ruptured by growth in diameter, new layers are formed from new phellogen which differentiates in parenchyma of the secondary phloem.

Periderm formation may also follow a different course. In some dicotyledons it first arises from subepidermal cells, but subsequently another is formed from a proliferated pericycle. In others, it originates only in the outer cortex. Many roots with little or no secondary thickening lack a periderm, the outer protective tissue being the persisting cutinized epidermis or an exodermis of subepidermal origin. The exodermis comprises layers of cells with suberized or lignified wall thickenings.

Root and shoot connection. The more or less cylindrical vascular system of the root is morphologically distinct from that of the shoot, which is composed of interconnected vascular bundles. In the hypocotyl (seedling axis below the cotyledons) the vascular tissues are arranged in a pattern that is intermediate or transitional between those of the two systems, root and stem. Because of its intermediate structure, the hypocotyl is often designated as the transition region. In this region, the rather solid vascular core of the root, with its radial tissue arrangement and exarch xylem, is merged with separate collateral, vascular bundles having endarch xylem. These bundles are the leaf traces of the cotyledons, the first foliar organs of the shoot. Structural features of the transition region are related to spatial adjustments which provide tissue continuity between the root and shoot systems of contrasting arrangements of tissues and patterns of differentiation.

The organization of the transition region varies greatly in different species. Essential features of a pattern common among dicotyledons are shown in **Fig. 5**. In this example the diarch xylem plate of the root is continuous with two leaf traces of the cotyledons, each obviously double in structure. At successively higher levels the relative position of protoxylem and metaxylem changes. Each phloem strand of the root is connected with each of the two cotyledonary traces. Initial union with the vascular system of the epicotyl (shoot above the cotyledons) is established indirectly with the root, the traces of the first leaves remaining distinct throughout the hypocotyl.

Other types of transition occur in plants with more than one cotyledonary trace or with three or more
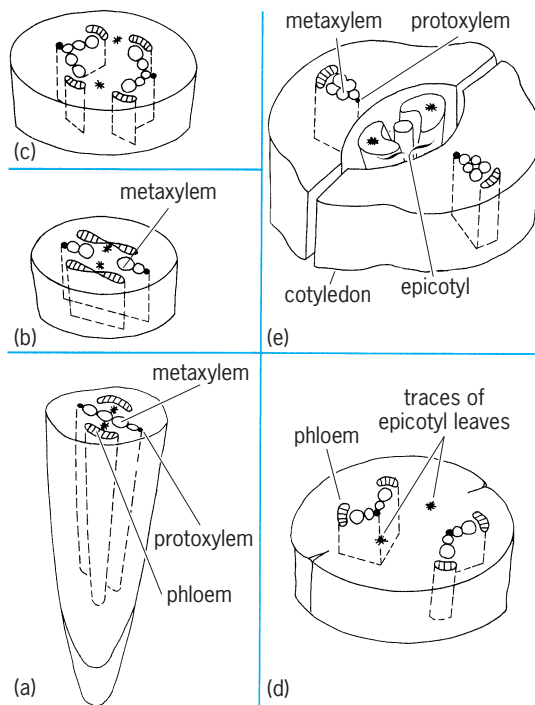
Fig. 5. Diagrams of vascular transition in dicotyledonous seedling (beet, *Beta vulgaris*) from root tip upward to level of cotyledons. (*a*) Root is diarch; its vascular tissues are continuous with those of (*e*) the cotyledons through (*b–d*) the transition region in the hypocotyl. (*After K. Esau, Plant Anatomy, John Wiley and Sons, 1953*)

poles of xylem and phloem in the root. Differences in the timing of epicotyl development and in the pattern of germination (epigean or hypogean cotyledons) also effect variations in the transition region. In plants that develop a cambium, secondary tissues are continuous between the root and stem.

The general organization of the transition region in gymnosperms is similar to that of dicotyledons, any variations being related to the usually greater number of cotyledons in gymnosperms and to the number of traces to each cotyledon. In monocotyledons the vascular system of the root is commonly connected with those of the single cotyledon and the epicotyl; transitional characteristics may occur in both connections or only in that with the cotyledon. A particularly complex transition characterizes the grass family, Gramineae. Vascular strands of the root join in a nodal plate at about the level of the scutellum or cotyledon. Also joined with the plate are vascular strands of the scutellum, coleoptile (first leaflike structure above the cotyledon), and first leaves. In this complex pattern the transition between vascular systems of root and shoot is relatively abrupt, and the root system is connected with more than one leaf above the scutellum. *See* SEED.

Charles Heimsch; Lewis Feldman

Root density characterization. The density of roots in a volume of soil is important in determining the ability of the plant to extract water and nutrients located there. The density is characterized by measuring either the length or mass of roots in a specific soil volume. A traditional means for obtaining root length

density (feet of roots per cubic foot of soil or meters of roots per cubic meter of soil) or root mass density (pounds of roots per cubic foot of soils or grams of roots per cubic meter of soil) has been to extract representative samples from the soil volume occupied by the roots, wash the soil from the roots, and measure the root length or root mass contained in the sample. This procedure is very time-consuming and labor-intensive if the plants are large. In addition, part or all of the root system is destroyed by the sampling process, making successive measurements impossible on the same roots.

Minirhizotrons, rhizotrons, computer-aided tomography, neutron radiography, and nuclear magnetic resonance are techniques which estimate root length density without disturbing the plant root system. Computer-aided analyses and enhancement of images obtained by these techniques also allow root mass density to be determined, sometimes without any destructive sampling of the root system.

Minirhizotrons. This technique involves installing a clear tube into the soil and lowering into the tube a viewing device that allows the operator to see roots that intersect the tube. Tubes are installed at 30–45° from the vertical a few days before measurements are to be started. The tubes are marked with permanent reference points prior to installation. A battery-powered light allows the roots to be observed through a mirror, a fiber-optic lens, and a borescope or a color video camera with a right-angle lens. Photographic or video tape images can be stored for later processing. Root length density can be determined by measuring the length of roots at the interface and estimating the volume of soil which these roots would have occupied if the tube had not been present, or by estimating the length of roots that would have grown in a specific volume of soil occupied by the tube. Root mass densities can be estimated from root length density, root diameter distribution, and specific weight values. Several tubes are required to estimate root densities accurately.

Rhizotrons. These are covered underground cellars or walkways with clear windows on one or both sides. The windows may be in contact with natural soil that surrounds the cellar or may be one wall of a soil-filled box. Plants are grown in soil adjacent to the window. Roots are observed either at the window–soil interface or by inserting a viewing device into clear tubes placed horizontally into the soil at different depths. Empirical relationships can be developed between the length of roots at the soil–window interface and root length density in soil away from the window when the horizontally placed tubes are used to view the roots. The number of roots that intersect the tubes are counted, and empirical relationships are developed between root counts and root length densities. Roots tend to be more concentrated near the window than in bulk soil, but light shining on the roots through the window tends to reduce root concentration at the interface, especially for most legumes.

Computer-aided tomography (CAT). This technique uses x-rays or gamma rays to measure the density of a soil

matrix and to infer the distribution of water in that matrix. Living plant material, such as a root, has a high concentration of water and can be seen readily in a computer-generated image. The soil sample is placed in a fixed position between a radiation source and a detector. The radiation source and detector, which are a fixed distance apart, travel a $360°$ circular path around the stationary sample. The radiation beam passes through the sample, and the transmitted radiation is measured by the detector. Electrical impulses from the detector are integrated by a computer that constructs an image representing the internal structure of the sample. *See* COMPUTERIZED TOMOGRAPHY.

Neutron radiography. This technique is based on the fact that hydrogen atoms will slow (thermalize) high-energy neutrons to a greater extent than most other atoms and that roots in soil are locally high concentrations of water, which is composed of hydrogen and oxygen. Water in roots effectively removes more neutrons from a collimated beam than does water in moist soil. The beam produces an image on a rare-earth transfer screen, which can be further transferred to x-ray film.

Small sample sizes, safety concerns for the users, and the need for access to a neutron source of at least 1.4×10^6 neutrons cm^{-2} s^{-1} are the principal problems associated with neutron radiography. *See* RADIOGRAPHY.

Nuclear magnetic resonance. This technique uses a strong magnetic field and a sequence of radio-frequency pulses and magnetic-field gradients to create a digital image of the spatial distribution of water in a soil sample that contains roots. These images allow the determination of time-dependent changes in water contents of the root, the rhizosphere, and adjacent soil. Resolution of the image increases as sample size decreases. *See* NUCLEAR MAGNETIC RESONANCE (NMR); RHIZOSPHERE.   Howard M. Taylor

Bibliography. H. M. Taylor (ed.), *Minirhizotron Observation Tubes*: *Methods and Applications for Measuring Rhizosphere Dynamics*, Amer. Soc. Agron. Spec. Publ. 50, 1987.

Root (mathematics)

If a function $f(x)$ has the value 0 for $x = a$, a is a root of the equation $f(x) = 0$. The fundamental theorem of algebra states that any algebraic equation of the form $a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$, where the a_k's are real numbers ($k = 0, 1, \ldots, n$), has at least one root. From this it follows readily that such an equation has roots, real or complex, in number equal to the index (here n) of the highest power of x.

Furthermore, if $a + ib$ (where $i = \sqrt{-1}$) is a complex root of the given equation, so is $a - ib$, the conjugate of $a + ib$. Equations of degrees up to four may be solved algebraically. This statement means that the roots may be expressed as functions of the coefficients, the functions involving the elementary arithmetical processes of addition, multiplication, raising

a number to a power, or extracting the root of a certain order of a given number. It was proved by H. Abel and by E. Galois that it is not possible to solve algebraically the general algebraic equation of degree higher than four. However, it is possible to determine the real roots of an algebraic equation to any desired degree of approximation.

The term zero is sometimes used in lieu of root when dealing with functions which are defined as infinite power series. Thus, one talks about the zeros of sin x, cos x, and $J_0(x)$ [the Bessel function of first kind and zeroth order]. Each of these three functions may be expressed as an infinite power series.

Numbers are called transcendentals if they cannot be the roots of any algebraic equation with integral or rational coefficients. The most important transcendental numbers are π, the ratio between the circumference of a circle and its diameter; and e, the base of the system of natural logarithms. *See* CALCULUS; EQUATIONS, THEORY OF; NUMERICAL ANALYSIS.
Arnold N. Lowan; Salomon Bochner

Root-mean-square

The square root of the arithmetic mean of the squares of a set of numbers is called their root-mean-square. If the numbers are $x_1, x_2, x_3, \ldots, x_n$, then

$$\text{Root-mean-square} = \sqrt{\frac{x_1^2 + x_2^2 + x_3^2 \cdots + x_n^2}{n}}$$

It is valuable as an average of the magnitudes of quantities, and it is not affected by the signs of the quantities.

Among applications of root-mean-square the most important is the standard deviation from the arithmetic mean $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + x_3 \cdots + x_n}{n}$$

Then the standard deviation

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

Thus standard deviation from the mean is the root-mean-square of the deviations from the mean. It has a great advantage over other measures of deviation because its computation can be greatly facilitated by the following transformation. If a set of values $x_1, x_2, x_3, \ldots, x_n$ are each reduced by the same number k and then the results are divided by the same number p, the root-mean-square of the deviations of the resulting quantities from their mean is equal to $1/p$ times the root-mean-square of the deviations of the x's from their mean. Thus if $x_i = pu_i + k$, Standard deviation of the x's = $p \times$ (standard deviation of the u's).

The most common method of minimizing errors in the use of a formula to fit empirical observations is that of determining the coefficients in the formula so as to make the mean of the squares of the errors

least. This is the principle of least squares. Its relation to root-mean-square is obvious since minimizing the root-mean-square of the errors accomplishes the same purpose. *See* CURVE FITTING; STATISTICS.

Hollis R. Cooley

Rope

A long flexible structure consisting of many strands of wire, plastic, or vegetable fiber such as manila. Rope is classified as a flexible connector and is used generally for hoisting, conveying, or transporting loads; transmitting motion; and occasionally transmitting power. For flexibility and to reduce stresses as the rope bends over the sheave (pulley), a rope is made of many small strands.

Wire rope. In making wire rope, the wires are preformed and wound into strands which are then twisted together to form the rope. A hemp center is usually used to hold lubricant for reducing friction and wear as the wires rub over each other when the rope is bent. *See* WIRE.

Figure 1 shows a 6 × 7 construction (6 strands, each with 7 wires). The hemp center is shown in black. Usual constructions are 6 × 7 coarse laid, 6 × 19 standard hoisting, and 8 × 19 and 6 × 37 extraflexible hoisting for elevator service. To keep bending stresses low, sheaves of large diameter must be used. Factors of safety based on ultimate strength of the rope vary from 3 to 8 depending on the application. Some manufacturers' data allow for bending stresses in normal size sheaves; otherwise these stresses must be allowed for separately.

There are two common ways of winding wire rope indicated by lay of the rope. Standard rope is called regular lay; the wires in the strand are laid to the left and the strands to the right. The wires are at right angles to the motion of the rope as it slides into the sheave groove so that the wire may wear excessively, particularly if the sheaves are not well aligned. In lang lay, wires in the strand and the strands themselves are laid in the same direction. This type of rope wears better, is more flexible, and lasts longer. Flat rope is made of right and left strands placed alternately beside each other and sewed together with soft steel wire. This rope can be used on a reel slightly wider than the rope; it is used to save space.

For aircraft service, 6 × 7, 7 × 7, and 7 × 19 cord and 19-wire strand have been developed. They are made of tinned or galvanized carbon steel or stainless steel.

Materials for commercial wire ropes in order of strength and cost are iron, mild plow steel, plow steel, and improved plow steel. Galvanized-wire rope
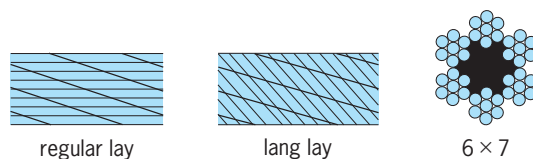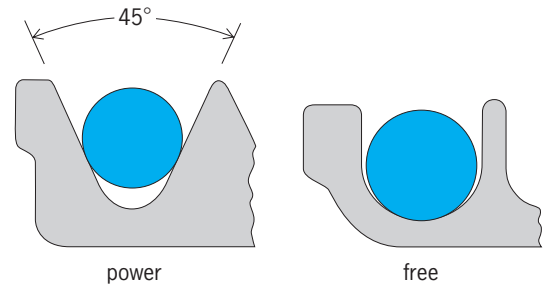

Fig. 1. **Examples of wire-rope construction.**

regular lay   lang lay   6 × 7

power   free

Fig. 2. **Groove profiles for rope.**

may be used for standing service, as for guy wires, but should not be used in hoisting because the galvanizing wears off. Failure of wire rope is usually caused by wear of wires as they rub on each other during bending or by high compressive stress as they bear against the grooves in power transmission (**Fig. 2**). Internal corrosion due to water or other corrosive substances causes failure. Profiles of grooves on the sheaves are semicircular at the bottom when the rope runs free for hoisting and transportation (Fig. 2). The rope should not be pinched at the bottom of the groove. For power transmission with wire ropes, the contacting part of the semicircular groove should be lined with rubber or similar resilient material.

Wire ropes may be attached to a hoisting drum by using a rope socket at the end of the rope; the socket is slipped into a pocket cast in the drum at the end of the last groove. The socket is held in place by a screwed plug, which can be easily removed to detach the rope. To attach the rope to its load or to a block-and-tackle anchor, various sockets, clips, and thimbles are available.

Manila rope. Manila fiber is twisted into yarn, the yarn is twisted into strands, and the strands are laid up into rope. Manila rope is used mainly for light service hoisting but sometimes for long-distance transmission of power. Being flexible, this rope accommodates itself to small-size pulleys and absorbs starting and shock loads, but the fiber at the center of the strand powders and pulverizes, although the rope may appear sound at the surface. This deterioration, from the rubbing of strands on each other as the rope flexes, can be retarded by such lubricants as acid-free paraffin or graphite worked into the strands during laying or twisting of the strands to form the rope. *See* HEMP; NATURAL FIBER.

For power transmission and for hoisting with manila rope see the groove profiles shown in Fig. 2. The 45° groove angle that is used for power transmission provides sufficient friction to prevent slippage between the rope and the surface of the groove. *See* PULLEY.

Paul H. Black

Rosales

An order of flowering plants in the eurosid I group of the rosid eudicots. In previous modern systems of classifications, this order was considered to be

much larger and to contain up to 25 families, but recently on the basis of DNA sequence studies the number of families was greatly reduced and changed. The order as now recognized contains only 11 families, many of which are small. Many families in this order are wind-pollinated (the former Urticales, for example Moraceae, Ulmaceae, and Urticaceae) and exhibit the typical syndrome of small petalless flowers with dangling anthers, whereas the other families are insect-pollinated with large, showy flowers in which the carpels are sometimes free. *See* FLOWER; MAGNOLIOPHYTA; MAGNOLIOPSIDA; POLLINATION.

The largest family is Rosaceae with nearly 3000 species. The great number of economically important trees, shrubs, and herbs in this family include apples (*Malus*), pears (*Pyrus*), almonds, cherries, plums, and prunes (*Prunus*), strawberries (*Fragaria*), blackberries, raspberries, and their relatives (*Rubus*), as well as many minor fruits such as quinces (*Cydonia*), loquat (*Eriobotrya*), and medlars (*Mespilus*). Many are grown as ornamental plants, including roses (*Rosa*), avens (*Geum*), cinquefoil (*Potentilla*), firethorn (*Pyracantha*), redbush (*Photinia*), and spirea (*Spiraea*). *See* ALMOND; APPLE; BLACKBERRY; CHERRY; ORNAMENTAL PLANTS; PEAR; PLUM; QUINCE; RASPBERRY; STRAWBERRY.

The second-largest family, Rhamnaceae (900 species), includes mostly woody species (shrubs and trees) that differ from Rosaceae in having fused carpels and only five stamens (rather than many) that are placed in the same position as the petals. A number of species in Rhamnaceae are of minor economic importance as timbers, some are of medicinal use or as dyes, and one is a minor fruit crop, the jujube (*Zizyphus*). A few are ornamentals, such as *Ceanothus* and *Colletia*.

Of the families formerly placed in Urticales, the largest are Moraceae (950 species) and Urticaceae (700 species), the exact limits of which are problematic. Many of these are large forest trees, but some are forest herbs and weeds. In Moraceae, many species are sources of timber and fruit, including figs (*Ficus*), mulberries (*Morus*), and breadfruits (*Artocarpus*). In Urticaceae, fiber-producing species are common, including hemp (*Cannabis*), ramie or China grass (*Boehmeria*), and bast-fiber (*Urtica*), the last infamous due to the stinging hairs present on many species. Ulmaceae and Celtidaceae are small families of north temperate forest trees, many of which provide useful timbers, including elm (*Ulmus*), hackberries (*Celtis*), and *Zelkova*. *See* ELM; FIBER CROPS; FIG; HACKBERRY; HEMP; MULBERRY; RAMIE.   Mark W. Chase

Rose curve

A type of plane curve that consists of loops (leaves, petals) emanating from a common point and that has a roselike appearance. Taking the common point O as the pole of a polar coordinate system (see **illus.**), these curves have equations of the form $\rho = a \cdot \sin n\theta$, where $a > 0$ and n is a positive integer (also
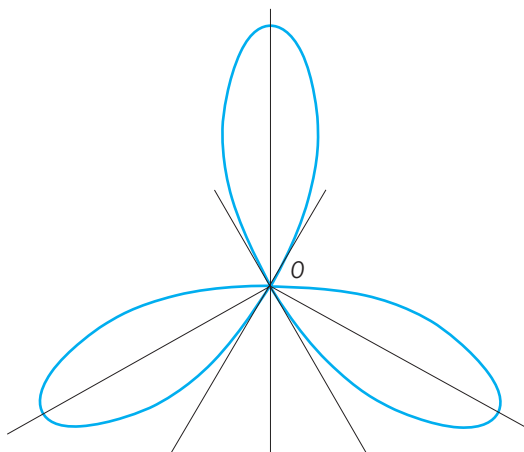

Diagram of rose curve.

$\rho = a \cdot \cos n\theta$, with a different choice of the initial line of the coordinate system). The curve is a circle of diameter a for $n = 1$. It has n or $2n$ leaves, according to whether n is an odd or even integer, respectively. The area enclosed by one leaf is $\pi a^2/n$; therefore the entire area of the three-leaved rose $\rho = a \sin 3\theta$ is $\pi a^2/4$, and that of the four-leaved rose $\rho = a \sin 2\theta$ is $\pi a^2/2$. The lemniscate is sometimes called a two-leaved rose, though its equation, $\rho^2 = a^2 \cos 2\theta$, is not of the form given above. *See* LEMNISCATE OF BERNOULLI.   Leonard M. Blumenthal

Rosemary

Rosmarinus officinalis, a member of the mint family, grown for its highly aromatic leaves and as an ornamental. Rosemary is an evergreen and perennial which can live as long as 20 years under favorable conditions. Although many varieties exist, *R. officinalis* is the only species. Most varieties are suitable for culinary use, although some (such as Pine Scented) contain high levels of terpenes and have a turpentinelike scent. Other varieties such as Forestii, Collingwood Ingram, and Lockwood de Forest have been developed as ornamentals which have large leaves and showy flowers. *See* LAMIALES.

Growth habit varies greatly. Some prostrate varieties form mats reaching only 6 in. (15 cm) in height, while other upright types, such as Erectus, become large shrubs as tall as 15 ft (4.5 m). Leaf size and shape also vary from short and needlelike to long flat blades. Plants commercially cultivated for their leaves or oil are generally upright in habit, which aids in mechanical harvesting.

Rosemary is native to the Mediterranean area, and much of the world production is harvested from wild plants growing there. Though it is widely cultivated mainly in France, Spain, and California, it will grow in most temperate areas not subject to hard frosts.

Rosemary can be propagated both by rooted cuttings and by seeds. In both instances the plants are usually started in flats and transplanted. Germination of rosemary seeds is low (only 5–20%), and it is slow

and erratic. Because of this and the genetic variability of seeds, they are little used for propagation.

Rosemary oil has been found to contain chemicals such as rosmanol which inhibit oxidation and bacterial growth. Rosemary is used in poultry seasoning and for flavoring soups and vegetables. *See* SPICE AND FLAVORING.   Seth Kirby

Bibliography. S. Arctander, *Perfume and Flavor Materials of Natural Origin*, 1960; L. H. Bailey, *Manual of Cultivated Plants*, rev. ed., 1975; F. Rosengarten, *The Book of Spices*, 1969.

Rosidae

A former widely recognized subclass of Magnoliophyta (angiosperms or flowering plants). Deoxyribonucleic acid (DNA) and morphological studies have demonstrated that this group includes many more families than previously thought (for example, a large number of families previously referred to the subclass Dilleniidae). Whereas formal reclassification awaits further evidence, it is clear that there is a large group that can be considered rosid. Approximately 136 families (roughly 30% of the total angiosperms) in 11 orders are included in the rosids. With the asterids (often recognized as Asteridae), the rosids represent the two most advanced lines of dicots.

The primary distinguishing features of rosids are their diplostemenous flowers (anthers in two whorls, each whorl typically numbering the same as the perianth parts, although in some groups the number of stamens has increased) with unfused parts, except for the carpels (asterids have haplostemenous flowers with fused parts, typically with the stamens fused at least to the bases of the petals). Rosids have ovules with two integuments (that is, they are bitegmic, as opposed to the unitegmic asterids) and a several-layered nucellus (that is, they are crassinucellate, as opposed to the tenuinucellate asterids, which have a single-layered nucellus). Ellagic acid is almost universally present (but this character occurs also in the asterid orders Cornales and Ericales). There is nuclear endosperm formation (cellular in asterids), a reticulate pollen layer, mucilaginous leaf epidermis, and generally simple perforation vessel end-walls in the wood (although the orders Malpighiales, Celastrales, and Oxalidales have more advanced scalariform perforations). The largest rosid orders are Malpighiales (16,100 species), Fabales (17,500), and Myrtales (10,600). *See* APIALES; CELASTRALES; CORNALES; EUPHORBIALES; FABALES; GERANIALES; HALORAGALES; LINALES; MAGNOLIOPSIDA; MYRTALES; PLANT KINGDOM; PODOSTEMALES; POLYGALALES; PROTEALES; RAFFLESIALES; RHIZOPHORALES; ROSALES; SANTALALES; SAPINDALES.   Mark W. Chase

Rosin

A brittle resin ranging in color from dark brown to pale lemon yellow and derived from the oleoresin of pine trees. Rosin is insoluble in water, but soluble in most organic solvents. It softens at about 80–90°C (180–190°F). Rosin consists of about 90% resin acids and about 10% neutral materials such as anhydrides, sterols (mainly sitosterol), and diterpene aldehydes and alcohols. The resin acids are diterpene acids of the abietic type—abietic (**1**), neoabietic, palustric, and dehydroabietic—and the pimaric type—pimaric (**2**), isopimaric, and sandaracopimaric. Isomerization
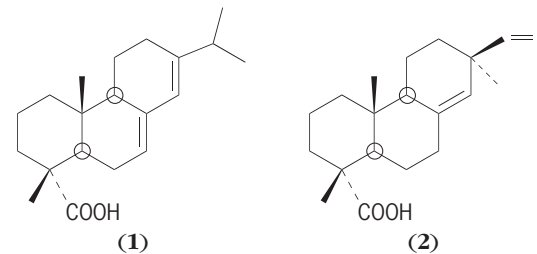
among these acids is facile, and rosin composition depends upon its origin and prior processing. However, the abietic resin acids predominate over the pimaric type by a 4:1 ratio.

Rosin is obtained by wounding living trees and collecting the exudate (gum rosin), by extraction of pine stumps (wood rosin), and as a by-product from the kraft pulping process (sulfate or tall oil rosin).

Most applications of rosin involve the use of a chemically modified form. Earlier destructive distillation to viscous rosin oils has given away to hydrogenation, disproportionation (dehydrogenation), esterification, polymerization, salt formation, or reaction with maleic anhydride or formaldehyde.

The largest single use of rosin is in sizing paper to control water absorption, an application in which fortified rosin (the maleic anhydride adduct) is important. Earlier use of rosin in laundry soaps evolved into application of rosin soaps as emulsifying and tackifying agents in synthetic rubber manufacture. Glycerol and pentaerythritol esters (ester gum) are used in coatings and lacquers. Other rosin uses include adhesives, printing inks, and chewing gum. *See* PAPER; PINE TERPENE; TALL OIL.   Irving S. Goldstein

Rotary engine

Internal combustion engine that duplicates in some fashion the intermittent cycle of the piston engine, consisting of the intake-compression-power-exhaust cycle, wherein the form of the power output is directly rotational.

Four general categories of rotary engines can be considered: (1) cat-and-mouse (or scissor) engines, which are analogs of the reciprocating piston engine, except that the pistons travel in a circular path; (2) eccentric-rotor engines, wherein motion is imparted to a shaft by a principal rotating part, or rotor, that is eccentric to the shaft; (3) multiple-rotor engines, which are based on simple rotary motion of two or more rotors; and (4) revolving-block engines, which combine reciprocating piston and rotary motion. Some of the more interesting engines of each type are discussed in this article.
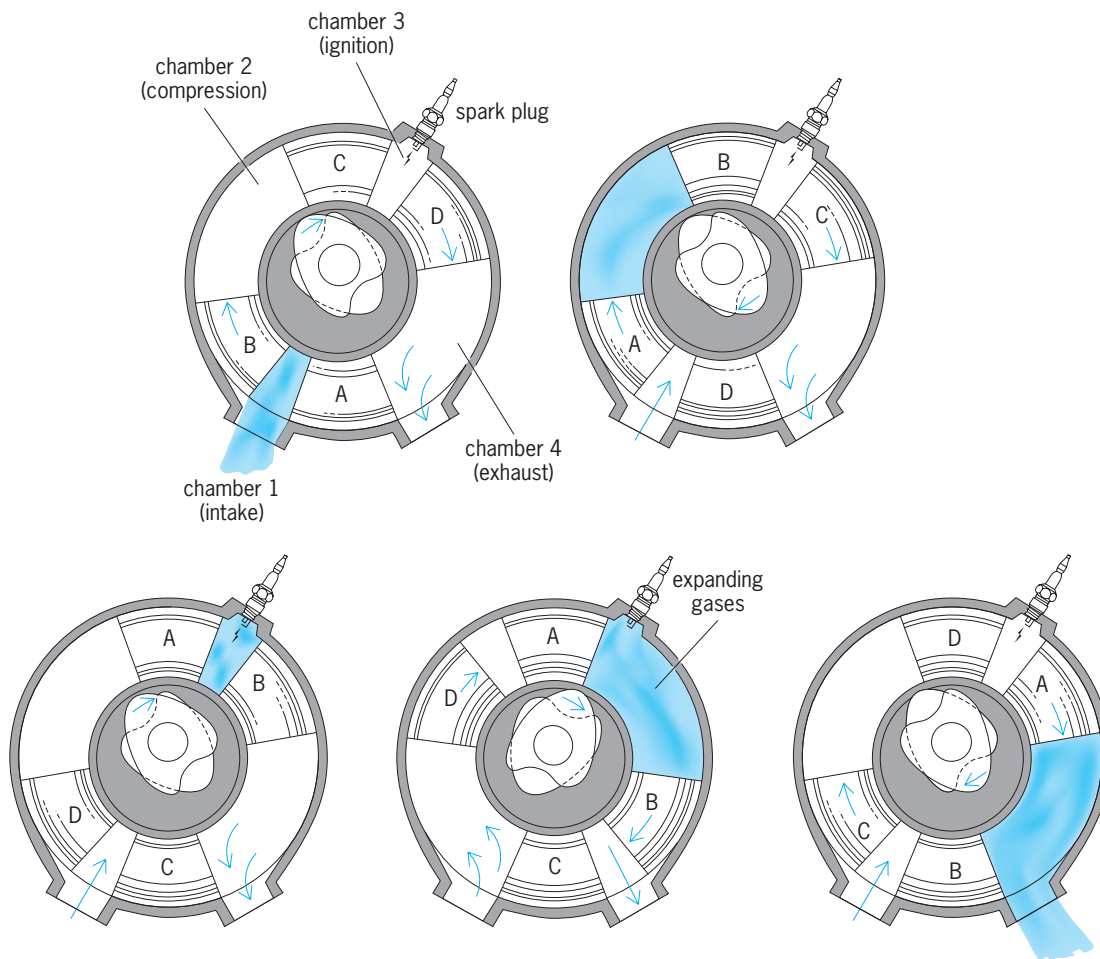
Fig. 1. Intake-compression-power-exhaust cycle of the Tschudi engine.

Cat-and-mouse engines. Typical of this class is the engine developed by T. Tschudi, the initial design of which goes back to 1927. The pistons, which are sections of a torus, travel around a toroidal cylinder. The operation of the engine can be visualized with the aid of **Fig. 1**, where piston A operates with piston C, and B with D. In chamber 1 a fresh fuel-air mixture is initially injected while pistons C and D are closest together. During the intake stroke the rotor attached to pistons B and D rotates, thereby increasing the volume of chamber 1. During this time the A-C rotor is stationary. When piston D reaches its topmost position, the B-D rotor becomes stationary, and A and C rotate so that the volume of chamber 1 decreases and the fuel-air mixture is compressed.

When the volume of chamber 1 is again minimal, both rotors move to locate chamber 1 under the spark plug, which is fired. The power stroke finds piston D moving away from piston C, with the A-C rotor again locked during most of the power stroke. Finally, when piston D has reached bottom, the B-D rotor locks, the exhaust port has been exposed, and the movement of piston C forces out the combustion product gas. Note that four chambers exist at any time, so that at each instant all the processes making up the four-strike cycle (intake-compression-power-exhaust) are occurring.

The motion of the rotors, and hence the pistons, is controlled by two cams which bear against rollers attached to the rotors. The cams and rollers associated with one of the rotors disengage when it is desired to stop the motion of that rotor. The shock loads associated with starting and stopping the rotors at high speeds may be a problem with this engine, and lubrication and sealing problems are characteristic of virtually all the engines discussed herein. However, the problem of fabricating toroidal pistons does not appear to be as formidable as was once believed.

An engine similar to the Tschudi in operation is that developed by E. Kauertz. In this case, however, the pistons are vanes which are sections of a right circular cylinder. Another difference is that while one set of pistons is attached to one rotor so that these two pistons rotate with a constant angular velocity, the motion of the second set of pistons is controlled by a complex gear-and-crank arrangement so that the angular velocity of this second set varies. In this manner, the chambers between the pistons can be made to vary in volume in a prescribed manner. Hence, the standard piston-engine cycle can be duplicated. Kauertz tested a prototype which was found to run smoothly and to deliver 213 hp (160 kW) at 4000 rpm. Here again, however, the varying angular velocity of the second set of pistons must produce

inertia effects that will be absorbed by the gear-and-crank system. At high speeds over extended periods, problems with this system are likely to be encountered.

An advanced version of the cat-and-mouse concept called the SODRIC engine has been developed by K. Chahrouri. Unlike the Tschudi engine, in which the four processes of intake-compression-power-exhaust are distributed over 360° of arc, the SODRIC engine performs these same four processes in 60° of arc. Hence, six power strokes per revolution are achieved, resulting in very substantial improvements in engine performance parameters. For example, Chahrouri has estimated that 225 hp (168 kW) can be achieved at only 1000 rpm using an engine having a toroid radius of 8 in. (20 cm) and 1-in.-radius (25-mm) pistons. Chahrouri has also improved upon the method by which alternate acceleration-deceleration of the pistons is achieved, and power is transmitted to the output shaft by using noncircular gears.

The "cat-and-mouse" and "scissors" characterizations of these engines should be clear once the picture of pistons alternately running away from, and catching up to, each other is firmly in mind. Other engines of this type, including the Maier, Rayment, and Virmel designs, differ principally in the system used to achieve the cat-and-mouse effect.

It should be noted that since the length of the power stroke is readily controlled in these engines, good combustion efficiencies (close to complete combustion) should be attainable.

Eccentric-rotor engines. The rotary engine which has received by far the greatest development to date is the Wankel engine, an eccentric-rotor type. The basic engine components are pictured in **Fig. 2**. Only two primary moving parts are present: the rotor and the eccentric shaft. The rotor moves in one direction around the trochoidal chamber, which contains peripheral intake and exhaust ports.

The operation of this engine can be visualized with the aid of Fig. 2. The rotor divides the inner volume into three chambers, with each chamber the analog of the cylinder in the standard piston engine. Initially, chamber AB is terminating the intake phase and commencing its compression phase, while chamber BC is terminating its compression phase and chamber CA is commencing its exhaust phase. As the rotor moves clockwise, the volume of chamber AB approaches a minimum. When the volume of chamber AB is minimal, the spark plug fires, initiating the combustion phase in that chamber. As combustion continues, the point is reached where the exhaust port is exposed, and the products of combustion are expelled from chamber AB.
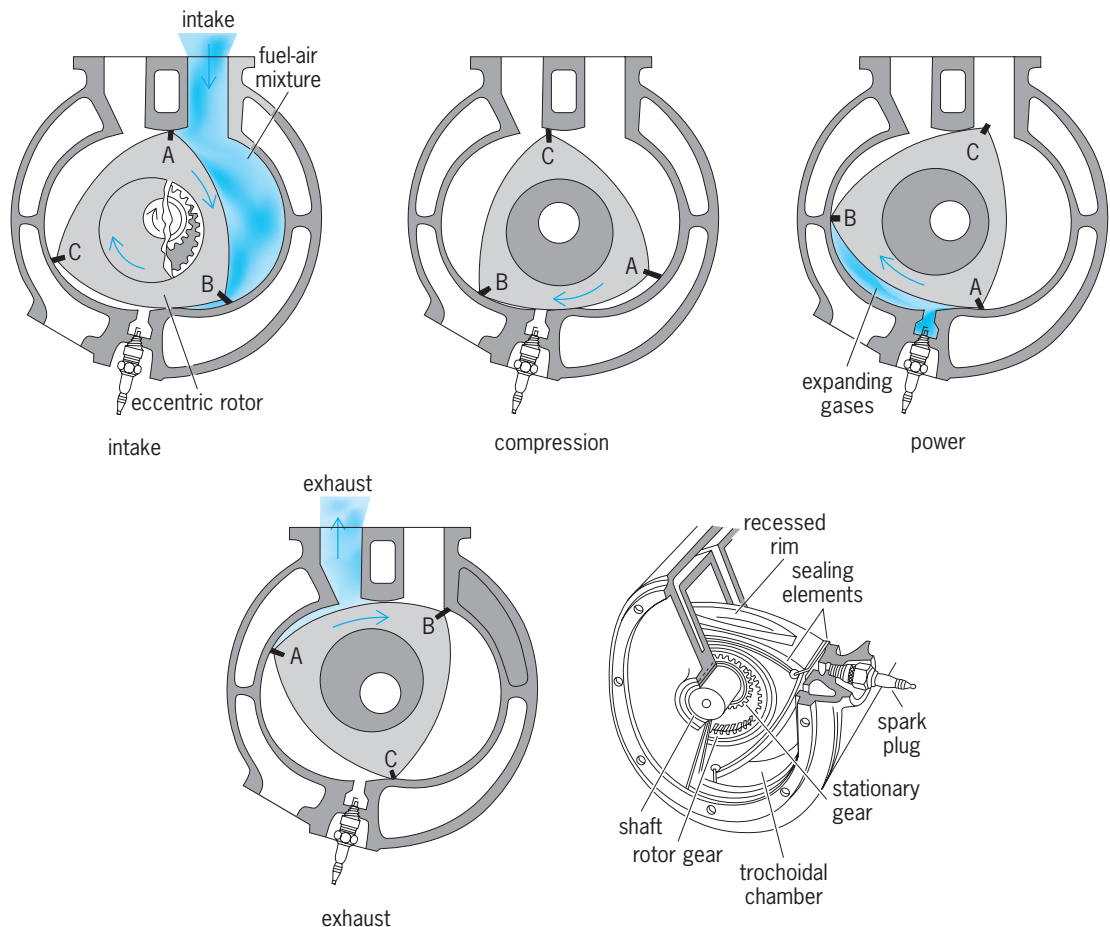
To increase the chamber volumes, each segment

Fig. 2. Operation and basic components of the Wankel engine.

of the rotor rim is recessed (Fig. 2). During the combustion-expansion phase, unburned gas tends to flow at high velocity away from the combustion zone toward the opposite corner. As a result, this engine has a tendency to leave a portion of the charge unburned, similar to the problem encountered in ordinary piston engines. In addition to reducing the engine performance, this unburned gas is a source of air pollution. Efforts have been directed toward increasing the turbulence in each chamber, thereby improving the mixing between the burned and unburned gases, leading to better combustion efficiency.

However, Wankel engines have demonstrated impressive advantages when contrasted with standard engines. Some of these are listed below.

1. The Wankel has superior power-to-weight ratio; that is, it generally produces more or at least comparable horsepower per pound of engine weight when compared with conventional piston engines.

2. To increase power output, additional rotor-trochoidal chamber assemblies can readily be added, which occupy relatively little space and add little weight. In piston engines, cylinder volumes must be increased, leading to substantial increases in weight and installation space.

3. The rotor and eccentric shaft assembly can be completely balanced; since they usually rotate at constant velocity in one direction, vibration is almost completely eliminated and noise levels are markedly reduced.

4. As with the cat-and-mouse engines, the intake and exhaust ports always remain open, that is, gas flow into and from the engine is never stopped, so that surging phenomena and problems associated with valves which open and close are eliminated.

5. Tests indicate that Wankel engines can run on a wide variety of fuels, including ordinary gasoline and cheaper fuels as well.

6. After considerable development, reasonably effective sealing between the chambers has been achieved, and springs maintain a light pressure against the trochoidal surface.

7. The Wankel has so few parts, relative to a piston engine, that in the long run it will probably be cheaper to manufacture.

The initial application of the Wankel engine as an automotive power plant occurred in the NSU Spider. In the early 1970s, however, the Japanese automobile manufacturer Mazda began to use Wankel engines exclusively. However, relatively high pollutant emissions, coupled with low gasoline mileage for automobiles of this size and weight, resulted in poor sales in the United States. Mazda ceased marketing Wankel-powered automobiles in the United States in the mid-1970s. Several American automobile manufacturers have experimented with Wankel-powered prototypes, but no production vehicles have emerged.

The Wankel engine is being used as a marine engine and in engine–electric generator installations, where its overall weight and fuel consumption have proved to be superior to those of a diesel engine or gas turbine generating equivalent power. Other
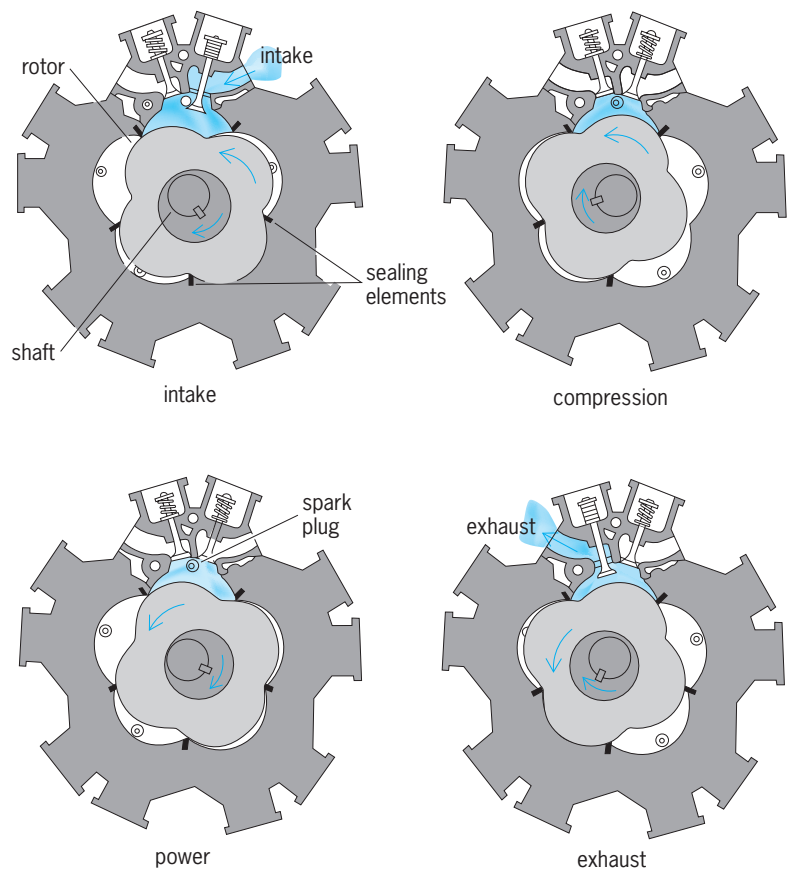

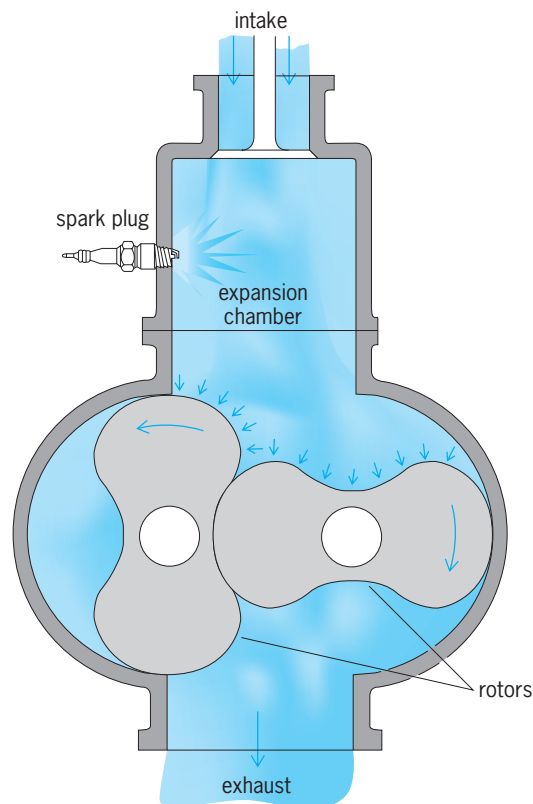Fig. 3. Renault-Rambler engine.
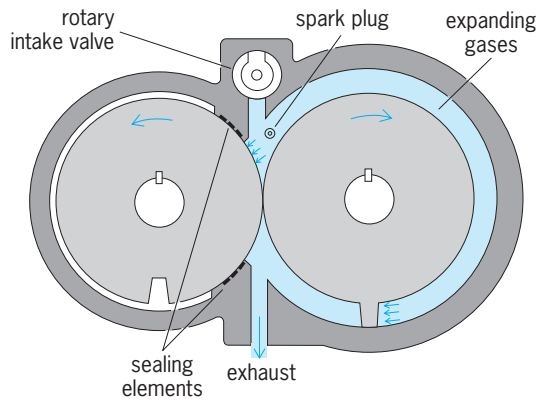

Fig. 4. Simple multirotor engine.

Fig. 5. Unsin engine, with two circular rotors.

projected applications include lawnmower and chainsaw engines. This wide range of applications is made possible by the fact that almost any size of Wankel engine is feasible.

An engine conceptually equivalent to the Wankel was developed jointly by Renault, Inc., and the American Motors Corp. It is sometimes called the Renault-Rambler engine. In this case, however, the rotor consists of a four-lobe arrangement, operating in a five-lobe chamber (**Fig. 3**). When a lobe moves into a cavity, which is analogous to the upward motion of the piston in the cylinder, the gas volume decreases, resulting in a compression process. The operation of this engine is detailed in Fig. 3. The fact that each cavity has two valves (intake and exhaust) represents a significant drawback. However, sealing between chambers may be simpler than in the Wankel; since each cavity acts as a combustion chamber, heat is evenly distributed around the housing, resulting in little thermal distortion.

It can be concluded that engines of the eccentric-rotor type are an integral part of the internal combustion engine scene. Their inherent simplicity, coupled with their advanced state of development, make them attractive alternatives to the piston engine in a number of applications.

Multirotor engines. These engines operate on some form of simple rotary motion. A typical design, shown in **Fig. 4**, operates as follows. A fuel-air mixture enters the combustion chamber through some type of valve. No compression takes place; rather, a spark plug ignites the mixture which burns in the combustion chamber, with a consequent increase in temperature and pressure. The hot gas expands by pushing against the two trochoidal rotors. The eccentric force on the left-hand rotor forces the rotors to rotate in the direction shown. Eventually, the combustion gases find their way out the exhaust.

The problems associated with all engines of this type are principally twofold: The absence of a compression phase leads to low engine efficiency, and sealing between the rotors is an enormously difficult problem. One theoretical estimate of the amount of work produced per unit of heat energy put into the engine (by the combustion process), called the thermal efficiency, is only 4%.

The Unsin engine (**Fig. 5**) replaces the trochoidal rotors with two circular rotors, one of which has a single gear tooth upon which the gas pressure acts. The second rotor has a slot which accepts the gear tooth. The two rotors are in constant frictional contact, and in a small prototype engine sealing apparently was adequate. The recommendation of its inventor was that some compression of the intake charge be provided externally for larger engines.

The Walley and Scheffel engines employ the principle of the engine in Fig. 4, except that in the former, four approximately elliptical rotors are used, while in the latter, nine are used. In both cases the rotors turn in the same clockwise sense, which leads to excessively high rubbing velocities. (The rotors are in contact to prevent leakage.) The Walter engine uses two different-sized elliptical rotors.

Revolving-block engines. These engines combine reciprocating piston motion with rotational motion of the entire engine block. One engine of this type is the Mercer (**Fig. 6**). In this case two opposing pistons operate in a single cylinder. Attached to each piston are two rollers which run on a track that consists of two circular arcs. When the pistons are closest together, the intake ports to the chambers behind the pistons are uncovered, admitting a fresh charge (Fig. 6a). At this moment a charge contained between the pistons has achieved maximum compression, and the spark plug fires. The pistons separate as combustion takes place between them, which results in a compression of the gases behind the pistons (Fig. 6b). However, the pistons moving apart force the rollers to move outward as well. This latter motion can only occur if the rollers run on their circular track, which consequently forces the entire engine block to rotate. When the pistons are farthest apart, the exhaust ports are uncovered and the combustion gases purged (Fig. 6c). At this same time the compressed fresh charge behind the pistons is transferred to the region between the pistons to prepare for its recompression and combustion, which must occur because of the continuing rotation of the block.

No doubt some of the fresh charge is lost to the exhaust during the transfer process. Stresses on the roller assembly and cylinder walls are likely to be
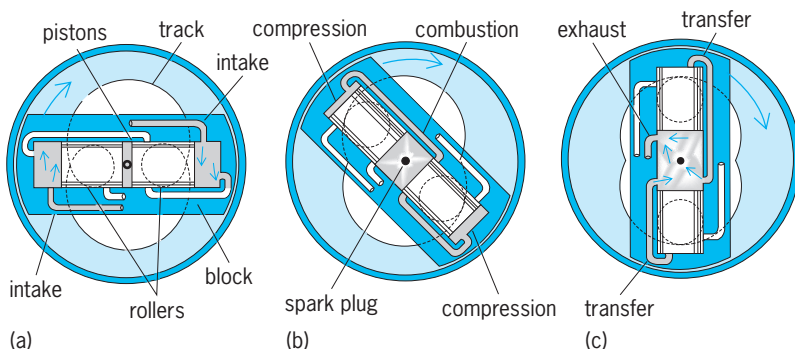

Fig. 6. Cycle of the Mercer engine. (*a*) Intake of fresh charge. (*b*) Compression of charge, following firing of spark plug. (*c*) Exhaust of combustion gases and transfer to compressed fresh charge.

quite high, which poses some design problems. Cooling is a further problem, since cooling of the pistons is difficult to achieve in this arrangement.

One the other hand, the reciprocating piston motion is converted directly to rotary motion, in contrast with the connecting rod–crank arrangement in the conventional piston engine. Also, no flywheel should be necessary since the entire rotating block acts to sustain the rotary inertia. Vibration will also be minimal.

The Selwood engine is similar in operation, except that two curved pistons 180° opposed run in toroidal tracks. This design recalls the Tschudi cat-and-mouse engine, except that the pistons only travel through 30° of toroidal track. This motion forces the entire block to rotate. The Leath engine has a square rotor with four pistons, each 90° apart, with a roller connected to each. As in the Mercer engine, the reciprocating motion of the pistons forces the rollers to run around a trochoidal track which causes the entire block to rotate. The Porshe engine uses a four-cylinder cruciform block. Again, rollers are attached to each of the four pistons. In this arrangement power is achieved on the inward strokes of the
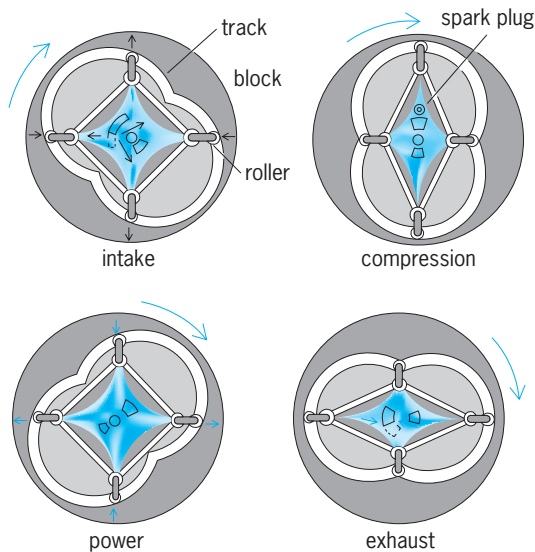

Fig. 7. Cycle of the Rajakaruna engine, showing distortion of the four-sided chamber as the surrounding housing rotates.
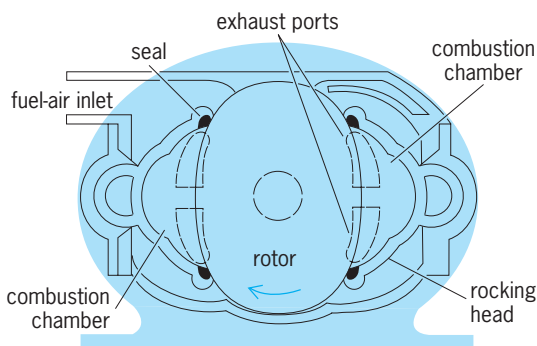

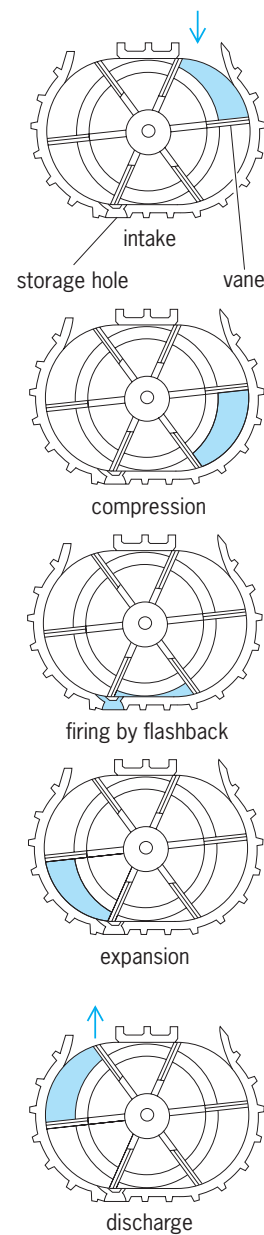Fig. 8. Walker engine, with elliptical rotor and two C-shaped rocking.

Fig. 9. Cycle of one pocket of combustion gases in the Heydrich vane-type rotary engine.

piston. Finally, the Rajakaruna engine (**Fig. 7**) uses a combustion chamber whose sides are pin-jointed together at their ends. Volume changes result from distortion of the four-sided chamber as the surrounding housing, which contains a trochoidal track, rotates. The huge pins are forced against the track. As usual, cooling and lubrication problems will be encountered with this engine, as will excessive wear of the hinge pins and track.

The Ma-Ho engine, invented by G. Hofmann, uses four cylinders welded concentrically around a central shft. As the pistons oscillate in the cylinders as a result of the intake-compression-power-exhaust processes undergone by the fuel-air mixture, the pistons rotate a barrel cam to which they are connected by cam followers. Hence, the entire block, including the central shaft, is forced to rotate.
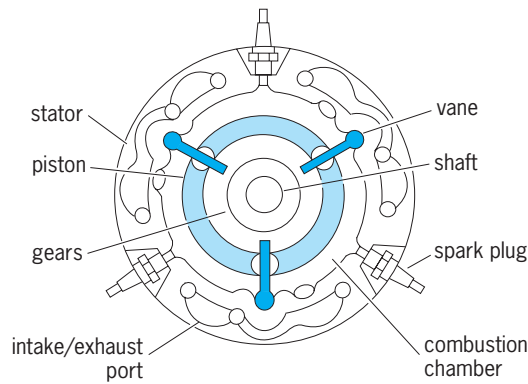
Fig. 10. Leclerc-Edmon-Benstead engine.

Engines of other types. Although the vast majority of rotary engines fall into one of the categories discussed above, several ingenious designs which do not are worthy of mention.

The Walker engine (**Fig. 8**) involves an elliptical rotor which rotates inside a casing containing two C-shaped rocking heads. The fuel-air mixture is drawn into combustion chambers on each side of the rotor; mixture cut-off occurs when the rotor is in the vertical position. As the rotor turns, the mixture undergoes compression, and combustion is initiated by spark plugs. Rotor momentum coupled with the expansion of the combustion gases forces the rotor to continue turning. Compensation for seal wear is made by adjusting the rocking heads closer to the rotor.

The Heydrich engine (**Fig. 9**) is a vane-type rotary engine which utilizes a small hole to "store" a quantity of high-temperature combustion gases from a previous firing. This gas is then used to ignite the subsequent fuel-air charge. The first charge is ignited by a glow plug. Floating seals make contact with the chamber wall.

The Leclerc-Edmon-Benstead engine (**Fig. 10**) has three combustion chambers defined by stationary vanes, a cylindrical stator, and end flanges. An output shaft passing through the chamber is geared to a circular, eccentrically mounted piston. As the chambers are fired sequentially, the movement of the piston forces the shaft to rotate. Slots in the flange control porting of the intake and exhaust gases. *See* COMBUSTION CHAMBER; DIESEL CYCLE; DIESEL ENGINE; GAS TURBINE; INTERNAL COMBUSTION ENGINE; OTTO CYCLE.   Wallace Chinitz

Bibliography. R. F. Ansdale, Rotary combustion engines, *Auto. Eng.*, vol. 53, no. 13, 1963, and vol. 54, nos. 1 and 2, 1965; W. Chinitz, Rotary engines, *Sci. Amer.*, vol. 220, no. 2, 1969; H. E. Dark, *The Wankel Rotary Engine*, 1974; W. Froede, *The Rotary Engine and the NSU Spider*, Soc. Automot. Eng. Pap. 650722, October, 1965; K. Matsumoto et al., *The Effects of Combustion Chamber Design and Compression Ratio on Emissions, Fuel Economy and Octane Number Requirement*, SAE Pap. 770193, 1977; R. H. Thring, Gasoline engines and their future, *Mech. Eng.*, vol. 15, no. 10, 1983; R. Wakefield, Revolutionary engines, *Road Track*, vol. 18, no. 3, 1966; F. Wankel, in R. F. Ansdale (ed.), *Rotary Piston Machines: Classification of Design Principles for Engines, Pumps and Compressors*, 1965.

Rotational motion

The motion of a rigid body which takes place in such a way that all of its particles move in circles about an axis with a common angular velocity; also, the rotation of a particle about a fixed point in space. Rotational motion is illustrated by (1) the fixed speed of rotation of the Earth about its axis; (2) the varying speed of rotation of the flywheel of a sewing machine; (3) the rotation of a satellite about a planet, in which both the speed of rotation and the distance from the center of rotation may vary; (4) the motion of an ion in a cyclotron, where the angular speed of rotation remains constant, but the radius of the circular motion increases; and (5) the motion of a pendulum, in which case the particles describe harmonic motion along a circular arc.

This discussion of rotational motion is limited to circular motion such as is exhibited by the first and second examples. For information concerning the other examples *see* CELESTIAL MECHANICS; EARTH ROTATION AND ORBITAL MOTION; HARMONIC MOTION; PARTICLE ACCELERATOR; PENDULUM.

Circular motion is a rotational motion in which each particle of the rotating body moves in a circular path about an axis. The motion may be uniform, that is, with constant angular velocity, or nonuniform, with changing angular velocity.

Uniform circular motion. The speed of rotation, or angular velocity, remains constant in uniform circular motion. In this case, the angular displacement θ experienced by the particle or rotating body in a time t is $\theta = \omega t$, where ω is the constant angular velocity.

Nonuniform circular motion. A special case of circular motion occurs when the rotating body moves with constant angular acceleration. If a body is moving in a circle with an angular acceleration of α radians/s^2, and if at a certain instant it has an angular velocity ω_0, then at a time t seconds later, the angular velocity may be expressed as $\omega = \omega_0 + \alpha t$, and the angular displacement as $\theta = \omega_0 t + \frac{1}{2}\alpha t^2$. *See* ACCELERATION; VELOCITY.

Banking of curves. When a car travels around a horizontal curve on a highway, the path is a circular arc of radius R, where R is the radius of curvature of the roadway. In order to have the car move in this circular arc, a horizontal external force must be applied to give the car an acceleration perpendicular to its path, that is, toward the center of rotation. This force must equal Mv^2/R, where M is the mass of the car and v its speed. This centripetal force is supplied by the friction between the tires and the road. If the force of friction is not great enough to produce this acceleration, the inertia of the car will tend to make it continue with its speed in a straight line, tangent to the road rather than around the curve, and this will cause the car to slide off the road. *See* CENTRIPETAL FORCE.
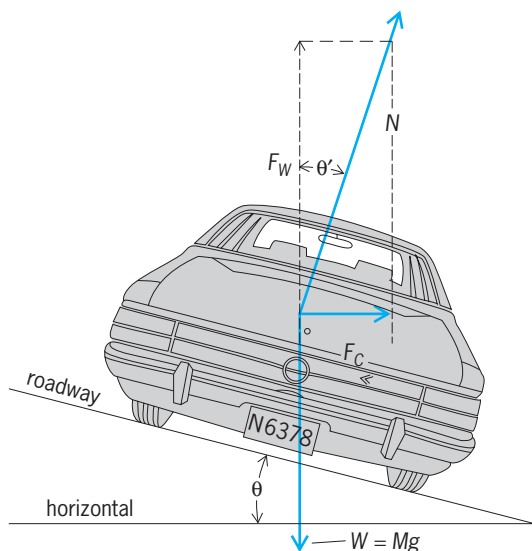
Fig. 1. Banking of a curve.

To reduce the probability of skidding roadways are customarily banked as illustrated in **Fig. 1**, which shows a car of mass M going away from the reader with a speed v and making a right-hand turn along an arc of radius R. The roadway must exert a vertical force F_W upward, equal and opposite to the weight $W = Mg$ of the car (g is the acceleration of gravity), and a horizontal centripetal force $F_C = Mv^2/R$ to make the car move in a circular arc. The net force N of the road on the car is the vector sum of these two forces.

From the diagram, it can be seen that the angle θ' which N makes with the vertical is given by Eq. (1). If

$$\tan \theta' = \frac{F_C}{F_W} = \frac{Mv^2/R}{Mg} = \frac{v^2}{gR} \qquad (1)$$

this angle θ' is equal to the bank angle θ of the road, the force N of the road on the car is perpendicular to the roadway, and there will be no tendency to skid. Equation (1) shows that the correct bank angle is proportional to the square of the speed and inversely proportional to the radius of the curve. For a given curve, there is no correct bank angle for all speeds; thus roadways are banked for the average speed of traffic. Bank angle enters into the design of railroads and into the banking of an airplane when it executes a turn.

Work and power relations. A rotating body posesses kinetic energy of rotation which may be expressed as $T_{\text{rot.}} = {}^1\!/_2 I\omega^2$, where ω is the magnitude of the angular
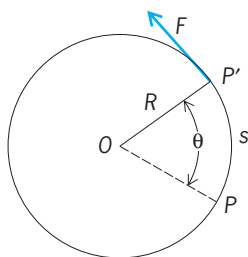

Fig. 2. Work in angular motion.

velocity of the rotating body and I is the moment of inertia, which is a measure of the opposition of the body to angular acceleration. The moment of inertia of a body depends on the mass of a body and the distribution of the mass relative to the axis of rotation. For example, the moment of inertia of a solid cylinder of mass M and radius R about its axis of symmetry is ${}^1\!/_2 MR^2$.

To impart kinetic energy to a rotating body, work must be done. In **Fig. 2** there is represented a solid cylinder of mass M and radius R, capable of rotation without friction about an axis perpendicular to the plane of the page through O. By means of a cord wrapped around the cylinder, a constant force F is applied, thus imparting angular acceleration to the cylinder. If the cylinder is originally at rest and the force F acts through a distance $s = R\theta$ equal to the arc PP', thus rotating the cylinder through the angle θ, the work W done is $W = Fs = FR\theta = L\theta$, where $L = FR$ is called the torque or moment of force. The action of this torque L is to produce an angular acceleration α according to Eq. (2), where $I\omega$, the

$$L = I\alpha = I\frac{d\omega}{dt} = \frac{d}{dt}(I\omega) \qquad (2)$$

product of moment of inertia and angular velocity, is called the angular momentum of the rotating body. This equation points out that the angular momentum $I\omega$ of a rotating body, and hence its angular velocity ω, remains constant unless the rotating body is acted upon by a torque. Both L and $I\omega$ may be represented by vectors.

It is readily shown by Eq. (3) that the work done by

$$W = L\theta = I\alpha\theta = I\alpha\,{}^1\!/_2(\alpha t^2)$$
$$= {}^1\!/_2 I(\alpha t)^2 = {}^1\!/_2 I\omega^2 \qquad (3)$$

the torque L acting through an angle θ on a rotating body originally at rest is exactly equal to the kinetic energy of rotation, because, for the case at hand, $\theta = {}^1\!/_2(\alpha t^2)$ and $\omega = \alpha t$.

Power is defined as rate of doing work, and the power P in rotational motion is given by Eq. (4).

$$P = \frac{dW}{dt} = \frac{d}{dt}(L\theta) = L\frac{d\theta}{dt} = L\omega \qquad (4)$$

See ANGULAR MOMENTUM; MOMENT OF INERTIA; MOTION; POWER; RIGID-BODY DYNAMICS; TORQUE; WORK.   Carl E. Howe; R. J. Stephenson

Rotifera

A phylum of microscopic, mainly free-living aquatic animals, characterized primarily by an anterior ciliary apparatus, the corona. When the cilia of the corona are in motion the animals appear to have a pair of rapidly rotating wheels on their heads, thus the name "Rotifera," Latin for "wheel bearers." Rotifers were first observed by Anton Leeuwenhoek more than 300 years ago; because they are easy to find and of a convenient size for observation with a
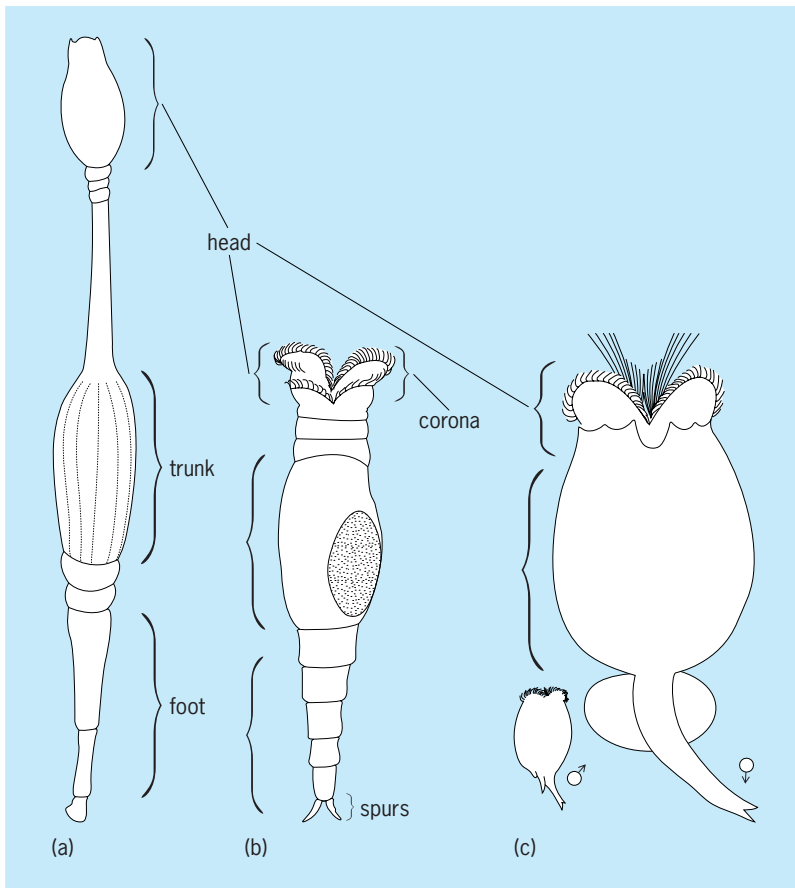
Fig. 1. Common species from the three rotifer groups. (*a*) The seisonid *Seison nebaliae*. (*b*) The bdelloid *Philodina roseola*. (*c*) The monogonont *Brachionus plicatilis*.

be a fourth group of Rotifera. Rotifera and Acanthocephala are often referred to as Syndermata, which in turn is part of an assemblage of microscopic animals including Micrognathizoa and Gnathostomulida called Gnathifera, because of the shared characteristic of jawlike structures made of sclero-protein.

Morphology. The body of a rotifer is usually divided into three parts: the head, trunk, and foot (**Fig. 1**). The head carries the corona, the mouth, and mastax of the digestive system, as well as the central ganglion (brain) of the nervous system. Most organs are located in the trunk, including the stomach, intestine, cloaca, anus, and gastric glands of the digestive system, the simple excretory system, and the reproductive organs. The excretory system consists primarily of paired nephridial tubes, provided with flame bulbs and opening posteriorly either into a bladder or directly into the cloaca. The female reproductive organs consist of paired ovaries and vitellaria; the ovaries open into the cloaca and are fused in monogononts. The foot contains cement glands whose secretions enable many species to attach themselves to the substratum. Sensory organs consist of one or more eyespots, absent in some species, as well as sensory bristles, papillae, and, in most species, a dorsal antenna. The majority of rotifer tissues and organs are syncytial.

Rotifers are covered by a thin; nonchitinous cuticle and an underlying epidermal intracytoplasmic lamina, unique to rotifers and acanthocephalans, which in some species is thickened into a hard case, or lorica, over much of the exterior epidermis. Loricate rotifers often have characteristic patterns and ornamentations, which are useful for species identification. The trunk and foot of many species, particularly bdelloids, have a segmented appearance; however, rotifers are not segmented in the classical sense, nor do they represent an early stage of segmentation. Sessile monogononts and some bdelloids construct a gelatinous tube around themselves, which in some species may be extensively ornamented with small debris and waste particles. Species in the genus *Floscularia* have modified mouthparts used to concentrate particles into pellets that are then attached to the tube in highly ordered rows.

The muscular system is complex. It includes bands of circular muscles in the body wall that allow the rotifer to elongate, longitudinal muscles in the body wall that control expansion and retraction of the head and foot, muscles associated with viscera, and others. Many rotifers can retract their head or foot sections into the trunk; when disturbed, bdelloids in particular will contract into a nondescript ellipsoid.

The pharynx is modified into a structure peculiar to rotifers, the muscular mastax, which contains a masticatory apparatus, the trophus. The terms mastax and trophus are used as synonyms by many workers. Typical trophi contain seven parts: three central pieces, making up the incus, and four lateral pieces, forming the paired mallei. The three parts of the incus include a posterior supporting structure, the fulcrum, and two rami. Each malleus consists of a head, or uncus, and a handlelike piece, or manubrium.

simple microscope, they have been avidly studied by amateur naturalists ever since.

Rotifers show considerable diversity in form and structure. They are bilaterally symmetrical animals with complete digestive, excretory, nervous, and reproductive systems but lack separate respiratory and circulatory systems. Most have about 1000 nuclei and are 0.1–1.0 mm in length. They possess three characteristic morphological features: the corona; a syncytial epidermis with an intracytoplasmic lamina; and the mastax, a gizzardlike structure derived from a modified pharynx.

Our understanding of the evolutionary relationships within Rotifera and between Rotifera and related phyla have been undergoing extensive revision. Older classifications placed rotifers in the phylum Aschelminthes, an incongruous assemblage including also the Nematoda, Priapulida, Kinorhyncha, and Gastrotricha, and divided rotifers into three groups: Monogononts, Bdelloids, and Seisonids. Because of new molecular and ultrastructural information, Aschelminthes is no longer considered a valid group and Rotifera is generally considered to be a phylum, with the three groups now given class status. Only Bdelloidea and Monogononta possess all the classical rotifer features and are therefore sometimes referred to as the Eurotatoria. The obligately parasitic Acanthocephala is a closely related phylum or may
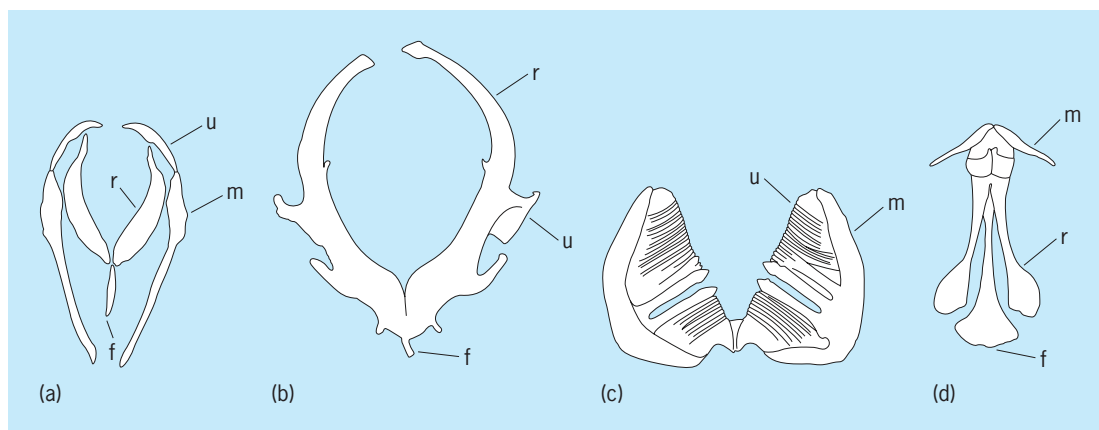
Fig. 2. Four of the 8 or 9 recognized types of rotifer mastax. (*a*) Forcipate mastax of the monogonont *Albertia vermiculus* has all parts of the mallei and incus. (*b*) Incudate mastax of the monogonont *Asplanchna brightwellii* has enlarged, pincerlike rami and reduced mallei. (*c*) Ramate mastax of the bdelloid *Macrotrachela quadricornifera* has a reduced incus and toothlike unci. (*d*) Fulcrate mastax of *Seison nebaliae* has an enlarged fulcrum and reduced mallei. Here u, unci; r, rami; f, fulcrum; m, mallei.

There is extensive diversification of mastax morphology between rotifer groups (**Fig. 2**). Some predatory rotifers have pincerlike rami and extend their trophus out of their mouth to seize prey; others have an enlarged fulcrum supporting a muscular piston that acts to suck out the contents of prey captured by toothed unci. In rotifers that feed exclusively on microphytoplankton and particulate matter, the unci are adapted for grinding and are often quite large, while other parts may be degenerate.

The most notable feature of rotifers is the corona, composed of a ventral, ciliated buccal field, which surrounds the mouth, and a more anterior circumapical band of cilia that extends around the head, enclosing an unciliated apical field. Large, powerful outer cilia of the circumapical band form the anterior trochus and posterior cingulum. As with the mastax, there is great diversity of form of the corona, with particular regions enlarged or reduced, reflecting habitat, means of locomotion, and feeding. Only a relatively small number of rotifer groups possess a corona that actually presents the impression of rotating wheels. The effect is most clearly seen among bdelloid species, where the apical field and trochus is divided into two trochal disks raised on pedicels. When the animal is disturbed or crawling, the pedicels are retracted into the head. When extended, rhythmic rapid strokes of the trochal cilia, followed by a slower recovery, give the striking appearance of a pair of spinning disks.

Feeding and locomotion. Most rotifers are free-living and feed on bacteria, algae, and organic debris. Water currents created by the beating of the corona draw particles into the mouth area, where each is examined by sensory cilia and either passed on to the mouth or rejected. Most bdelloids are of this type. Predatory rotifers may trap prey in an enlarged corona (for example, *Collotheca*) or, more commonly (for example, *Asplanchna, Dicranophorus, Philodinavus*), extrude their modified pincerlike mastax to seize prey such as ciliates, nematodes, and smaller rotifers. Some rotifers, particularly in the genus *Proales*, are parasitic on aquatic animals such as cladocerans, oligochaetes, insect larvae, or snail eggs, or on algae such as *Vaucheria* and *Volvox*. Those in the genera *Albertia, Balatro*, and *Drilophage* are nearly all obligately parasitic, feeding on body fluids of oligochaetes, leeches, or earthworms. They remain attached to their host by their mastax and have highly reduced corona and feet. Many more species are epizoic commensal or opportunistically parasitic symbionts: The only marine bdelloid species, *Zelinkiella synaptae*, is found on the surface or in the body cavity of sea cucumbers, and many species of *Embata* and *Proales* are found on the gills and legs of crustaceans and cladocerans. The only three described species of Seisonid rotifers are found on the gills and legs of the marine crustacean Nebalia; one of these, *Paraseison annulatus*, is parasitic, feeding on the body fluids and eggs of the host, while the other two feed on bacteria.

Rotifers move about by crawling in a leechlike manner (common in many bdelloids) or by swimming at speeds up to 0.5 mm/s using the water currents created by their coronal cilia. Many species in the monogonont groups Flosculariidae and Collothecidae are sessile: after hatching they rapidly settle on a substrate and produce a gelatinous tube, followed by extensive changes in the morphology of the corona (often incorrectly described as metamorphosis).

Reproduction and development. The three major groups of rotifers each have a characteristic form of reproduction (**Fig. 3**). Seisonids reproduce only sexually: males are fully developed and only slightly smaller than females. The monogonont life cycle is a complex combination of asexual and sexual forms of reproduction: Monogononts generally reproduce by parthenogenesis (females produce diploid eggs by mitosis that develop, without fertilization, into females) but in response to certain environmental cues, such as crowding, females are produced that produce haploid eggs by meiosis. If unfertilized, these eggs develop into haploid males (generally much small than females and often with a highly reduced or absent digestive tract). If a male
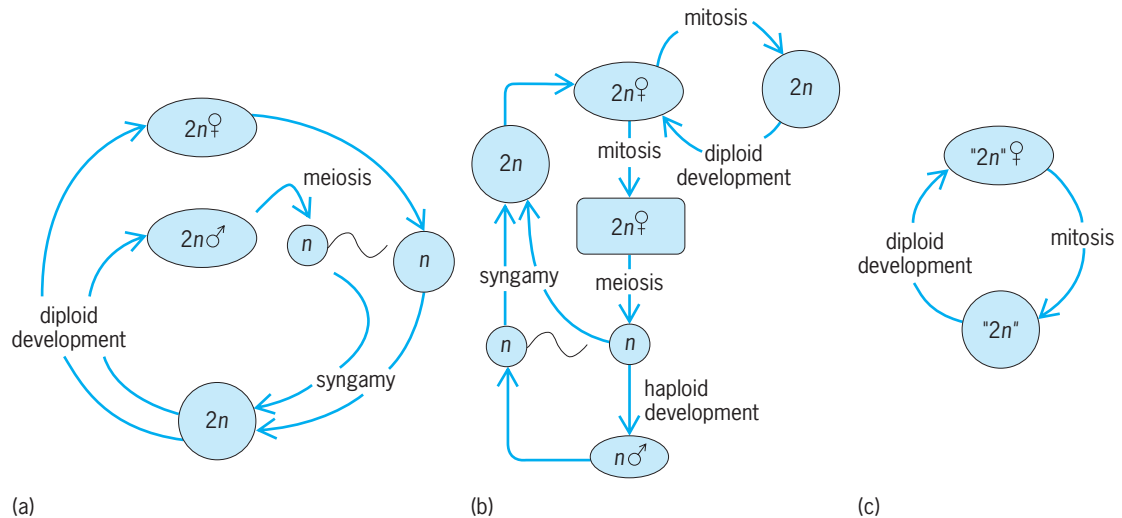
Fig. 3. Life cycles of the three rotifer groups. (*a*) Gonochoristic reproduction in seisonids. (*b*) Parthenogenesis alternating with sexual reproduction in monogononts. (*c*) Obligate parthenogenesis in bdelloids. 2*n* represents diploidy, and *n*, haploidy; bdelloids are designated "2*n*" because the concept of diploidy loses meaning in the absence of meiosis.

mates with a female and fertilizes a haploid egg, the female pumps additional nutrients and cytoplasm into the developing diploid embryo, which develops a thick shell and is called a resting egg. The resting egg can remain dormant (diapause) for years until, in response to environmental cues, it hatches into a female that will reproduce parthenogenetically. Bdelloid rotifers reproduce parthogenetically; no males or hermaphrodites are known and cytological and molecular evidence suggest that the group has evolved without sexual reproduction over many millions of years. Monogononts and Bdelloids represent two of the most successful examples of facultative and obligate parthenogenesis.

Seisonids are oviparous; monogononts and bdelloids are oviparous or viviparous depending on species. Embryonic development proceeds through a modified spiral cleavage and is eutelic; juveniles develop directly into adults. Fecundity and lifespan are highly variable between species: monogonont and bdelloid females possess approximately 10–50 oocytes and live for approximately 5–45 days.

Ecology. Rotifers are among the most widely distributed of aquatic animals. Monogononts are the most common nonarthropods in most benthic and pelagic fresh- and brackish water environments, and are a food source for copepods, insect larvae, and fish fry. Bdelloids are extremely widely distributed in all forms of fresh water, from the littoral zones of large permanent lakes to temporary pools, mud holes, water films on mosses or lichens, and similar ephemerally moist environments. They are often the most commonly encountered invertebrates in these habitats. Bdelloids can be found in hot springs, withstanding temperatures as high as $45°C$ ($113°F$) in Antarctic lakes where they can survive being frozen, and in highly acidic or alkaline streams. *See* MARINE ECOLOGY.

Most bdelloid species are capable of surviving extreme desiccation by a process known as anhydrobiosis. As water slowly evaporates from their surroundings, bdelloids contract into tight oblong balls and exude nearly all free water. In this state bdelloids are metabolically quiescent and can survive vacuum, large doses of ionizing radiation, and temperatures of -270 to $150°C$ ($-454°$ to $-320°F$). A bdelloid can remain desiccated for years and, upon rehydration, resumes its normal lifespan with little or no loss in fecundity. *See* ACANTHOCEPHALA; BDELLOIDEA; MONOGONONTA; SEISONACEA.   J. B. Jennings; David B. Mark Welch

Bibliography. L. H. Hymen, *The Invertebrates*, vol 3, 1951; Th. Nogrady (ed.), *Guides to the Identification of the Microinvertebrates of the Continental Waters of the World: Rotifera*, 1993; S. D. Rundle, A. L. Robertson, and J. M. Schmid-Araya, *Freshwater Meiofauna: Biology and Ecology*, 2002; D. G. Smith, *Pennak's Freshwater Invertebrates of the United States: Porifera to Crustacea*, 4th ed, 2001; J. H. Thorp and A. P. Covich (eds.), *Ecology and Classification of North American Freshwater Invertebrates*, 2nd ed., 2001.

Rous sarcoma

The first filterable agent (virus) known to cause a solid tumor in chickens. It was discovered in 1911 by P. F. Rous, who won the Nobel prize in 1967 for his discovery. Little advance was made in the study of this virus until technological advances made more detailed studies possible. Certain strains of the virus cause tumors in hamsters, rabbits, monkeys, and other species, thus effectively eliminating an age-old notion that tumor viruses are more or less species-specific. Although the virus cannot be demonstrated in the tumors in foreign species, transfer of the tumors back to the chick results in reappearance of the active virus. Tissue culture of chick embryos inoculated with Rous sarcoma virus produce foci of

proliferating cells composed of minute aggregations of tumor cells. It was found that the Rous virus by itself was incapable of bringing about this change but needed another closely related virus of the avian leukosis group to act as a "helper" for the production of the foci. Thus the Rous virus is known as a "defective" virus in that it is incapable of producing tumors by itself. *See* AVIAN LEUKOSIS; TISSUE CULTURE.

It is a ribonucleic acid virus and belongs to the avian leukosis group. When inoculated into chick embryos, it causes fatal hemorrhagic disease, and produces sarcomas in older chickens, pheasants, and ducklings. *See* ANIMAL VIRUS; TUMOR VIRUSES; VIRUS, DEFECTIVE.   Alice E. Moore

Bibliography. J. W. Beard (ed.), *International Conference on Avian Tumor Viruses*, Nat. Cancer Inst. Monogr. 17, 1964.

Rubber

Originally, a natural or tree rubber, which is a hydrocarbon polymer of isoprene units. With the development of synthetic rubbers having some rubbery characteristics but differing in chemical structure as well as properties, a more general designation was needed to cover both natural and synthetic rubbers. The term elastomer, a contraction of the words elastic and polymer, was introduced, and defined as a substance that can be stretched at room temperature to at least twice its original length and, after having been stretched and the stress removed, returns with force to approximately its original length in a short time.

Three requirements must be met for rubbery properties to be present in both natural and synthetic rubbers: long threadlike molecules, flexibility in the molecular chain to allow flexing and coiling, and some mechanical or chemical bonds between molecules.

A useful way of visualizing rubber structure is to consider as a model a bundle of wiggling snakes in constant motion. When the bundle is stretched and released, it tends to return to its original condition. If there were no entanglements, stretching the bundle would pull it apart. The more entanglements, the greater the tendency to recover, corresponding to cross-links in rubber. In rubbers, the characteristic property of reversible extensibility results from the randomly coiled arrangement of the long polymer chains. Upon extension the chains are elongated in a more or less orderly array. The tendency to revert to the original coiled disarray upon removal of the stress accounts for the elastic behavior. Vulcanizing rubber increases the number of cross-links and improves the properties.

Natural rubber and most synthetic rubbers are also commercially available in the form of latex, a colloidal suspension of polymers in an aqueous medium. Natural rubber comes from trees in this form; many synthetic rubbers are polymerized in this form; some other solid polymers can be dispersed in water. *See* POLYMER.

Latexes are the basis for a technology and production methods completely different from the conventional methods used with solid rubbers.

Processing

In the crude state, natural and synthetic rubbers possess certain physical properties which must be modified to obtain useful end products. The raw or unmodified forms are weak and adhesive. They lose their elasticity with use, change markedly in physical properties with temperature, and are degraded by air and sunlight. Consequently, it is necessary to transform the crude rubbers by compounding and vulcanization procedures into products which can better fulfill a specific function.

Curatives and vulcanization. After the addition of curing or vulcanizing agents to rubber, the application of heat causes cross-linking, yielding a durable product by binding the long chains together. Sulfur, the first successful curing ingredient, is the basis of vulcanization. However, various chemicals or combinations of chemicals are also capable of vulcanizing rubber. These include oxidizing agents, such as selenium, tellurium, organic peroxides, and nitro compounds, and also generators of free radicals, such as organic peroxides, azo compounds, and certain organic sulfur compounds such as the alkyl thiuram disulfides.

Sulfur alone results in very slow cure rates and develops less than optimum physical properties in the rubber. Several classes of compounds have been found which accelerate the rate of vulcanization, improve the efficiency of the vulcanization reaction, and reduce the sulfur requirements, in general enhancing the physical properties of the vulcanized rubber.

Accelerators are substances which act as catalysts in the vulcanization process, initiating free radicals. The first accelerators to be used in the rubber industry were inorganic chemicals, such as the basic carbonates and oxides of lead supplemented by magnesium or lime. Since the introduction of the first organic accelerator in 1906, several thousand have been patented. Only a small number of these are in general use. Most accelerators can be grouped into the following general classifications: aldehyde-amines, guanidines, thiuram sulfides, thiazoles, thiazolines, dithiocarbamates, and mercaptoimidazolines. *See* FREE RADICAL.

Some rubbers can be vulcanized by gamma radiation, but it is not an economically feasible process. For example, butyl rubber cannot be vulcanized by irradiation because degradation rather than cross-linkage occurs. The structure of the chain molecule is the determining factor. Cross-linking predominates if the polymer has the structure $(-CH_2CH_2-)_n$ or $(-CH_2CHR-)_n$. If the structure is $(-CH_2CR_1R_2-)_n$, degradation is the rule. There is great variation in the effects of gamma-radiation vulcanization, depending on the type of rubber being treated, the nature of the compounding ingredients, and the process conditions.

Although the vulcanization process was discovered by Charles Goodyear in 1839, an understanding of the mechanism of this solid-state reaction is still incomplete. There are formations of loose ends, rings, and other structures, but the main vulcanization reaction is a cross-linking of the long-chain polymer molecules by sulfur bridges or covalently bonded carbon-carbon linkages to form an extensive three-dimensional structure. This network structure modifies the essentially thermoplastic properties of the raw rubber and confers predominantly elastic properties on the resulting vulcanizate by preventing slippage of the long-chain molecules past each other.

Pigments. Vulcanization improves the elasticity and aging properties of rubber, but in most cases it is necessary to enhance further such properties as tensile strength, abrasion resistance, and tear resistance by incorporation of fillers. Those which improve specific physical properties are known as reinforcing fillers; those which serve primarily as diluents are classed as inert fillers. The physical properties of the resulting vulcanizates are affected by both the type and amount of filler.

The most universally used reinforcing filler in the rubber industry is carbon black. The types used commercially in the greatest bulk for this purpose include furnace and thermal blacks. Each may be further classified according to particle size and surface structure, and then selected according to specific properties which are required in the end product. *See* CARBON BLACK.

In addition to the carbon blacks, inorganic reinforcing agents, such as zinc oxide and the silicas, are used for reinforcement of light-colored end products. Although zinc oxide is not used extensively in the rubber industry, its incorporation does enable the resulting product to withstand extended exposure at high temperatures, and it also functions as an activator during the vulcanization process. The silicas are used in those products in which high abrasion resistance is an essential requirement.

The inert fillers, such as whiting and various types of clays, serve merely as extenders and to facilitate the processing of the compound. In general, they do not improve the tensile strength of the resulting vulcanizate, but do increase hardness and modulus.

Protective agents. The aging stability of rubber compounds is influenced by such factors as heat, light, and atmospheric conditions. The chemical unsaturation (presence of —C=C— groups) of natural and synthetic rubber compounds provides a focal point for oxidation. The initial oxidative degradation involves hydroperoxide formation and scission of polymer chains. These reactions are accompanied by network-forming or cross-linking reactions which occur to an extent, depending on the process conditions and the composition of the polymer. Chain scission is dominant in natural rubber, which becomes soft and sticky and loses tensile strength on degrading. In SBR (styrene-butadiene rubber), the predominant reaction is cross-linking, and so it becomes hard and brittle when attacked by oxygen. With either chain scission or cross-linking, the intermediates are free radicals which set off many repetitive reaction cycles. Some synthetic rubbers such as butyl have low unsaturation and so are resistant to oxygen and ozone.

While vulcanization improves resistance to attack by oxygen, further improvements have been accomplished by the incorporation of chemical compounds, known as antioxidants and antiozonants, into the rubber compound. Broadly speaking, antioxidants may be classified into two general groups: aromatic amines and phenols. The choice of the antioxidant is governed by the rubber and the purpose for which the rubber is intended. Antioxidants function by combining with one or more of the free radicals which arise and thus prevent further degradation of the rubber chains. Amine antioxidants have staining characteristics which do not interfere with their use in rubber compounds for tires, hose, and belts. However, light-colored end products or rubber compounds which come in contact with light-colored surfaces require nonstaining antioxidants. These compounds are usually highly hindered phenols, obtained by alkylation of phenols or cresols, or are derivatives of aromatic phosphite esters. Unfortunately, the most effective anoxidants usually tend to stain or discolor the vulcanizate. *See* ANTIOXIDANT.

The cracking of rubber compounds under stress when exposed to very low ozone concentrations, such as exist in the atmosphere, also originates from the unsaturated chemical structure of the rubber. The deteriorating effects of ozone can be diminished by the incorporation of various types of insoluble wax which bloom to the surface to form a physical barrier. If rubber is deformed, however, the protective surface is broken and is useless in many applications such as tires.

Certain accelerators have antiozonant properties, but the majority of commercial antiozonants are substituted *p*-phenylenediamines. These can be subdivided into the diaryl-, aryl-alkyl-, and dialkyl-*p*-phenylenediamines. Another type is formed by condensation of amines and ketones. The mechanism of antiozonant action has not definitely been established, but there is evidence to indicate that it is different from the reaction mechanism of antioxidants. The protective effect is believed to be the result of preferential combination of the ozone with the antiozonant instead of the rubber. Two parameters are believed to define ozone cracking behavior: the critical energy necessary to rupture the rubber surface, and the rate of crack growth. A large number of antiozonants decrease the rate of crack growth but only the N,N'-dialkyl-*p*-phenylenediamines increase the critical energy needed for the crack growth to occur.

Photooxidation of rubber compounds, resulting from exposure to bright sunlight, causes a chemical reaction which gives the surface a resinous or crazed appearance. This deleterious effect, caused mainly by ultraviolet light, is not reduced to any appreciable extent by antioxidants; indeed, some amine antioxidants are known to accelerate the degradation of natural rubber. Although complete protection against photooxidation is difficult, stocks containing

carbon black are sufficiently resistant for most applications. Ultraviolet-absorbing materials, such as salicylates and benzotriazoles, provide a certain degree of protection for light-colored products.

Gamma radiation causes degradation by initiating free-radical cross-linking and chain-scission processes in the rubber. The net change in physical properties depends on the extent of each process, which in turn is dependent on the molecular structure of the rubber, radiation dosage, and exposure conditions. Since the changes induced by radiation are the result of free-radical mechanism, the type of compounding can affect the results.

The degradative effects of radiation exposure may be reduced by incorporation of protective agents known as antirads, whose efficiency does not correlate with their behavior as antioxidants or antiozonants. Test results have shown that the following compounds impart radiation stability to various rubbers: N,N'-cyclohexylphenyl-p-phenylenediamine, a mixture of di-phenyl-p-phenylenediamine and phenyl-α-naphthylamine, quinhydrone, diphenylamine, and p-methoxyphenol. These represent just a few of the numerous compounds which can function as radiation inhibitors.

Physical testing. The physical testing of rubber compounds plays an essential role in rubber technology. Because available rubbers exhibit a wide range of properties which can be further varied by compounding techniques, standard methods of evaluation based upon physical properties have been established. Physical tests screen, measure, and evaluate the desired physical properties of a product, according to the information requested. In the rubber industry, tests are conducted on natural and synthetic rubbers, rubberlike materials, plastics, fabrics, and cords. What seem like trivial changes in the formulation, compounding, or processing of a material such as rubber are often quite discernible in the physical testing data. Frequent physical testing helps to ensure a standard product.

Commercial articles are often so complicated in structure that only actual service tests will suffice to prove their utility. In the most notable instance, new tire designs are introduced only after literally millions of miles of road testing.

Standards of the American Society for Testing and Materials (ASTM) are used to test materials whenever possible. This standardization of testing procedures simplifies the demands of the manufacturer and the compounder, and guarantees universal, comparable data.

In screening for physical properties, the first step is the evaluation of unvulcanized compounds. The raw polymers may be evaluated by techniques appropriate to any scientific investigation. These include chemical analyses and examination by infrared spectroscopy, nuclear magnetic resonance, pyrolysis, gel permeation chromatography, thermal analysis, and others. Polymer molecular weight may be estimated by a viscosity or relative plasticity measurement such as that done by the Mooney plastometer. Other instruments such as the Instron extrusion rheometer

measure plasticity at different shear rates.

In vulcanized compounds, stress-strain properties, that is, the relation of the force required to the sample elongation, are of importance. The elastic modulus, the ratio of stress to strain at low elongations, and the ultimate elongation to sample rupture are usually reported. Stress-strain properties are evaluated by the use of a constant-power-driven machine such as the Scott tensile tester. With the advancement in electronics, the Instron tester and similar apparatus employing load cells have come into practical use for obtaining more detailed information regarding such physical testing. *See* STRESS AND STRAIN.

All physical testing is conducted at a standard temperature of 25°C (77°F) and 50% relative humidity, except where higher or lower temperature testing is requested. Accelerated aging and conditioning of materials prior to testing is conducted in specially designed and automatically controlled ovens and refrigerators. Physical tests may be requested for properties such as hardness or softness, tear resistance, abrasion, flex, resistance to ozone, ultraviolet, salt spray and water, compression and recovery, permeability to gases, volatility, rebound, elasticity, dynamic modulus and resilience, and low-temperature properties. Each specific test employs a specifically designed apparatus for measuring the physical property or properties desired.

An example of a modern piece of equipment is the Gehman torsion apparatus. It has automatic controlling devices whereby measurements can be made at any desired temperature from ambient to −320°F (−195.5°C) by the use of liquid gases such as nitrogen. The Gehman torsion test determines the change in twist modulus with temperature and may be used to estimate the glass transition temperature of the rubber. This is the temperature at which there is enough free volume in the rubber that the polymer chain segments may slip or rotate past one another. This temperature is of fundamental importance since many rubbery properties depend on how far above it tests are conducted. Extensive research has attempted to uncover the complicated relation between many different physical properties and the chemical structure of a polymer. Rubber testing involves so many factors and variables that computer techniques have been developed to cope with their complexity. *See* INSPECTION AND TESTING.

Specific Types of Rubber

One rubber is obtained from natural sources in commercial quantities. In addition, a number of synthetic compounds are classified as rubbers.

Smoked sheet and pale crepe represent the forms in which the major portion of natural rubber is commercially available. The smoked sheet is obtained by acid coagulation of the tree latex, sheeting the coagulum, and then drying and smoking the sheets of rubber. The chemicals in the smoke preserve the rubber from molds and other organisms. Pale crepe is obtained by treating the latex, before or after coagulation, with sodium bisulfite and then washing, drying, and sheeting. Synthetic rubbers are available
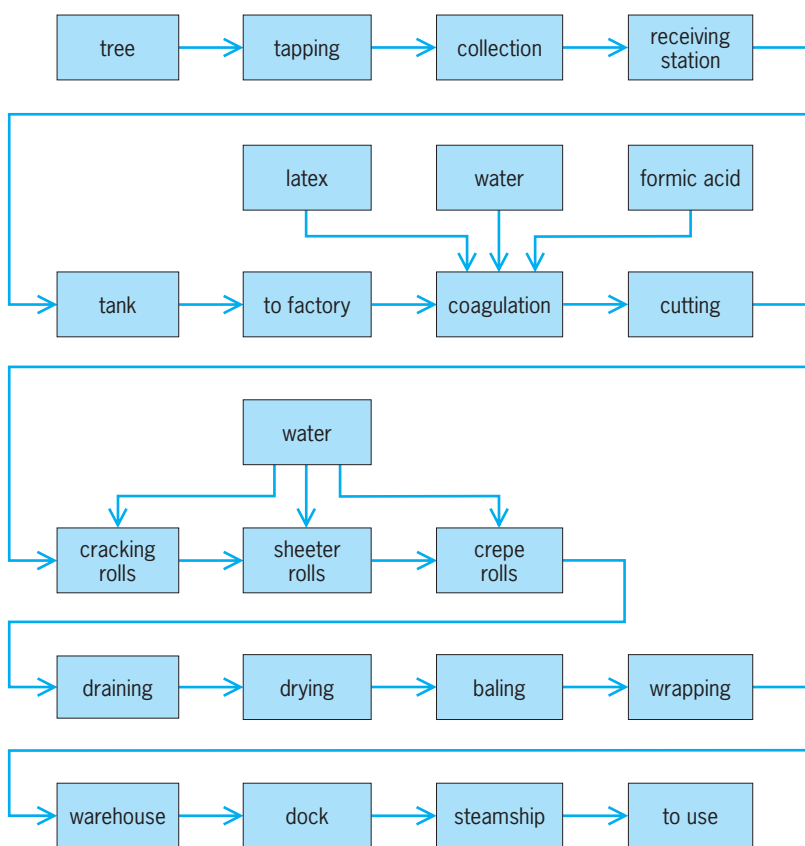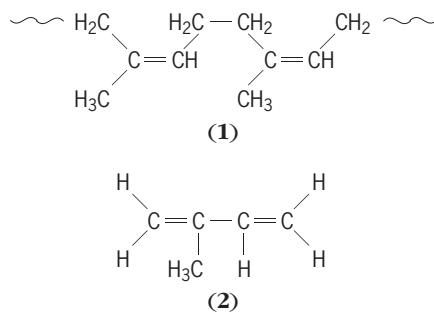
Fig. 1. Steps in natural-rubber production.

as bales, pellets, or flakes made from the polymerization process.

Natural rubber (NR). Although natural rubber may be obtained from hundreds of different plant species, the most important source is the rubber tree (*Hevea brasiliensis*). Natural rubber is *cis*-1,4-polyisoprene (**1**), containing approximately 5000 isoprene units (**2**) in the average polymer chain. World annual pro-

duction of natural rubber increased from 50,500 tons (45,900 metric tons) in 1900 to over 3.9×10^6 tons (3.5×10^6 metric tons) in 1980. *See* RUBBER TREE.

The economic competition from synthetic rubbers has stimulated research and development in natural rubber by increasing productivity in the field, improving uniformity and quality of the product and packaging, and developing modified natural rubbers with specific properties.

Increased productivity has been achieved by increasing the yield of the trees by cross-pollination of high-rubber-yielding clones of *H. brasiliensis*, use of

chemical stimulants, and better tapping and collection methods. These methods have resulted in large increases in yields on Malaysian estates, with much greater increases anticipated.

With improved productivity and better processing methods and controls, much better quality and uniformity have been reached. This has made possible the development of standard Malaysian rubbers (SMR) to meet specifications on a number of properties, including dirt and ash content, viscosity, and copper and manganese content.

A number of modified forms of natural rubber have been developed, including the following: superior processing (SP), oil extended (OENR), special tire rubber, deproteinized (DPNR), pellet form, filler master batches, and chemically modified.
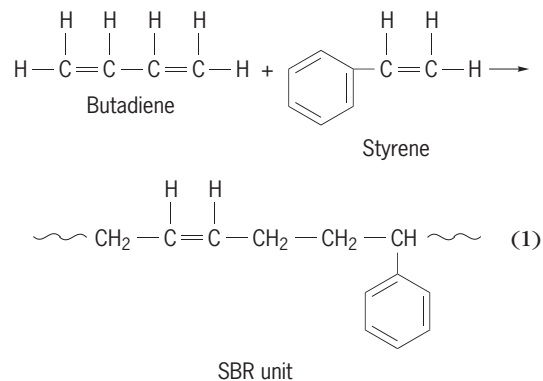
In addition, experimental work has been done on improved collection of latex by micro tapping and collecting in polyethylene bags, and trees having three parts: a high-yielding trunk system grafted onto a strong root system and an improved, more prolific leaf system grafted onto the trunk.

Despite competition from synthetic rubbers and a great reduction in percentage use worldwide of natural rubber, the tonnage of natural rubber continues a steady growth due to the growth of the industry. Steps in the production of natural rubber are shown in **Fig. 1**.

During World War II, considerable development work was done on cultivation of the Guayule shrub as a source of rubber, and some excellent rubber was produced experimentally from this plant in the United States but the process was not economically feasible. Improved cultivation and processing of Guayule coupled with the sharply increased cost of natural and synthetic rubber revived interest in this type of rubber. *See* GUAYULE.

Styrene-butadiene rubbers (SBR). The extensive development of the synthetic rubber industry originated with the World War II emergency, but continued expansion has been the result of the superiority of the various synthetic rubbers in certain properties and applications. The most important synthetic rubbers and the most widely used rubbers in the entire world are the styrene-butadiene types.

Formerly designated GR-S, SBRs are obtained by the emulsion polymerization of butadiene and styrene in varying ratios. However, in the most commonly used type, shown in reaction (1), the ratio

of butadiene to styrene is approximately 78:22. The polymerization is conducted in water emulsion also containing soap, a molecular weight regulator, and an initiator which generates the free radicals necessary to promote the polymer chain growth. At 122°F (50°C) conversion to polymer occurs at about 5–6%/h. The reaction is terminated at 70–75% conversion, since higher conversions lead to cross-linked polymer formation in the emulsion particle. The termination is effected by adding a so-called shortstop such as hydroquinone (about 0.1%), which reacts rapidly with free radicals and oxidizing agents. This destroys the initiator and prevents formation of any new polymer chains. Antioxidant is then added, the remaining butadiene and styrene are removed, and the latex is coagulated by adding salt and dilute sulfuric acid or aluminum sulfate. The coagulated crumb is then washed, dried, and baled for shipment. This general procedure is basic for production of all emulsion polymers. *See* POLYMERIZATION.

The SBR formed is a copolymer of styrene and butadiene units linked randomly in long chains. The butadiene units enter the polymer chain randomly in three different conformations known as *cis*-1,4; *trans*-1,4; and 1,2 or vinyl. The original or hot SBR is polymerized at a temperature of 122°F (50°C) or higher and is used primarily for mechanical goods because of its better color retention. The cold type, polymerized at 41°F (5°C), can be made to a higher molecular weight and therefore exhibits an improvement over the hot types in most physical properties. When used in light-duty tires such as passenger car tires, the cold-rubber compounds have proved equal or superior to natural rubber treads. However, they are inferior to natural rubber for truck tires because of the greater heat buildup during flexing. The cold, oil-extended type, prepared by the replacement of a portion of the polymer by a heavy-fraction oil, accounts for more than 50% of the cold-rubber production. Its advantage is primarily economic. Master batches, containing both oil and carbon black, have the advantage of process simplification.

In general, the compounding and processing methods for all the SBR types are similar to those of natural rubber. Although natural rubber is superior with respect to lower heat buildup, resilience, and hot tear strength, the SBR types are more resistant to abrasion and weathering. However, carbon black or some other reinforcing fillers must be added to the SBR to develop the best physical properties. Unlike natural rubber, SBR does not crystallize on stretching and thus has low tensile strength unless reinforced. The major use for SBR is in tires and tire products. Other uses include belting, hose, wire and cable coatings, flooring, shoe products, sponge, insulation, and molded goods.

cis-1,4-Polybutadiene (BR). Work on the duplication of natural rubber stimulated interest in stereoregulated polymerization of butadiene, particularly of the high *cis*-1,4 structure. Structure (**3**) shows the *cis*-1,4-polybutadiene units in the polymer chain. Several different catalyst systems were developed to give polybutadiene rubbers of varying *cis*-1,4 content. A

(**3**)

cobalt-containing catalyst, for example, will result in a polymer of 98% *cis*-1,4; a catalyst of an aluminum alkyl and titanium tetraiodide, 90–95% *cis*-1,4; and a catalyst of an alkyl lithium, 40–50% *cis*-1,4. All these systems employ air-free, moisture-free polymerization in hydrocarbon solvents.

Black-loaded vulcanizates of *cis*-1,4-polybutadiene rubbers exhibit good physical properties, such as lower heat generation, higher resilience, improved low-temperature properties, and greatly improved abrasion resistance. Processing properties are rather poor, but can be greatly improved by employing blends with natural rubber or SBR. A 1:1 blend of *cis*-1,4-polybutadiene and natural rubber displayed hysteresis properties equivalent to the natural rubber control and exhibited satisfactory modulus, tensile strength, and hardness. Tests on retreaded passenger tires gave outstanding abrasion resistance and increased resistance to cracking as compared to natural rubber. In passenger tires, *cis*-1,4-polybutadiene improves treadwear by about 1% for each percent of polybutadiene in the tread compound. In truck tires, a blend with natural rubber gives 14% more wear than natural rubber alone.

Butyl rubber. Isobutylene and isoprene or butadiene obtained from cracked refinery gases are the primary raw materials required for the manufacture of butyl rubber, as shown in reaction (2). Butyl is es-

Isobutylene   Isoprene

(2)

Butyl rubber unit

sentially a polyisobutylene except for the presence of 0.5–4.5% diolefin, usually isoprene, to provide the unsaturation necessary for vulcanization. *See* PETROCHEMICAL.

A solution polymerization system is used, the reaction being conducted at approximately −150°F (−101°C) in methyl chloride with aluminum chloride ($AlCl_3$) and a trace of water as the cationic catalyst. The low temperature is necessary to obtain a high-molecular-weight product and to control the almost explosive reaction rate. Butyl rubbers are processed according to conventional methods used for natural rubber and SBR. An antioxidant, sulfur, and an ultraaccelerator are compounded into the stock.

H H H H   H H
H—C=C—C=C—H + H—C=C—C≡N ⟶
Butadiene   Acrylonitrile

⁓ CH₂—CH=CH—CH₂—CH₂—C⁓ (4)
                                    |
                                    CN

Nitrile rubber unit

    H CH₃   H H
·· —C—C— ···   ··· —C—C— ··
    | | | |
    H CH   H C—CH₃
       ‖ |
       CH₂   CH₂
   (6)   (7)

of the rubber because the glass transition temperature increases with nitrile content. Both sulfur and nonsulfur vulcanizing agents may be used to cure these rubbers. Carbon black or other reinforcing agents are necessary to obtain the optimum properties. If proper processing methods are followed, the nitrile rubbers can be blended with natural rubber, polysulfide rubbers, and various resins to provide characteristics such as increased tensile strength, better solvent resistance, and improved weathering resistance.

The nitrile rubbers have outstanding oil, grease, and solvent resistance. Consequently, the commercial usage of these rubbers is largely for items in which these properties are essential. Another major usage is the utilization of the latex form for adhesives and for the finishing of leather, impregnation of paper, and the manufacture of nonwoven fabrics.

cis-1,4-Polyisoprene (IR). In 1954, synthetic *cis*-1,4-polyisoprene was made from isoprene with two different classes of catalysts. The first class includes lithium and the lithium alkyls. The second class uses a mixture of an aluminum alkyl and titanium tetrachloride, the system first used for the low-pressure polymerization of ethylene by Karl Ziegler. Both catalyst system polymerizations are carried out in hydrocarbon solution and require highly purified monomer and solvent. Traces of air, moisture, and most polar compounds adversely affect reaction rates, polymer properties, and structure.

The *cis*-1,4 polymer structure obtained with these catalysts is also characteristic of natural *Hevea* rubber. The presence of high *cis* content appears necessary for the desirable physical properties with *Hevea* in contrast to the inferior properties of emulsion-type polyisoprene, which contains mixed *cis*-, *trans*-, 1,2-, and 3,4-isoprene units [structures (4)–(7)] in the

H₃C   H
    \\   /
     C = C
    /   \\
··· —H₂C   CH₂— ···
         (4)

H₃C   CH₂— ···
    \\   /
     C = C
    /   \\
··· —H₂C   H
         (5)
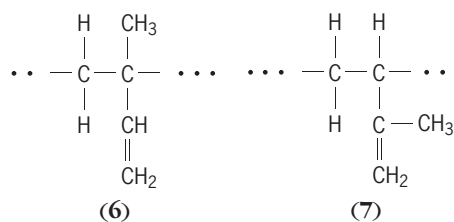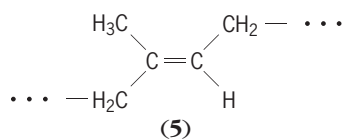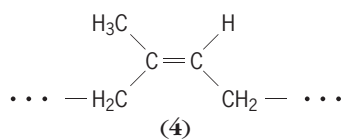
polymer chain. When polymerization is complete, the catalyst is destroyed, for example, by treating with alcohol, and the solvent removed. The solid polymer thus obtained may be handled and compounded as any other dry rubber is.

This polymer and the corresponding butadiene polymer discussed above are called stereorubbers because of their preparation with stereospecific catalysts. The emulsion polymers are formed by a free-radical mechanism that does not permit control of the molecular structure; stereorubbers are formed by anionic mechanisms that permit nearly complete control of the structure of the growing polymer chain in stereoregular fashion. *See* STEREOSPECIFIC CATALYST.
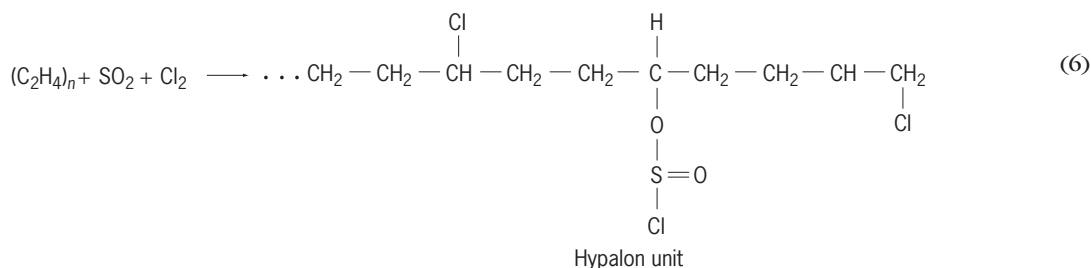
The type of catalyst employed influences the structure of the polymer. The aluminum-titanium-catalyzed rubbers exhibit, on stretching, a gradual crystallization somewhat slower than natural, and they are readily processed since the molecular weight range tends to be less than that of natural rubber. The lithium-catalyzed rubbers contain about 93% *cis*-1,4 structure; they exhibit very little tendency to crystallize and cannot be processed satisfactorily until they have undergone substantial mastication to reduce their very high molecular weight values to a level more nearly resembling that of well masticated natural rubber.

The differences in structure influence the properties of these rubbers. The aluminum-titanium-catalyzed rubber, since it crystallizes more readily, exhibits better hot tensile strength. The lithium-catalyzed polymer, being of higher molecular weight, exhibits higher resilience and less heat generation. Both types, when compounded and cured, produce physical properties which closely approach, or are equivalent to, natural rubber.

Tire tests with heavy-duty tires for trucks, buses, and airplanes have shown that with respect to wear and heat buildup the isoprene rubbers are comparable to natural rubber during tire operation. Polyisoprene rubber has passed qualification tests in high-speed jet aircraft tires to withstand landing speeds as high as 250 mi/h (112 m/s).

Other rubbers. There are other specialty rubbers available which are important because of specific properties, but in the aggregate they make up only approximately 2% of the world production of synthetic rubbers.

Silicone rubbers. Silicone rubber is a linear condensation polymer based on dimethyl siloxane. In the preparation, dimethyl dichlorosilane is hydrolyzed to form dimethyl silanol, which is then condensed to dimethyl siloxane, and this, upon further

CH₃ structures — reaction (5) and (6) diagrams:

$$\text{Cl}-\underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{Si}}-\text{Cl} \xrightarrow{H_2O} \text{HO}-\underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{Si}}-\text{OH} \xrightarrow{-H_2O} \text{HO}-\underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{Si}}-\text{O}-\underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{Si}}-\text{OH} \xrightarrow{-H_2O}$$

Dimethyl dichlorosilane → Dimethyl silanol → Dimethyl siloxane

$$\text{HO}\cdot\left[\cdots \underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{Si}}-\text{O}-\underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{Si}}-\text{O}-\underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{Si}}-\text{O}-\underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{Si}}-\text{O}\cdots\right]\cdot\text{H} \quad (5)$$

Dimethyl polysiloxane

$$(C_2H_4)_n + SO_2 + Cl_2 \longrightarrow \cdots CH_2-CH_2-\underset{\underset{Cl}{|}}{CH}-CH_2-CH_2-\underset{\underset{\underset{\underset{\underset{Cl}{|}}{S=O}}{|}}{\overset{\overset{H}{|}}{C}}}{}-CH_2-CH_2-\underset{\underset{Cl}{|}}{CH}-CH_2 \quad (6)$$

Hypalon unit
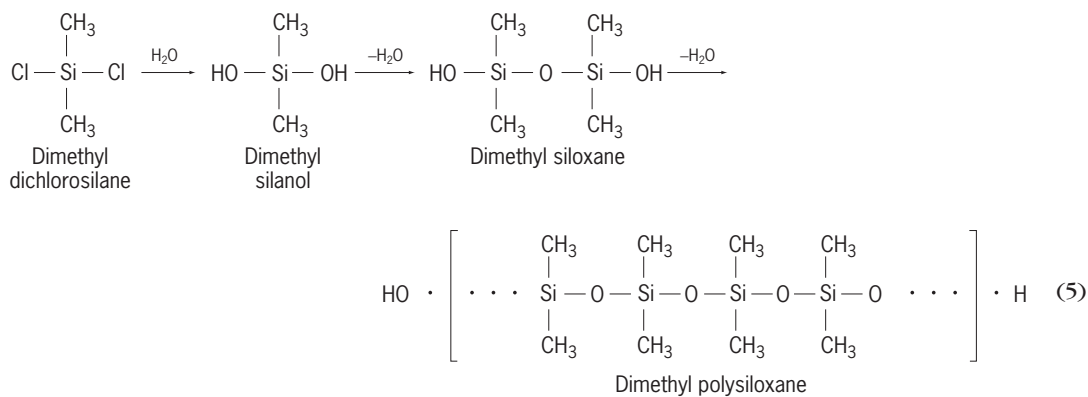
condensation, yields dimethyl polysiloxane, the standard silicone rubber [reaction (5)].

Various types of silicone rubbers are produced by substituting some of the methyl groups in the polymer with other groups such as phenyl or vinyl. Advantages of this type of substitution are evidenced by improvements in specific properties. For example, the presence of phenyl groups in the polymer chain gives further improvement in low-temperature properties. Fluorine-containing side groups improve chemical resistance. Many types are commercially available, ranging from fluid liquids to tough solids.

Because sulfur is not effective for the vulcanization of most silicone rubbers, a strong oxidizing agent, such as benzoyl peroxide, is used; the cross-linking produced is random. However, when unsaturated groups, such as vinyl or allyl, are present, these products are vulcanized with sulfur, the vinyl groups serving as a control of the degree of cross-linking.

Although the standard silicone rubbers are not reinforced by carbon black, the physical properties can be improved by the incorporation of various inorganic fillers such as titania, zinc oxide, iron oxide, and silica, which act as reinforcing and modifying agents. The physical, chemical, and electrical properties can be altered by varying the type and amount of these fillers. Carbon black can be used as a filler with vinyl-containing polymers.

In general, the silicone rubbers have relatively poor physical properties and are difficult to process. However, they are the most stable of rubbers and are capable of remaining flexible over a temperature range of −130 to 600°F (−90 to 316°C). They are unaffected by ozone, are resistant to hot oils, and have excellent electrical properties. Their most extensive uses are for wire and cable insulation, tubing, packings, and gaskets in aerospace and aircraft applications. In the form of dispersions and pastes, they are used for dip-coating, spraying, brushing, and spreading. Silicone rubbers are important in medical and surgical devices because of their property, unique among elastomers, of being compatible with body tissues. Fast, automatic, economical injection molding of liquid silicones has been developed. *See* SILICON; SILICONE RESINS.

Hypalon. Hypalon is the Du Pont trade name for a family of chlorosulfonated polyethylenes prepared by treating polyethylene with a mixture of chlorine and sulfur dioxide, whereby a few scattered chlorine and sulfonyl chloride groups are introduced into the polyethylene chain [reaction (6)]. By this treatment, polyethylene is converted to a rubberlike material in which the undesirable degree of crystallinity is destroyed but other desirable properties of polyethylene are retained. The outstanding chemical stability of Hypalon results from the complete absence of unsaturation in the polymer chain.

Vulcanization is accomplished by means of metallic oxides, such as litharge, magnesia, or red lead, in the presence of an accelerator. The sulfonyl chloride groups provide the sites of reactivity at which the bases react with the chlorine of the sulfonyl chloride group to give cross-links. The incorporation of an organic acid, such as hydrogenated wood rosin or stearic acid, is also necessary for optimum cure. Carbon black or other fillers are not needed to obtain optimum strength properties, and antioxidants are required only for maximum resistance to heat. In addition to its use as a stock in itself, Hypalon can be blended with other types of rubbers to provide a wide range of properties.

Hypalon has extreme resistance to ozone. Its chemical resistance to strong chemicals, such as nitric acid, hydrogen peroxide, and strong bleaching

CH_2—CH—CH_2Cl
 \\ /
  O

Epichlorohydrin

CH_2—CH_2
 \ /
  O

Ethylene oxide

$$\left(\begin{array}{c} H \quad H \\ | \quad\quad | \\ -C-C-O- \\ | \quad\quad | \\ H \quad H\text{-}C\text{-}H \\ | \\ Cl \end{array}\right)_n$$

Homopolymer
(n = about 5000)

$$\left(\begin{array}{c} H \quad H \quad H \quad H \\ | \quad | \quad | \quad | \\ -C-C-C-C-O \\ | \quad | \quad | \quad | \\ H \quad H\text{-}C\text{-}H \; H \quad H \\ | \\ Cl \end{array}\right)_n$$
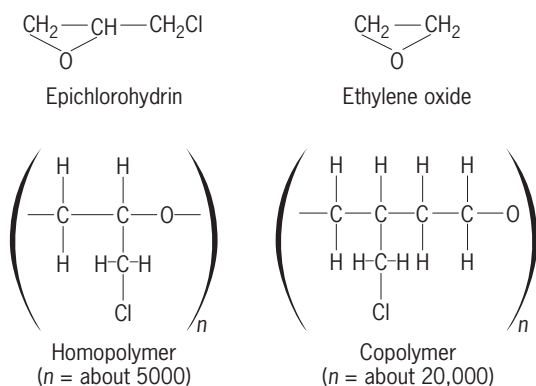
Copolymer
(n = about 20,000)

Fig. 2. Two main types of epichlorohydrin elastomers, showing the monomer and the repeating unit.

agents, is superior to any of the commonly used rubbers. These vulcanizates also have good heat resistance, mechanical properties, and unlimited colorability. Typical applications include white sidewall tires and a variety of sealing, waterproofing, insulating, and molded items.

Epichlorhydrin elastomers. These elastomers are polymers of epichlorhydrin. The two main types of epichlorhydrin elastomers are the homopolymer (CO) and the copolymer (ECO) of epichlorhydrin and ethylene oxide (**Fig. 2**). The copolymer has a lower brittle point, better resilience, and a lower specific gravity than the homopolymer, but is not as good in high-temperature properties.

Since the backbone of the molecule is saturated, these elastomers cannot be vulcanized with sulfur. Cross-linking is achieved by reaction of the chloromethyl side group with diamines or thioureas. A metal oxide is also required.

They exhibit outstanding ease of processing; extreme impermeability to gases; moderate tensile strength and elongation; high modulus; good abrasion resistance; good heat aging; excellent resistance to hot oil, water, perchloroethylene, acids, bases, and ozone; good low-temperature properties; and electrical properties ranging between those of a poor insulator and a good conductor, depending on compounding.

These properties make the epichlorhydrin elastomers useful in gaskets for oil field specialties, diaphragms, pump and valve parts; hose for low-temperature flexibility, oil and fuel resistance, or gas impermeability applications, and mechanical goods such as belting, wire, and cable.

Thermoplastic elastomers. Proper choice of catalyst and order of procedure in polymerization have led to development of thermoplastic elastomers. The leading commercial types are styrene block copolymers having a structure which, unlike the random distribution of monomer units in conventional polymers, consists of polystyrene segments or blocks connected by rubbery polymers such as polybutadiene, polyisoprene, or ethylene-butylene polymer. These types can be designated SBS, SIS, and SEBS, respectively. Other types of thermoplastic elastomers, such as the polyester type, have also been developed.

SBS, SIS, and SEBS polymers, when heated to a temperature above the softening point of the styrene blocks, can be processed and shaped like thermoplastics but, when cooled, act like vulcanized rubber, with the styrene blocks serving as cross-links. Thus, it is unnecessary to vulcanize. Also scrap or excess material can be reused.

Thermoplastic elastomers are very useful in providing a fast and economical method of producing a variety of products, including molded goods, toys, sporting goods, and footwear. One of the disadvantages for many applications is the low softening point of the thermoplastic elastomers. Attempts to modify the composition and structure to raise the softening temperature have generally resulted in inferior rubbery properties.

Fluoroelastomers. These are basically copolymers of vinylidene fluoride and hexafluoropropylene. Because of their fluorine content, they are the most chemically resistant of the elastomers and also have good properties under extremes of temperature conditions. They are useful in the aircraft, automotive, and industrial areas.

Polyurethane elastomers. Polyurethane elastomers are of interest because of their versatility and variety of properties and uses. They can be used as liquids or solids in a number of manufacturing methods. The largest use has been for making foam for upholstery and bedding, but difficulties have been encountered in public transportation due to flammability. The basic repeating structure (**8**) is shown for

(8)
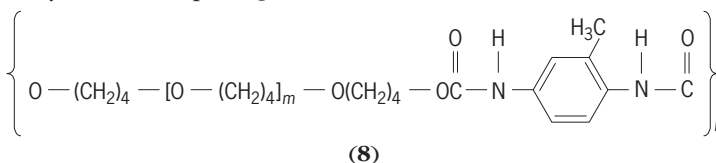
urethane rubber. *See* POLYURETHANE RESINS.

Polysulfide rubbers. These rubbers have a large amount of sulfur in the main polymer chain and are therefore very chemically resistant, particularly to oils and solvents. They are used in such applications as putties, caulks, and hose for paint spray, gasoline, and fuel. *See* POLYSULFIDE RESINS.

Polyacrylate rubbers. Polyacrylate rubbers are useful because of their resistance to oils at high temperatures, including sulfur-bearing extreme-pressure lubricants.   Edward G. Partridge

Products

A wide variety of products have a rubber, or elastomer, as an essential component. Most rubber products contain a significant amount of nonrubber materials, used to impart processing, performance, or cost advantages.

The automotive industry is the biggest consumer of rubber products. About 60% of all rubber used goes into the production of passenger, bus, truck and off-the-road tires. In addition, a typical automobile contains about 150 lb (68 kg) of rubber products such as belts, hose, and cushions. Outside the automotive area, a wide variety of rubber products are

produced, including such familiar items as rubber bands, gloves, and shoe soles.

The formulation and manufacture of rubber products is as diverse as the wide variety of applications for which they are intended. This discussion is limited to the more common methods of manufacture and those which apply to the widest variety of products. In general, the manufacture of rubber products involves compounding, mixing, processing, building or assembly, and vulcanization.

Compounding. The technical process of determining what the formulation components of a rubber product should be is called compounding. This term is also applied to the actual process of weighing out the individual components in preparing for mixing.

Rubber is the backbone of any rubber product. Natural rubber, obtained from the latex of rubber trees, accounts for about 23% of all rubber consumed in the United States. The balance is synthetic rubber. Of the synthetic rubbers, SBR (styrene-butadiene rubber) and polybutadine rubber are most important, accounting for 71% of synthetic rubber production. A variety of specialty synthetic rubbers, such as butyl, EPDM, polychloroprene, nitrile, and silicone, account for the balance of synthetic rubber production.

Choice of the rubber to be used depends on cost and performance requirements. The specialty rubbers often give superior performance properties but do so at higher product cost.

Many rubber products contain less than 50% by weight of rubber. The balance is a selection of fillers, extenders, and processing or protective chemicals. A typical formulation for the tread of a passenger car tire is shown in the **table**.

In the formulation shown in the table, the mixture of SBR and polybutadiene rubber represents a balance between cost, wear, grip, and aging factors. The carbon black adds strength to the rubber (reinforcing effect). A variety of specialty carbon blacks

Formulation for passenger car tire tread	
Component	Parts
Styrene-butadiene rubber	65
Polybutadiene rubber	35
Carbon black	67
Oil	40
Zinc oxide	3
Stearic acid	1
Accelerator	1.2
Sulfur	2
Wax	2
Antidegradant	2

are available to give a range of processing and performance characteristics. Oil is used as an extender for the elastomer and may be used as a softener or processing aid. In many cases, part of the oil may be incorporated by the synthetic rubber producer directly in the rubber and supplied as a rubber oil master batch.

Zinc oxide, stearic acid, and an accelerator are used to control the rate and degree of vulcanization. Mercaptobenzothiazoles and benzothiazole sulfenamide derivatives are the most commonly used accelerators for sulfur-cured products, although a wide variety of other chemical classes may be used. Sulfur is used to provide the cross-linking between rubber molecules. Wax and antidegradant are used to protect the finished product against thermal and oxidative aging, fatigue cracking, and ozone attack. Substituted quinolines, amines, or phenolics provide protection against thermal or oxidative aging.

In products other than tires, clays may be added as extenders, silicas as reinforcing agents, and plasticizers for flex or fire-retardant properties; and colors or brightening agents may also be used.

Mixing. This step accomplishes an intimate and homogeneous mix of the formulation components. Most large-volume stocks are mixed in internal mixers; these operate with two winged rotors with the compound ingredients forced into the rotors by an external ram.

In many cases, sulfur and accelerator systems are added in a separate mixing operation on a two-roll mill (**Fig. 3**); this mixing operation is normally carried out at lower temperatures than in an internal mixer to prevent the cross-linking reaction between sulfur and the rubber.

A rubber mill consists of two steel rollers. The rollers rotate at different surface speeds toward each other, thus giving a high shear condition that aids in mixing of the formulation.

There are thousands of different recipes, each designed for different purposes and each requiring a mixing procedure of its own. In addition, various manufacturers differ in their ideas as to precisely how the mixing operation should be conducted, and the equipment in various plants differs widely. The trend is toward more automation in this process, both in weighing and charging the ingredients into the mixer and in handling the stock after discharge.

Forming. Usually, forming operations involve either extrusion into the desired shape or calendering

Fig. 3. Two-roll mill. (*B. F. Goodrich Co.*)

to sheet the material to some specified gage or to apply a sheet of the material to a fabric (**Fig. 4**).

Building. Building operations, varying from simple to complex, are required for products such as tires, shoes, fuel cells, press rolls, conveyor belts, and life rafts. These products may be built by combining stocks of different compositions or by combining rubber stocks with other construction materials such as textile cords, woven fabric, or metal. For a few products no building operations are required. For example, many molded products are made by extruding the rubber, cutting it into lengths, and placing them in the mold cavities. Also, extruded products that are given their final shape by the die used in the extruder are ready for the final step of vulcanization, with no intermediate building operations.

In building tires, the "green" tire is built on a rotatable, collapsible drum, the diameter of which is slightly larger than the diameter of the bead rings. In producing bias-ply tires, a ply of coated cord fabric cut on a bias is applied to the drum so that the cords form an angle of about $40°$ with a circumferential line around the drum. A second ply is then applied, with the cords running at about $90°$ to those in the first ply. The bead assemblies are then applied, and the edges of the first two plies are folded over the beads. A third and then a fourth ply are added, with alternation of the direction of the cords, and their edges are folded over the beads. The extruded tread strip, which has previously been cut to the proper length and spliced, is then applied and rolled down. The drum is then collapsed and the green tire removed from the drum. **Figure 5** shows a cross section of a building drum with a green tire on it. Following the building, the tire may be shaped or "lifted" by inserting a rubber curing bag. This operation increases the diameter of the green tire in the crown or tread area and draws the two beads together, thus forming a section resembling that of the finished tire. Sometimes this operation is carried out automatically in the curing press.

The building operations for tires vary widely, depending on the kind of tire to be built and the equipment available in the particular plant. For example, the cord reinforcement may be of polyester, nylon, or steel wire; the number of plies required increases as the size and service requirements increase; the tread proper may be of a different composition than the tread base and sidewalls; one of the sidewalls may be made of a white stock; the construction may be of the tubeless variety, in which case an air-barrier ply of rubber is applied to the inside surface of the first ply; a puncture-sealing layer may be applied to the

Fig. 4.   A calender. (*B. F. Goodrich Co.*)

inside surface; or there may be several bead assemblies instead of only one.

Increasing popularity of radial tires has caused some basic changes in tire manufacturing operations. In a radial tire, the reinforcement cords in the body, or carcass, of the tire are placed in a radial direction from bead to bead. Radial tires require special building drums and result in a green tire of a different appearance from the barrel-shaped green bias-ply tire.

Vulcanization. The final process, vulcanization, follows the building operation or, if no building operations are involved, the forming operation.

Vulcanization is the process that converts the essentially plastic, raw rubber mixture to an elastic state. It is normally accomplished by applying heat for a specified time at the desired level. The most common methods for vulcanization are carried out: in molds held closed by hydraulic presses and heated by contact with steam-heated platens, which are a part of the press in open steam in an autoclave; under water maintained at a pressure higher than that of saturated steam at the desired temperature; in air chambers in which hot air is circulated over the product; or by various combinations of these methods.

The vast majority of products are sulfur-cured; that is, sulfur cross-links join the rubber chains together. For special applications, vulcanization may take place without the use of sulfur, for example,
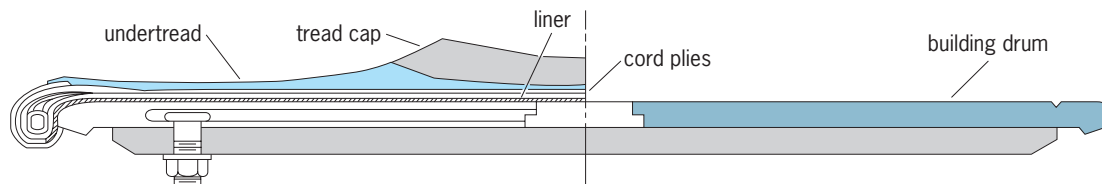

Fig. 5.   Cross section of building drum and green tire.

in resin- or peroxide-initiated cross-links or metal-oxide-cured polychloroprene.

The time and temperature required for vulcanication of a particular product may be varied over a wide range by proper selection of the vulcanizing system. The usual practice is to use as fast a system as can be tolerated by the processing steps through which the material will pass without "scorching," that is, without premature vulcanization caused by heat during these processing steps. Rapid vulcanization effects economies by producing the largest volume of goods possible from the available equipment. This is particularly the case for products made in molds, because molds are costly, and their output is determined by the number of heats which can be made per day.

The rate of vulcanization increases exponentially with an increase in temperature; hence, the tendency is to vulcanize at the highest temperature possible. In practice, this is limited by many factors, and the practical curing temperature range is 260–340°F (127–171°C). There are numerous exceptions both below and above this range, but it probably covers 95% of the products made.

Finishing operations following vulcanization include removal of mold flash, sometimes cutting or punching to size, cleaning, inspection for defects, addition of fittings such as valves or couplings, painting or varnishing, and packing.

Latex technology. In addition to the technology of solid rubber, there is a completely different, relatively small but important part of the rubber industry involving the manufacture of products directly from natural rubber latex from the tree, synthetic latex from emulsion polymerization, or aqueous dispersions made from solid rubbers.

In latex technology, materials to be added to the rubber are colloidally dispersed in water and mixed into the latex, a process involving the use of lighter equipment and less power than the mixing of solid rubber compounds. The latex compound can then be used in a variety of processes such as coating or impregnating of cords, fabrics, or paper; in paints or adhesives; molding (such as in toys); dipping (for thin articles like balloons, or household and surgeon's gloves); rubber thread (for garments); and production of foam. Latex technology is particularly important in producing articles for medical and surgical uses.   John J. Leucken; Edward G. Partridge

Bibliography. M. Ash, *Handbook of Plastic and Rubber Additives*, 2004; A. K. Bhowmick and H. L. Stephens (eds.), *Handbook of Elastomers (Plastics Engineering)*, 2000; D. C. Blackley, *Synthetic Rubbers*: *Their Chemistry and Technology*, 1983; J. Brydson, *Rubbery Materials and Their Compounds*, 1989; C. A. Harper, *Handbook of Plastics, Elastomers & Composites*, 2002; C. Hepburn (ed.), *Rubber Technology*, 3d ed., 1990; W. Hofmann, *Rubber Technology Handbook*, 1990; J. E. Mark, B. Erman, and F. R. Eirich (eds.), *Science and Technology of Rubber*, 2d ed., 1994; A. D. Roberts (ed.), *Natural Rubber Science and Technology Handbook*, 1988; M. R. Sethuraj and N. M. Mathew (eds.), *Natural Rubber: Biology, Cultivation, and Technology*, 1992; J. R. White and S. K. De (eds.), *Rubber Technologist's Handbook*, 2001.

Rubber tree

Hevea brasiliensis, a member of the spurge family (Euphorbiaceae) and a native of the Amazon valley. It is the natural source of commercial rubber. The tree may become 60–100 ft (18–30 m) tall and 8–10 ft (2.4–3 m) in circumference. *See* EUPHORBIALES.

Economic importance. It has been introduced into all the tropical countries supporting the rainforest type of vegetation, and is grown extensively in established plantations, especially in Malaysia. The latex from the trees is collected and coagulated. The coagulated latex is treated in different ways to produce the kind of rubber desired. Rubber is made from the latex of a number of
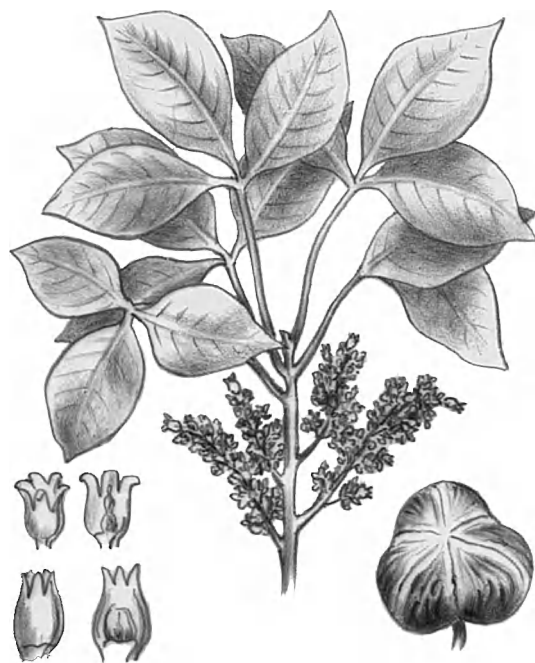
Foliage and flowers of the rubber tree.

other plants, but *Hevea* (see **illus.**) is the rubber plant of major importance. *See* RAINFOREST; RUBBER.
                     Perry D. Strausbaugh; Earl L. Core

Diseases. South American leaf blight, restricted to Central America and northern South America, remains a limiting major factor in the development of a rubber industry in the Western Hemisphere. Symptoms caused by the fungus *Microcyclus ulei* (three spore forms) are strictly foliar but vary according to leaf age. Young, expanding, reddish leaves (to 10 days of age) are most susceptible, but leaves 16 days old are effectively immune. Airborne conidia germinate immediately and initiate a disease cycle completed in about a week. Incidence and severity of the disease have been correlated with occurrence

and length of an annual dry season combined with seasonal and cumulative rainfall. Control of the disease in nurseries has been achieved with a fungicide (zineb) but is not economically feasible in the field. Field control has been accomplished through the use of resistant material crown-grafted onto susceptible stocks that give high yields of latex.

A complex of rubber diseases caused by *Phytophthora palmivora* includes vertical black lines in renewing bark (black thread or black stripe), discolored areas of untapped bark 0.4–0.8 in. (1–2 cm) above the soil line (patch canker), and defoliation of green pods and leaves of all ages (Phytophthora leaf fall, the only disease of mature rubber leaves). All aspects of this complex are most severe in wet weather. Incidence of black thread is reduced by opening panels in dry weather, extending the frequency of tapping (renewing bark becomes increasingly resistant to infection after 4–6 days), and applying fungicide after each tapping cut. Excision of diseased tissues and application of a fungicide has been practiced to control patch canker. Foliar application of cuprous oxide or copper oxychloride has been used to reduce leaf fall.

Secondary leaf fall caused by *Glomerella cingulata*, troublesome only under conditions of excessive moisture, can be controlled by repeated application of a fungicide (zineb, copper oxychloride, or chlorothalonil). Powdery mildew (*Oidium hevea*) affects young leaves in cool humid conditions, but can be controlled by early application of sulfur dust. Pink disease (*Botryobasidium salmonicolor*) is more prevalent in wet weather, and serious loss may result from bark damage. Brush application of 2% tridemorph in a natural latex base has provided control).

Bird's-eye spot, caused by *Drechslera hevea*, occurs in most rubber-growing areas but is more serious in nurseries. The disease is less severe in shade, and repeated applications of zineb or maneb have provided satisfactory control.

Symptoms of root diseases on rubber trees often are not apparent until affected trees have died. Infections by *Rigidoporus lignosus* (white root), *Ganoderma philippii* (red root), *Phellinus noxius* (brown root), or *Armillariella mellea* (*Armillaria* root disease) result from root contact with debris remaining after clearing (diseased) endemic flora. A rationale has been adopted in some areas that scattered, early, root disease losses of young rubber trees are acceptable if the disease complex is eliminated by the time that trees reach maturity. Control in recent years involves killing endemic trees in place by ring barking and the application of a herbicide, which also kills roots and renders them nonsusceptible to infection by root-disease fungi. Rubber is planted after clearing. Early collar and root infections by root-disease fungi may be controlled by excision of diseased tissues and the application of 20% pentachloronitrobenzene in bitumen (white root disease), 10% drazoxolon (red root disease), and 10% tridemorph (brown root disease). *See* PLANT PATHOLOGY.   A. A. Cook

Rubella

A benign, infectious virus disease of humans characterized by coldlike symptoms and transient, generalized rash. This disease, also known as German measles, is primarily a disease of childhood. However, maternal infection during early pregnancy may result in infection of the fetus, giving rise to serious abnormalities and malformations. The congenital infection persists in the infant, who harbors and sheds virus for many months after birth.

Virus characteristics. Rubella virus is a ribonucleic acid–containing, ether-sensitive virus about 50–70 nanometers in diameter. It has an internal nucleoid with a double membrane; virus grown in certain kinds of cell cultures will agglutinate red blood cells. Rubella virus can be propagated in tissue cultures, but is difficult to detect directly since it produces no cytopathic effect in cells in which it multiplies best—cercopithecus monkey kidney cells. Various tests for detection of the virus or its antibody are based upon the ability of the virus to interfere with the growth of certain other viruses which are cytopathogenic. The presence of unneutralized rubella virus prevents destruction of the cultures by the challenge virus. *See* ANIMAL VIRUS; ANTIBODY; TISSUE CULTURE.

Immunology. In rubella infection acquired by ordinary person-to-person contact, the virus is believed to enter the body through respiratory pathways. Antibodies against the virus develop as the rash fades, increase rapidly over a 2–3-week period, and then fall during the following months to levels that are maintained for life. One attack confers life-long immunity, since only one antigenic type of the virus exists. Immune mothers transfer antibodies to their offspring, who are then protected for approximately 4–6 months after birth. *See* IMMUNITY.

In the congenital infection, however, even though the fetus produces antibody in response to virus transmitted from the mother, the infant continues to harbor and shed the virus for as long as 18–24 months after birth, and is capable of transmitting the infection to susceptible contacts. Congenitally infected infants may have cataracts, deafness, mental retardation, microcephaly, congenital heart disease, growth retardation, and disorders of blood, liver, spleen, long bones, and skull, as well as interstitial pneumonitis. As a result of the severe epidemic of rubella in the United States in the spring of 1964, about 30,000 babies born in that fall and winter were afflicted with the congenital rubella syndrome.

Epidemiology. Rubella is a worldwide disease. Epidemics occur less frequently than those of some of the other childhood diseases, such as measles, and therefore more members of the population remain susceptible into young adulthood. Serological studies indicate that in most areas of the world, about 50% of the female population have already experienced infection by the age of 6–8 years and are therefore immune, and 80–85% are immune by 17–22 years of age.

Diagnosis. Since rubella infection in adults resembles a number of other illnesses, diagnosis cannot

be made with certainty on the basis of symptoms alone. Laboratory-proved rubella in the crucial period of pregnancy, the first 10 weeks, is associated with fetal infection in virtually 100% of cases, and there is a very high probability of serious abnormality or malformation of the fetus. Developments in the cultivation and testing of the rubella virus and its antibody make it feasible for pregnant women to be tested for preexisting rubella antibody and consequent immunity to infection, or for laboratory confirmation of suspected rubella infection.

Immunization. Live attenuated rubella vaccines have been available since 1969. The vaccine, which is now grown in human diploid cells, induces high antibody titers and an enduring and solid immunity. It may also induce secretory immunoglobulin (IgA) antibody in the respiratory tract and thus interfere with establishment of infection by wild virus. This vaccine is available as a single antigen or combined with measles and mumps vaccines (MMR vaccine).

The vaccine virus multiplies in the body and is shed in small amounts, but it does not spread to contacts. Vaccinated children pose no threat to mothers who are susceptible and pregnant. In contrast, nonimmunized children can contract and spread wild virus to susceptible family contacts. The vaccine induces immunity in at least 95% of recipients, and that immunity endures for at least 10 years.

The vaccine is safe and causes few side effects in children. In postpubertal females, the vaccine may produce self-limited arthralgia and arthritis in about one-third of vaccinees.

In the United States, control of rubella is being attempted by routine vaccination of children aged 1–12 years and selective immunization of adolescents and women of childbearing age. Before the vaccine became available, about 70,000 cases were being reported annually. Vaccination decreased the incidence of rubella to only 550 cases in 1986, a reduction of 99%. However, the decrease occurred primarily in children; in persons 15 years of age and older, only a small reduction in incidence was observed. Since the introduction of vaccine, scattered outbreaks of rubella still occur, chiefly among nonvaccinated adolescents in high school and college.

If vaccine is inoculated into pregnant women, the living vaccine virus may cross the placenta and infect the fetus. However, it is not teratogenic. In a study conducted between 1971 and 1986, more than 1000 women were vaccinated within 3 months of conception. None gave birth to children with congenital rubella syndrome. Therefore, accidental immunization during pregnancy is not an indication for termination of the pregnancy. Nevertheless, it is still considered prudent to avoid vaccination during pregnancy, and nonpregnant women vaccinees should be advised to delay conception for at least 3 months.

Because immunity may wane in those who were vaccinated as children, some women who become pregnant may be at risk of infection. Therefore it has been proposed that vaccine be given to prepubertal girls and to women in the immediate postpar-

tum period. It may be wise for all pregnant women to undergo a serum antibody test for rubella and, if found to be susceptible, receive a vaccination immediately after delivery. *See* BIOLOGICALS; VACCINATION.   Joseph L. Melnick

Bibliography. A. S. Evans, *Viral Infections of Humans*: *Epidemiology and Control*, 4th ed., 1997; D. M. Knipe et al. (eds.), *Fields Virology,* 4th ed., 2001; W. A. Orenstein et al., The opportunity and obligation to eliminate rubella from the United States, *J. Amer. Med. Ass.*, 251:1988–1994, 1984; M. K. Serdula et al., Serological response to rubella revaccination, *J. Amer. Med. Ass.*, 251:1974–1977, 1984.

Rubellite

The red to red-violet variety of the gem mineral tourmaline. Perhaps the most sought-for of the many colors in which tourmaline occurs, it was named for its resemblance to ruby. The color is thought to be caused by the presence of lithium. Fine gemquality material is found in Brazil, Madagascar, Maine, southern California, the Ural Mountains, and elsewhere. Gem material is almost exclusively a product of pegmatite dikes. Although rubellite is relatively inexpensive, it is regarded by many as one of the loveliest of gemstones. It has a hardness of 7–7.5 on Mohs scale, a specific gravity near 3.04, and refractive indices of 1.624 and 1.644. *See* GEM; TOURMALINE.   Richard T. Liddicoat, Jr.

Rubiales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Asteridae of the class Magnoliopsida (dicotyledons). The order consists of the large family Rubiaceae, with about 6500 species and the family Theligoniaceae with only 3 species. The Rubiales are marked by their inferior ovary; regular or nearly regular corolla with the petals grown together by their margins; stamens (equal in number to the petals) which are attached to the corolla tube alternate with the lobes; and opposite leaves with interpetiolar stipules or whorled leaves without stipules. The most familiar species of temperate regions are herbs with whorled leaves, such as madder (*Rubia tinctorium*, the traditional source of red dye), but opposite-leaved tropical shrubs such as *Coffea* (the source of coffee) are more typical of the group. *See* COFFEE; IPECAC; MAGNOLIOPHYTA; MAGNOLIOPSIDA; PLANT KINGDOM.   Arthur Cronquist; T. M. Barkley

Rubidium

A chemical element, Rb, atomic number 37, and atomic weight 85.47. Rubidium is an alkali metal. It is a light, low-melting, reactive metal. The physical properties of rubidium metal are summarized in the **table**. *See* PERIODIC TABLE.

	Temperature		Metric (scientific) units	British (engineering) units
Physical properties of rubidium metal				
Property	°C	°F		
Density	20	68	1.53 g/cm^3	95.5 lb/ft^3
Melting point	39	102		
Boiling point	688	1270		
Heat of fusion	39	102	6.1 cal/g	10.95 Btu/lb
Heat of vaporization	688	1270	212 cal/g	381 Btu/lb
Viscosity	50	122	6.26 millipoises	4.1 kinetic units
	220	428	3.23 millipoises	
Vapor pressure	294	561	1 mm	0.019 lb/in.^{2}
	628	1162	400 mm	7.75 lb/in.^{2}
Thermal conductivity	39	102	0.07 cal/(s)(cm^2)(cm)(°C)	16.9 Btu/(h)(ft^2)(°F)
	50	122	0.075 cal/(s)(cm^2)(cm)(°C)	18.1 Btu/(h)(ft^2)(°F)
Heat capacity	39–126	102–259	0.0913 cal/(g)(°C)	0.0913 Btu/(lb)(°F)
Electrical resistivity	50	122	23.15 microhm-cm	
	100	212	27.47 microhm-cm	

Most uses of rubidium metal and rubidium compounds are the same as those of cesium and its compounds. The metal is used in the manufacture of electron tubes, and the salts in glass and ceramic production.

Rubidium is a fairly abundant element in the Earth's crust, being present to the extent of 310 parts per million (ppm). This places it just below carbon and chlorine and just above fluorine and strontium in abundance. Sea water contains 0.2 ppm of rubidium, which (although low) is twice the concentration of lithium. Rubidium is like lithium and cesium in that it is tied up in complex minerals; it is not available in nature as simple halide salts as are sodium and potassium.

Rubidium has a density of 1.53 g/cm^3 (95.5 lb/ft^3), a melting point of 39°C (102°F), and a boiling point of 688°C (1270°F).

Rubidium is so reactive with oxygen that it will ignite spontaneously in pure oxygen. The metal tarnishes very rapidly in air to form an oxide coating, and it may ignite. The oxides formed are a mixture of Rb_2O, Rb_2O_2, and RbO_2. The molten metal is spontaneously flammable in air.

Rubidium reacts violently with water or ice at temperatures down to −100°C (−148°F). It reacts with hydrogen to form a hydride which is one of the least stable of the alkali hydrides. Rubidium does not react with nitrogen. With bromine or chlorine, rubidium reacts vigorously with flame formation. Organorubidium compounds can be prepared by techniques similar to those used for sodium and potassium. *See* ALKALI METALS; CESIUM.   Marshall Sittig

Bibliography. D. R. Lide, *CRC Handbook of Chemistry and Physics*, 85th ed., CRC Press, 2004.

Ruby

The red variety of the mineral corundum, in its finest quality the most valuable of gemstones. Only medium to dark tones of red to slightly violet-red or very slightly orange-red are called ruby; light reds, purples, and other colors are properly called sapphires. In its pure form the mineral corundum, with composition Al_2O_3, is colorless. The rich red of fine-quality ruby is the result of the presence of a minute amount of chromic oxide, usually well under 1%. The chromium presence permits rubies to be used for lasers producing red light. *See* LASER; SAPPHIRE.

The mineral corundum is commonly a constituent of basic igneous rocks, but it rarely occurs in a transparent form suitable for gem use. It is also known as a constituent of the type of marble formed in the contact zone of an igneous intrusion into an impure limestone. It is in this type of deposit that the finest rubies (those of Mogok, Burma) were formed. Today, most of those mined in the Mogok region are taken from the famous gem gravels of that area. There are only two other sources of significance, Ceylon and Thailand. In each of these countries, the finest quality obtained is far less valuable than fine Burma material, for the Ceylon ruby is too light in color and the so-called Siam ruby is a darker red. The finest ruby is the transparent type with a medium tone and a high intensity of slightly violet-red, which has been likened to the color of pigeon's blood. Star rubies do not command comparable prices, but they, too, are in great demand.

The ruby was among the first of the gemstones to be duplicated synthetically and the first to be used extensively in jewelry. A French chemist, A. Verneuil, announced successful reproduction in 1902 by a

flame-fusion process. (However, it is certain that flame-fusion synthetic rubies were available much earlier.) Many years later, C. Chatham produced synthetic ruby by flux fusion. Several others have also been successful. Bell Laboratories made synthetic rubies by a hydrothermal process. *See* CORUNDUM; GEM.   Richard T. Liddicoat, Jr.

Rugosa

One of the two principal orders of Palaeozoic corals, the other being the Tabulata. Rugose corals first appeared in the Middle Ordovician. The small number of known Cambrian corals include at least one species with superficial similarity, but it is not considered a direct ancestor of the Rugosa. The Rugosans diversified steadily during the Early and Middle Paleozoic to reach their greatest diversity in the Middle Devonian. A series of extinction events in the Late Devonian almost resulted in the disappearance of the group, but a handful of survivors rediversified to almost match peak diversity by the Late Mississippian. Thereafter, the group declined, slowly at first but more rapidly in the Permian, and became extinct at the end of the Permian. *See* PALEOZOIC; PERMIAN.

The rugose coral skeleton (or corallum) was secreted by a polyp (coral animal) assumed to have been similar in many respects to the polyps of living corals. Rugose corals existed in either solitary or colonial form, though solitary forms were more common.

Solitary forms. The skeleton of a solitary rugose coral (corallite) consists of a cup, cylinder, or (more rarely) a plate that is almost always defined by a solid outer wall (epitheca), in or on which radial vertical blades or spines (septa) are disposed (**Figs. 1***a*, *b*, **2***a*, *b*). The upper surface, to which the soft parts were attached, is the calice. Subhorizontal elements are usually deposited between or beneath the septa and vary, including simple, more or less flat plates only (tabulae), a peripheral series of smaller vesicular plates (dissepiments), and an axial series of larger

tabulae which may vary in form considerably from genus to genus (Figs. 1*c* and 3*b*). An axial structure, constructed of modified septal ends and/or axial tabulae, may be present (Fig. 2*d*).

Colonial forms. In about one-third of rugose corals, the initial corallite (protocorallite) produces offsets to form a colony. There was no specialization of individuals in rugose coral colonies and all the corallites are essentially similar in appearance. The offsets formed either fasiculate colonies, characterized by discrete cylindrical corallites, either parallel or divergent, that were held in a cluster by the skeleton beneath them (Fig. 2*c*) or massive colonies, in which the corallites remained in contact. In many massive colonies, the corallites retained their individual epithecae, and packed together to form a polygonal (often hexagonal) cross section (Fig. 2*d*). However, in a small number of these colonies, the epithecae are lost, and are replaced by a continuous holotheca (external wall), which surrounds the entire colony. The corallites then merge into one another in various ways (Figs. 2*e* and 3*b*). Often, the peripheral ends of septa are confluent between corallites, or (more rarely) withdrawn from the intercorallite areas so that corallite centers are joined by dissepimental tissue only. This is thought to reflect integration of the polyps secreting the skeleton with continuous neural networks across the colony surface and, possibly in some rare cases, confluent gastric cavities.

Classification. Classification of rugose corals can be difficult, as the skeletons are often relatively simple and the same structures may appear again in different lineages. The presence and arrangement of septa, dissepiments, and tabulae; the form of any axial structure that may be present; and to some extent the microstructure of the skeletal elements are all important in classification. The number of major subdivisions is relatively stable, but the phylogenetic relationships between them are disputed.

Rugosa versus Scleractinia. The superficially similar scleractinian corals of the post-Paleozoic are not directly related to the Rugosa, although the
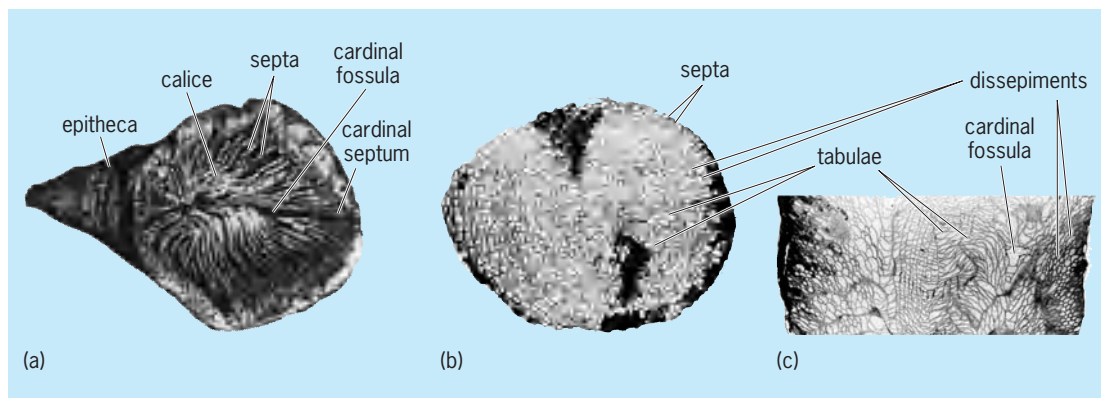

Fig. 1. Rugose corals. (*a***)** *Aulacophyllum* **sp of horn coral showing the major external features of a solitary coral skeleton; septa pinnately arranged about the cardinal fossula. Devonian, North America, ×0.75. [Most rugose corals need to be studied in thin section for accurate identification, as shown in parts *b* and *c*.] (***b***) Cross section of** *Phaulactis angusta***, showing septa and the intersects of dissepiments and tabulae between the septa, ×1. (***c***) Longitudinal section of** *P. angusta***, showing dissepiments, tabulae, and the depression that is the cardinal fossula. Silurian, U.K., ×1.**
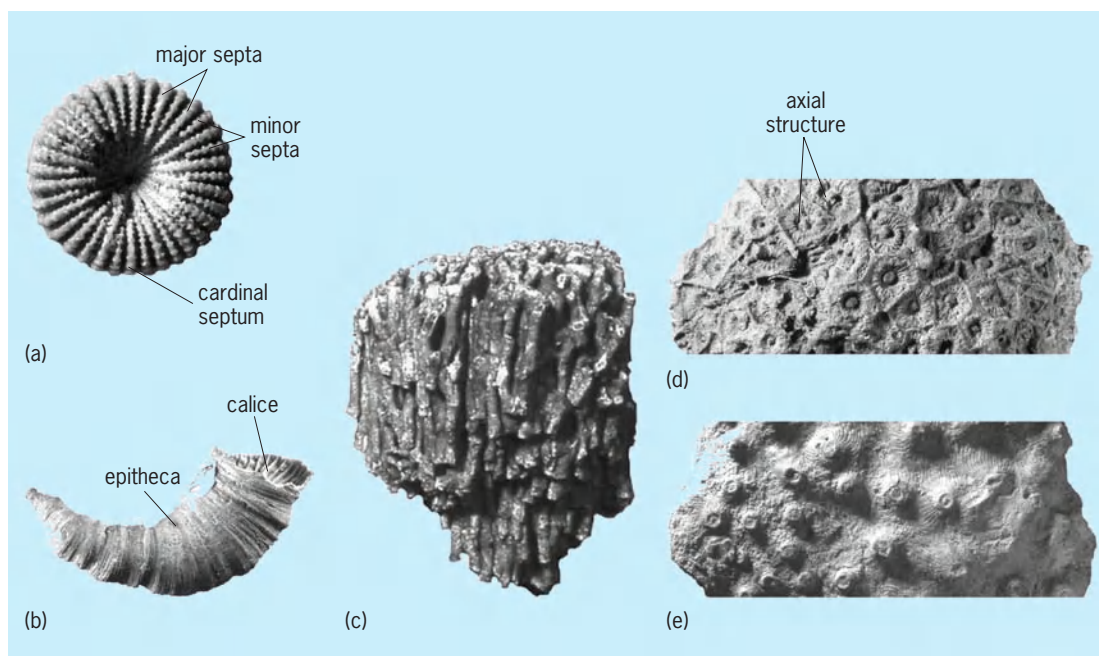
Fig. 2. Rugose coral skeletons. (*a*) *Palaeocyclus porpita*, discoidal corallite, calical view with platelike epitheca beneath, cardinal septum shortened. Silurian, Sweden, ×2. (*b*) *Caninia* sp., typical horn coral in growth orientation with septa visible in the calice. Carboniferous, Ireland, ×1. (*c*) *Siphonodendron* sp., fasciculate colony. Carboniferous, UK, ×0.5. (*d*) *Actinocyathus floriformis*, massive colony, individual corallite walls retained. Carboniferous, UK, ×0.6. (*e*) *Arachnophyllum murchisoni*, massive colony, individual corallite walls lost; peripheral ends of septa either abut or are continuous between corallites. Silurian, U.K., ×0.4.

relationship of the polyp to the skeleton in living scleractinian corals can be used to some extent to interpret the biology of the rugose corals.

Morphology. The main differences between the Rugosa and Scleractinia are (1) a contrasting mode of formation of the septa, together with a second order of minor septa (and rarely additional orders) in Rugosa; (2) a different mineralogy of the skeleton, calcitic in Rugosa, aragonitic in Scleractinia; and (3) the relative rarity of an epitheca surrounding the corallite in Scleractinia. In the Scleractinia, septal insertion is related to the sequence of appearance of fleshy radial partitions (mesenteries) in the polyp; a similar relationship is assumed for the Rugosa. Both orders have six protosepta (the initial septa); however, septal insertion in scleractinians is cyclical in all six sectors defined by the protosepta, whereas in the rugosans, septa are inserted serially in only four of the sectors. This can result in a pinnate arrangement of septa, particularly around the protoseptum (known as the cardinal septum), which may be shortened (Figs. 1*a* and 2*a*) and associated with a pit (known as the cardinal fossula) in rugose corals. However, in most mature corallites, the arrangement of septa is essentially radial.

Paleoecology and ecology. Another major difference between the Rugosa and the Scleractinia is in their ecology and paleoecology. The vast majority of reef-building scleractinian coral polyps are symbiotic with algae (zooxanthellae), which play an important role in their metabolism and reef-building potential. Indirect evidence suggests that this relationship was not developed in the rugosans. The Rugosa were not able to dominate reef building in the Paleozoic as the scleractinians do in the post-Paleozoic. This appears to be the result of the almost universal presence of an epitheca or holotheca in the Rugosa, which seems to have inhibited secure attachment to a hard substrate. Rugosans mostly have very small attachment scars, and almost all were essentially free-living as adults.

Paleoecological adaptations. Paleoecological adaptations predominantly relate to ensuring stability on a soft substrate.

Solitary rugosans. Among solitary rugosans, long cylindrical corallites had much of their length buried in soft sediment to hold them upright. Storm events resulted in uprooting, survivors often showing sharp changes in their direction of growth. Most cylindrical corals were ultimately preserved lying on their sides (Fig. 3*a*). The common horn-shaped coralla lived convex-side down partially buried in soft sediment (Fig. 2*b*). Some of these developed noncircular cross sections to aid stability. Discoidal corallites of low relief could spread their weight across the substrate (snow-shoe strategy).

Colonial rugosans. Fasciculate colonial rugosans relied on sediment building up between their corallites to ensure stability. Massive rugosans may show the effects of partial mortality due to the movement of soft sediment, with breaks in growth followed by recolonization over a layer of sediment from surviving areas of the colony (Fig. 3*b*). Massive corals followed two principal strategies: (1) spreading, low relief, disk- to domelike colonies adapted to areas of episodic sedimentation and (2) less-spreading, high relief, conical colonies adapted to areas with more
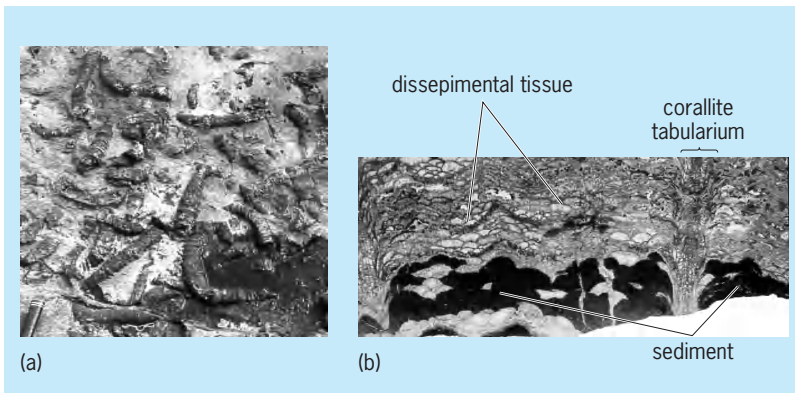
Fig. 3. Rugose coral skeletons. (*a*) *Siphonphyllia gigantea*, solitary cylindrical corallites prone on bedding plane; individual bottom-right showing sharp change of growth direction. Carboniferous, Ireland; width of photo, 1 m. (*b*) *Arachnophyllum pentagonum*, massive colony with extensive sediment fouling in lower part, as indicated by the darker sediment infill. Corallite tabularia connected by continuous dissepimental tissue. Silurian, North America, ×1.

continuous sedimentation. In both cases, sediment fouling was the most likely cause of death. Colonial corals, particularly fasciculate forms, are much more likely to be found in position of growth than solitary corals.

Distribution. Thus, rugose corals played only a minor part in Paleozoic reef-building, although they were more common in peri-reefal environments. Middle Paleozoic reef core facies were dominated by stromatoporoids. Massive tabulate corals often contributed to the structure and rugose corals might be found scattered through the fabric. In some cases, fasciculate rugosans were overgrown by stromatoporoids in an apparently symbiotic relationship. Massive rugosans occasionally occupied niches in the reef core but were more common on the reef flat, and specialized solitary and fasciculate rugosans were often present in more restricted back-reef environments. Rugose corals were most abundant in bioherms (lenslike to moundlike organic structures) and biostromes (sheetlike accumulations of skeletal organisms) usually associated with fine-grained carbonate sediments or calcareous muds. Some biostromes are close to monospecific, while others, and particularly many bioherms, may contain a great diversity of species and growth forms. Rugosans remain reasonably common in deposits interpreted as formed in deeper shelf environments, but little is known of their penetration to off-shelf depths. Their diversity was highest in warm, tropical and subtropical waters, decreasing markedly toward higher temperate and polar regions. Faunas in high latitudes and in deeper waters tend to be dominated by solitary corals, and any colonial corals are relatively small. There was also a specialized fauna of small, thick-walled, solitary rugose and tabulate corals (the laccophyllid fauna) that colonized marginal environments. *See* CNIDARIA; SCLERACTINIA; TABULATA.

Colin Scrutton

Bibliography. D. Hill, 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 corals groups and the origin of the Scleractinia, ed. by G. D. Stanley, Jr., *Paleobiology and Biology of Corals: The Paleontological Society Papers*, vol. 1, pp. 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, 52:1–57, 1998; J. E. Sorauf, Biocrystallisation models and skeletal structure of Phanerozoic corals, in G. D. Stanley, Jr. (ed.), *Paleobiology and Biology of Corals: The Paleontological Society Papers*, vol. 1, pp. 159–185, 1996; G. E. Webb, Morphological variation and homoplasy: The challenge of Paleozoic coral systematics, in G. D. Stanley, Jr. (ed.), *Paleobiology and Biology of Corals: The Paleontological Society Papers*, vol. 1, pp. 135–157, 1996.

Runge vector

The Runge vector describes certain unchanging features of a nonrelativistic two-body interaction for which the potential energy is inversely proportional to the distance r between the bodies or, alternatively, in which each body exerts a force on the other that is directed along the line between them and proportional to r^{-2}. Two basic interactions in nature are of this type: the gravitational interaction between two masses (called the classical Kepler problem), and the Coulomb interaction between like or unlike charges (as in the hydrogen atom). Both at the classical level and the quantum-mechanical level, the existence of a Runge vector is a reflection of the symmetry inherent in the interaction. *See* COULOMB'S LAW; KEPLER'S EQUATION; NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; SYMMETRY LAWS (PHYSICS).

Classical Kepler problem. The nonrelativistic two-body problem can be transformed into an equivalent problem for one body attracted to a fixed center. In the equivalent one-body problem, the motion is governed by the relation in Eq. (1), where $\hat{\mathbf{r}} = \mathbf{r}/r$,

$$\frac{d\mathbf{p}}{dt} = -\lambda r^{-2}\hat{\mathbf{r}} \qquad (1)$$

$\mathbf{p} = m\, d\mathbf{r}/dt$, m is the reduced mass, and λ is a constant (positive for attractive force) that characterizes the strength of the interaction. It follows that the energy E and the angular momentum $\mathbf{L}$ are constants of the motion and that Eqs. (2) hold. The Runge vector, defined by Eq. (3), is also conserved (that is, does

$$E = \frac{p^2}{2m} - \frac{\lambda}{r} \qquad (2)$$
$$\mathbf{L} = r \times p$$

$$\mathbf{R} = \mathbf{p} \times \mathbf{L} - \lambda m\hat{\mathbf{r}} \qquad (3)$$

not change with time). Since $\mathbf{R} \cdot \mathbf{L} = 0$, it follows
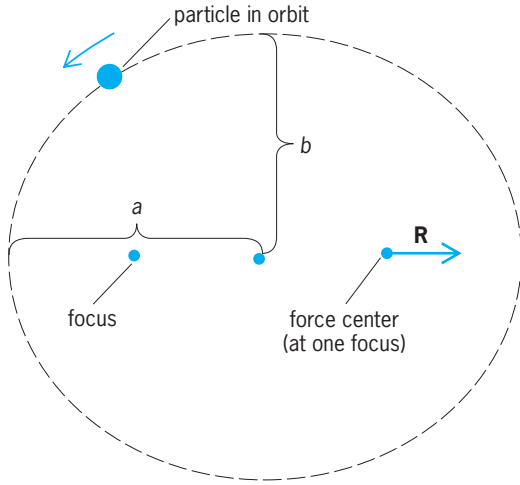
Runge vector for elliptical motion ($R = \lambda m \cdot \sqrt{a^2 - b^2}/a$).

that **R** is a fixed vector in the plane of the orbit. If **R** is taken to be the fixed direction from which the azimuthal angle θ is measured, Eq. (4) holds. Here

$$\frac{1}{r} = \frac{\lambda m}{L^2}\left(1 + \frac{R}{\lambda m}\cos\theta\right) \qquad (4)$$

$R = |\mathbf{R}| = [2mEL^2 + \lambda^2 m^2]^{1/2}$. Hence **R** points in the apsidal direction and its magnitude determines the eccentricity $\epsilon = R/\lambda m$. The **illustration** shows the situation for an elliptical orbit. Since the particle in orbit endlessly retraces the same path (broken line in the illustration), the shape of the elliptical path (described by the eccentricity) and its orientation (described by the direction **R**/R) are constant. A single-valued constant quantity giving a complete description of the orbit is possible only because of the periodic nature of the motion; that is, the orbit is reentrant. Moreover, a single-valued conserved vector in the plane of motion occurs only in the non-relativistic problem in which the force varies as r^{-2}. In the corresponding relativistic problem, the ellipse precesses.

This symmetry (constancy of the orbit path in the bound case), which leads to the Runge vector, is reflected in other aspects of the Kepler problem. The fact that the corresponding Hamilton-Jacobi equation is separable not only in spherical coordinates but also in parabolic coordinates implies the existence and form of the Runge vector. The components of the Runge vector may also be related to the generating functions of canonical transformations that transform a given orbit pertaining to a fixed energy into a different orbit of the same energy. The group of transformations comprising those generated by rotation in three-dimensional space (namely, generated by **L**) as well as by the Runge vector is isomorphic to rotations in four-dimensional space. *See* CANONICAL TRANSFORMATIONS.

Quantum-mechanical problem. The hamiltonian of the Schrödinger equation analogous to the classical problem is given by Eq. (5), where **p** is the oper-

$$H = \frac{\mathbf{p} \cdot \mathbf{p}}{2m} - \frac{\lambda}{r} \qquad (5)$$

ator $-i\hbar\nabla$. In addition to the angular momentum operator $\mathbf{L} = \mathbf{r} \times \mathbf{p}$, there exists a conserved hermitian Runge vector operator given by Eq. (6). Its

$$\mathbf{R} = {}^1/_2(\mathbf{p} \times \mathbf{L} - \mathbf{L} \times \mathbf{P}) - \lambda m\hat{\mathbf{r}} \qquad (6)$$

time independence is assured by the fact that it commutes with H. Its existence and form can be inferred from the fact that the time-independent Schrödinger equation is separable in both spherical and parabolic coordinates.

Instead of the Runge vector given above, the Runge-Lenz vector operator $\mathbf{A} = (-2mH)^{-1/2}\mathbf{R}$ is usually used. It is well defined on all eigenstates and is hermitian for bound states ($E' < 0$). The square of the Runge-Lenz operator is given by Eq. (7), where

$$\mathbf{A} \cdot \mathbf{A} = -\lambda^2 m(2H)^{-1} - (\mathbf{L} \cdot \mathbf{L} + \hbar^2) \qquad (7)$$

$\hbar$ is Planck's constant divided by 2π. *See* EIGENFUNCTION; EIGENVALUE (QUANTUM MECHANICS).

The Runge-Lenz vector operator **A** and the angular momentum operator **L** have the commutation properties characteristic of the Lie algebra 0(4) [orthogonal group of rotations in four dimensions]. Explicitly, Eqs. (8)–(10) hold, where ϵ_{ijk} is antisymmetric in its

$$A_iL_j - L_jA_i = i\hbar\epsilon_{ijk}A_k \qquad (8)$$

$$A_iA_j - A_jA_i = i\hbar\epsilon_{ijk}L_k \qquad (9)$$

$$L_iL_j - L_jL_i = i\hbar\epsilon_{ijk}L_k \qquad (10)$$

indices with $\epsilon_{123} = 1$. Moreover, since both A_3 and L_3 commute with each other as well as with the hamiltonian, a simultaneous eigenfunction of these operators can be found. This corresponds to separating the Schrödinger equation in parabolic coordinates $\xi = r(1 + \cos\theta)$, $\eta = r(1 - \cos\theta)$, and ϕ, instead of the usual spherical coordinates r, the polar angle θ, and the azimuthal angle ϕ. These simultaneous eigenfunctions are not also eigenfunctions of the angular momentum since L_2 does not commute with A_3. For energy $E = k^2(2m)^{-1}$ which is greater than zero, one particular eigenfunction is of special interest. It is the one whose L_3 and A_3 eigenvalues are respectively zero and $(\hbar - im\lambda k^{-1})$. It is defined by Eq. (11), where F_{11} is the confluent hypergeomet-

$$\psi = e^{ikr\cos\theta}{}_1F_1[im\lambda(\hbar k)^{-1}, 1, ik\eta] \qquad (11)$$

ric function. As $r \to \infty$, this wave function behaves as if it were a plane wave plus an outgoing spherical wave, and hence this single eigenfunction has the asymptotic properties of a scattering solution for the Schrödinger equation. The scattering solution is useful in a discussion of scattering of a beam of noninteracting charged nonrelativistic particles off

a Coulomb force center since, asymptotically, the plane wave represents that which is scattered by the Coulomb force center. The same eigenfunction is expressible as an infinite sum of angular momentum eigenfunctions, that is, the so-called partial wave expansion of the scattering eigenfunction. *See* DISPERSION RELATIONS.

Symmetry and invariance group. Since the potential for the hydrogen atom is spherically symmetric, it is apparent that both the classical and quantum-mechanical problems enjoy three-dimensional rotational symmetry. The additional symmetry peculiar to the $1/r$ potential extends that symmetry to a four-dimensional symmetry. The space of the four dimensions is that of the variables $2(-2mE)^{1/2}\mathbf{p}/(p^2 - 2mE)$ and $(2mE + p^2)/(p^2 - 2mE)$, where $\mathbf{p}$ is the ordinary three-dimensional momentum. It can be shown that the Schrödinger equation in momentum space can be converted into an integral equation symmetric in these four variables by projecting the momentum space stereographically onto a four-dimensional unit sphere. The abstract Kepler problem in n spatial dimensions has been analyzed with similar results.

The additional internal symmetry is responsible for the degeneracy (that is, number of distinct states of a single energy) characteristic of the hydrogen atom. The group 0(4) is an invariance group of the hydrogen atom since all the bound states of a given energy constitute the basis for an irreducible representation of that group. In turn, all the bound states constitute an irreducible representation of the noncompact DeSitter group 0(4,1). This larger group (whose generators do not all commute with the hamiltonian) is called the noninvariance group of the problem. The conceptual analogy between energy levels and mass, hydrogen atom states and elementary particle states, has given hope that perhaps the spectrum of elementary particles can also be understood in terms of an appropriate invariance and noninvariance group. *See* ELEMENTARY PARTICLE; ENERGY LEVEL (QUANTUM MECHANICS).   D. M. Fradkin

Bibliography. I. M. Bander and C. Itzykson, Group theory and the hydrogen atom, *Rev. Mod. Phys.*, 38:330–346, 1966; G. F. Bassani, M. Inguscio, and T. W. Hansch (eds.), *The Hydrogen Atom*, 1989; H. Goldstein, *Classical Mechanics*, 3d ed., 2001; L. D. Landau and E. M. Lifshitz, *Mechanics*, 3d ed., 1976; E. C. G. Sudarshan, N. Mukunda, and L. O'Raifeartaigh, Group theory of the Kepler problem, *Phys. Lett.*, 19:322, 1965.

Running fit

The intentional difference in dimensions of mating mechanical parts that permits them to move relative to each other. A free running fit has liberal allowance; it is used on high-speed rotating journals or shafts. A medium fit has less allowance; it is used on low-speed rotating shafts and for sliding parts. Running fits are affected markedly by their surface finish and the effectiveness of lubrication. *See* ALLOWANCE.

Running and sliding fits are standardized into nine classes. Close sliding fits accurately locate parts with some sacrifice in free motion; they permit no perceptible play. Sliding fits permit the parts to move but are not intended for freely running parts or moving parts subject to appreciable temperature change. Precision running fits permit parts to run freely at low speeds and at light journal pressures, provided temperature differences are limited. Close, medium, and free running fits are intended for progressively higher surface speeds, journal pressures, and temperature ranges. Loose running fits are for use with cold-rolled shafting and tubing made to commercial tolerances.   Paul H. Black

Rural sanitation

Those procedures, employed in areas outside incorporated cities and not governed by city ordinances, that act on the human environment for the purpose of maintaining or improving public health. Their purpose is the furtherance of community cleanliness and orderliness for esthetic as well as health values.

Water. Purification of water supplies since 1900 has helped to prolong human life more than any other public-health measure. Organisms which produce such diseases as typhoid, dysentery, and cholera may survive for a long time in polluted water, and prevention of water contamination is imperative to keep down the spread of diseases. Watertight covers for wells are important means of preventing surface contamination to the water supply in rural areas.

Purification of a surface water supply, such as from a lake or a farm pond, is accomplished by sedimentation, filtering, and chlorination. Sedimentation can be effected in a storage chamber by adding aluminum sulfate, which flocculates the finer particles of soil and other undesirable matter in suspension in the water. Filtering through fine sand removes the flocculated particles. Finally, the water is purified by a chlorine solution at the rate of $1/2$–1 part of chlorine to 10^6 parts of water. The system of treatment for farm-pond water in **Fig. 1** comprises (1) hedge post or pipe; (2) screen suspended from post 3 ft or 90 cm under water; (3) flexible pipe; (4) hand valve; (5) aspirator, or alum feeder; (6) float valve; (7) hinged wood cover; (8) hand valve; (9) reinforced concrete top; (10) foot valve and strainer; (11) insulated pump house; (12) automatic pump; (13) automatic chlorinator; (14) pressure tank; (15) 2-in. or 5-cm iron pipe or plastic pipe muting; (16) concrete
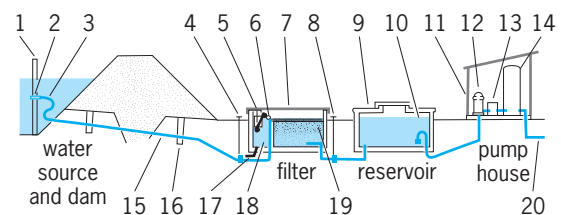

Fig. 1. Farm-pond water-treatment system. The numbers are explained in text.
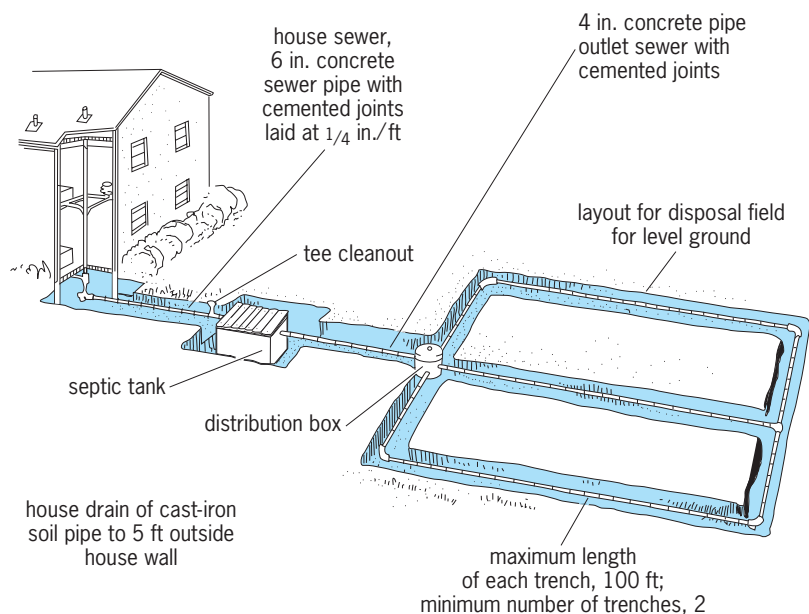
Fig. 2. Typical family-size sewage-disposal system. 1 in. = 2.5 cm; 1 ft = 0.3 m.

cutoff collar; (17) drain when needed; (18) coagulation sedimentation chamber; (19) washed river sand screened through $\frac{1}{8}$-in. or 3-mm sieve; and (20) purified water to house, below frost line.

The colon bacillus is the usual indicator of pollution of water supplies by human waste. Chlorine kills such organisms, and therefore it is widely used for purification of water supplies.

Sewage disposal. The problem of safe disposal of sewage becomes more complex as population increases. The old practice of piping sewage to the nearest body of water has proved to be dangerous. Sanitary engineering techniques are now being used in rural areas, as well as in cities, for the disposal of household and human wastes. Where sewage-plant facilities are not available, the most satisfactory method of sewage disposal is by means of the septic tank system (**Fig. 2**).

The septic tank system makes use of a watertight tank for receiving all sewage. Bacterial action takes place in the septic tank and most of the sewage solids decompose, are given off as gases, or go out into the drainage lines as liquid. The gas and liquid are then released from the top 2 ft (60 cm) of soil without odor or sanitary problems. The solids that do not decompose settle to the bottom of the tank, where they can be easily removed and disposed of safely. Such a sewage-disposal system has made it possible for all farm homes and rural communities to have modern bathroom equipment and sanitary methods of sewage disposal. *See* SEPTIC TANK; SEWAGE DISPOSAL.

Harold E. Stover

Rutabaga

The plant *Brassica napobrassica*, a cool-season, hardy biennial crucifer of European origin, belonging to the order Capparales and probably resulting from the natural crossing of cabbage and turnip. Unlike turnip, it has smooth nonhairy leaves and 38 chromosomes. Propagation is by seed, commonly sown in early summer. The fleshy roots (see **illus.**) are cooked and eaten mashed as a vegetable. Rutabagas have been widely grown as a livestock feed in northern Europe and eastern Canada. Popular yellow-fleshed varieties (cultivars) are Laurentian and American Purple Top; a leading white variety is Macomber. Rutabagas have a high requirement for boron. High temperatures cause misshapen root growth. Commercial production is limited to Canada and the northern part of the United States.

Fleshy roots and leaves of rutabaga plants. (*Asgrow Seed Co., Subsidiary of The Upjohn Company*)

Harvesting generally begins after frost and when the roots are 4–6 in. (10–15 cm) in diameter, commonly 90–100 days after planting. The disease clubroot and the root maggot insect are the most common problems. For diseases of rutabaga *see* TURNIP; CABBAGE; CAPPARALES.   H. John Carew

Ruthenium

A chemical element, Ru, atomic number 44. The element is a brittle gray-white metal of low natural abundance, usually found alloyed with other platinum metals in nature. Ruthenium occurs as seven

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

stable isotopes, and more than ten radioactive (unstable) isotopes are known. Four allotropes of the metal are known. Ruthenium is a hard white metal, workable only at elevated temperatures. It can be melted with an electric arc or an electron beam. *See* ISOTOPE; METAL; PERIODIC TABLE; RADIOISOTOPE; TRANSITION ELEMENTS.

The metal is not oxidized by air at room temperature, but it does oxidize to give a surface layer of ruthenium dioxide (RuO_2) at about 900°C (1650°F); at about 1000°C (1830°F) the volatile compounds ruthenium tetraoxide (RuO_4) and ruthenium monoxide (RuO) form, which can result in loss of the metal. Metallic ruthenium is insoluble in common acids and aqua regia up to 100°C (212°F). The principal properties of ruthenium ore given in the **table**. *See* AQUA REGIA.

Ruthenium is relatively rare, having a natural abundance in the Earth's crust of about 0.0004 part per million. It is always found in the presence of other platinum metals. The major commercial sources of the element are the native alloys osmiridium and iridiosmium and the sulfide ore laurite. The element is also separated from other platinum metals by an intricate process, involving treatment with aqua regia (in which ruthenium, osmium, rhodium, and iridium are insoluble), to yield the pure metal.

Ruthenium is used commercially to harden alloys of palladium and platinum. The alloys are used in electrical contacts, jewelry, and fountain-pen tips. Application of ruthenium to industrial catalysis (hydrogenation of alkenes and ketones) and to automobile emission control (catalytic reduction of nitric oxide) and detection have been active areas of research. In medicine, ruthenium complexes have attracted some attention as potential antitumor reagents and imaging reagents. Ruthenium tetraoxide is finding increasing use as an oxidant for organic compounds. *See* CATALYSIS; CATALYTIC CONVERTER; HYDROGENATION; OSMIUM; PALLADIUM; PLATINUM; TECHNETIUM.   Carol Creutz

Bibliography. C. Bruneau et al. (ed.), *Ruthenium Catalysts and Fine Chemistry*, Springer, 2005; F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; J. R. Davis (ed.),

Principal properties of ruthenium	
Property	Value
Atomic number	44
Atomic weight	101.07
Crystal structure	Hexagonal close-packed
Lattice constant a at 25°C (77°F), nm c/a at 25°C (77°F)	0.27056
	1.5820
Density at 25°C (77°F), g/cm^3	12.37
Thermal neutron capture cross section, barns (10^{-28} m^2)	2.50
Melting point	2310°C (4190°F)
Boiling point	4080°C (7380°F)
Specific heat at 0°C, cal/g (J/kg)	0.0551 (231)
Thermal conductivity, 0–100°C, cal cm/cm^2°C	0.25
Linear coefficient of thermal expansion, 20–100°C, μin./(in.)(°C)	
or μm/(m)(°C)	9.1
Electrical resistivity at 0°C, microhm-cm	6.80
Temperature coefficient of electrical resistance, 0–100°C/°C	0.0042
Tensile strength (annealed), kN · m^{-2}	4.96×10^5
Young's modulus at 20°C (68°F), lb/in.^{2} (Pa)	
Static	60×10^6 (4.1×10^{11})
Dynamic	69×10^6 (4.75×10^{11})
Vickers hardness number (diamond pyramid hardness)	200–350

Metals Handbook: Desk Edition, 2d ed., ASM International, 1998; E. A. Seddon and K. R. Seddon, *The Chemistry of Ruthenium*, Elsevier, 1984.

Rutherfordium

A chemical element, symbol Rf, atomic number 104. Rutherfordium is the first element beyond the actinide series. In 1964 G. N. Flerov and coworkers at the Dubna Laboratories in Russia claimed the first identification of rutherfordium. A. Ghiorso and coworkers made a definitive identification at the Lawrence Radiation Laboratory, University of California, Berkeley, in 1969. *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

The Dubna group claimed the preparation of rutherfordium, mass number 260, by irradiating plutonium-242 with neon-22 ions in the heavy-ion cyclotron. The postulated nuclear reaction was ^{242}Pu $+$ ^{22}Ne $\rightarrow$ ^{260}Rf $+$ 4 neutrons. By 1969 the Berkeley group had succeeded in discovering two alpha-emitting isotopes of rutherfordium with mass numbers 257 and 259 by bombarding ^{249}Cf with ^{12}C and ^{13}C projectiles from the Berkeley heavy-ion linear accelerator (HILAC).

A number of years after the discovery at Berkeley, a team at Oak Ridge National Laboratory confirmed discovery of the isotope ^{257}Rf by detecting the characteristic nobelium x-rays following alpha decay. *See* ACTINIDE ELEMENTS; NOBELIUM; TRANSURANIUM ELEMENTS.

Albert Ghiorso

Bibliography. S. Hofmann, *On Beyond Uranium: Journey to the End of the Periodic Table*, 2002; G. T. Seaborg and W. D. Loveland, *The Elements Beyond Uranium*, 1990.

Rydberg atom

An atom which possesses one valence electron orbiting about an atomic nucleus within an electron shell well outside all the other electrons in the atom. Such an atom approximates the hydrogen atom in that a single electron is interacting with a positively charged core. Early observations of atomic electrons in such Rydberg quantum states involved studies of the Rydberg series in optical spectra. Electrons jumping between Rydberg states with adjacent principal quantum numbers, n and $n - 1$, with n near 80 produce microwave radiation. Microwave spectral lines due to such electronic transitions in Rydberg atoms have been observed both in laboratory experiments and in the emissions originating from certain low-density partially ionized portions of the universe called HII regions. *See* ELECTRON CONFIGURATION; INTERSTELLAR MATTER; MICROWAVE.

The valence electron's classical orbit radius increases as n^2, while the orbital velocity decreases as n^{-1}. The binding energy between the electron and the core decreases as n^{-2}. Thus Rydberg atoms with $n = 80$ have diameters of 10^{-6} m and are as large as some bacteria. Yet they are very delicate, being easily distorted by weak electric and magnetic fields. They are also easily destroyed in collisions with other atomic particles. The natural lifetime of an undisturbed Rydberg atom increases as n^3 for a given electronic angular momentum and has millisecond values for n near 80.

Production and applications. The advent of the laser has made possible the production of sizable numbers of Rydberg atoms within a bulb containing gas at low pressures, 10^{-2} torr (1.3 pascals) or less. The rapid energy-resonance absorption of several laser light photons by an atom in its normal or ground state results in a Rydberg atom in a state with a selected principal quantum number. Aggregates of Rydberg atoms have been used as sensitive detectors of infrared radiation, including thermal radiation. They have also been observed to collectively participate in spontaneous photon emission, called superradiance. Such aggregates form the active medium for infrared lasers that operate through the usual laser mechanism of collective stimulated photon emission. All these developments are based upon the great sensitivity of Rydberg atoms to external electromagnetic radiation fields. Atoms with n near 40 can absorb almost instantaneously over a hundred microwave photons and become ionized at easily achievable microwave power levels. Isotope separation techniques have been developed that combine the selectivity of laser excitation of Rydberg states with the ready ionizability of Rydberg atoms. Such applications have been pursued for atoms ranging from deuterium through uranium. *See* INFRARED RADIATION; ISOTOPE SEPARATION; LASER.

Classical behavior. Quantum systems that have large values for their quantum numbers have many properties that are adequately described by the equations of classical physics. Thus the picture of an atomic electron orbiting in an elliptical trajectory about the nucleus can be increasingly appropriate as n is increased. Experiments can produce localized Rydberg electron wave packets that actually orbit a nucleus. These employ a short laser pulse to initially create an appropriate coherent combination of Rydberg atom energy eigenstates. Although the center of the packet follows the classical electron trajectory, the shape of the packet exhibits quantum wave

interference effects. These include time recurrences of the initial packet localization. The wave effects take longer times to develop as n is increased. This also occurs theoretically when the classical limit is approached by letting Planck's constant pass to zero. As classical physics is considered to be a limiting case of quantum physics, there should be correspondences between quantum and classical quantities and phenomena. *See* QUANTUM MECHANICS.

Quantum chaos. The field of quantum chaos is concerned with those properties of quantum systems that are related to chaos in the classical limit. In this limit, several perturbed Rydberg atom systems exhibit deterministic chaos, a randomlike property characterized by an extreme sensitivity of time evolution to initial conditions. In these chaotic classical systems, it is impossible to predict the future of the Rydberg atom's electron without actually following its evolution. The classical differential equations of motion must be solved continually, by using a computer that must be increasingly precise as longer and longer times are considered.

When a highly excited hydrogen atom is in a sufficiently strong external static magnetic field or microwave field, the competition between the external force on the electron and the atomic Coulomb force on it is so great that the trajectory of the electron becomes chaotic. The optical absorption spectrum of the real quantum-mechanical atom in the external field then exhibits a special distribution of nearest-neighbor energy-level spacings. This distribution is one manifestation of chaos in the classical limit. Another arises when a strong microwave field is applied for a very short time of about 100 microwave oscillation periods. During this time the atom with an initially well-defined value for n quickly evolves into a probability distribution in n that then broadens with time according to a classical diffusion equation, the second manifestation of chaos. A third is that the diffusion stops when a destructive quantum wave interference becomes complete. The interference is highly destructive because of the random character of the underlying classical chaos. All three effects related to classical chaos have been observed in quantal computer calculations that have been confirmed by laboratory experiments. However, the property of quantum systems that corresponds to chaotic behavior in the classical limit is yet undiscovered. *See* CHAOS.

Related systems. The positively charged core attracting a Rydberg valence electron may be a molecular ion rather than an atomic ion. Even more complex Rydberg-like quantum states exist in solids, in which a valence electron can be weakly bound either to an impurity ion or to a hole vacancy in a semiconductor. Thus many of the phenomena observed in studies with Rydberg atoms also can occur in molecules and in the solid state. The consequences of this for physics, chemistry, biology, and technology largely remain unexplored. *See* CRYSTAL DEFECTS; SEMICONDUCTOR.

Experiments have been carried out on doubly excited atoms, in which laser radiation is used to drive simultaneously two atomic electrons out of their normal atomic shells into different Rydberg valence shells. The resulting planetary atoms have two or more valence electrons in widely separated classical orbits. *See* ATOMIC STRUCTURE AND SPECTRA.

James E. Bayfield

Bibliography. G. F. Bassani, M. Inguscio, and T. W. Hansch, *The Hydrogen Atom,* 1989; G. Casati et al., Relevance of classical chaos in quantum mechanics: The hydrogen atom in a monochromatic field, *Phys. Rep.,* 154:77–123, 1987; T. G. Gallagher, *Rydberg Atoms*, 1994; P. W. Milonni, M. L. Shih, and J. R. Ackerhalt, *Chaos in Laser-Matter Interactions*, 1987; R. F. Stebbings and F. B. Dunning, *Rydberg States of Atoms and Molecules*, 1983.

Rydberg constant

The most accurately measured of the fundamental constants; it is a universal scaling factor for any spectroscopic transition and an important cornerstone in the determination of other constants.

Empirical definition. This constant was introduced empirically. J. Balmer's formula described the visible spectral lines of atomic hydrogen, while J. Rydberg's formula applied to the spectra of many elements. Their results may be summarized by Eq. (1), where

$$\frac{1}{\lambda} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right) \tag{1}$$

λ is the wavelength of the spectral line and R is a constant. In Balmer's account of the visible hydrogen spectrum, n_1 was equal to 2, while n_2 took on the integer values 3, 4, 5, and so forth. In Rydberg's more general work, n_1 and n_2 differed slightly from integer values. A remarkable result of Rydberg's work was that the constant R was the same for all spectral series he studied, regardless of the element. This constant R has come to be known as the Rydberg constant.

Atomic theory. Applied to hydrogen, Niels Bohr's atomic model leads to Balmer's formula with a predicted value for the Rydberg constant given by Eq. (2), where m_e is the electron mass, e is the elec-

$$R_\infty = \frac{m_e e^4}{8 h^3 \epsilon_0^2 c} \tag{2}$$

tron charge, h is Planck's constant, ϵ_0 is the permittivity of vacuum, and c is the speed of light. The equation expresses the Rydberg constant in SI units. To express it in cgs units, the right-hand side must be multiplied by $(4\pi\epsilon_0)^2$. The subscript ∞ means that this is the Rydberg constant corresponding to an infinitely massive nucleus. To account for the fact that the electron and the nucleus of finite mass m_N orbit around a common center of gravity, the electron mass m_e must be replaced by the reduced mass given by Eq. (3). This changes the energy levels by

$$\mu = \frac{m_e m_N}{m_e + m_N} \tag{3}$$

the ratio of electron mass to nuclear mass, which

is about 1/2000 in the case of hydrogen. This prediction was found in excellent agreement with the empirical value found by Rydberg.

E. Schrödinger's wave mechanics predicts the same energy levels as the simple Bohr model, but the relativistic quantum theory of P. A. M. Dirac introduces small corrections or fine-structure splittings of order α^2, where α is the fine-structure constant and is approximately 1/137. The modern theory of quantum electrodynamics predicts further corrections of order α^3 and higher. Additional small hyperfine-structure corrections account for the interaction of the electron and nuclear magnetic moments. *See* FINE STRUCTURE (SPECTRAL LINES); HYPERFINE STRUCTURE; NONRELATIVISTIC QUANTUM THEORY; QUANTUM ELECTRODYNAMICS; RELATIVISTIC QUANTUM THEORY.

The Rydberg constant plays an important role in its relationship to other fundamental constants. It may, for instance, be expressed in terms of the fine-structure constant α or the Bohr radius a_0 by Eqs. (4) and (5). Because of its extreme accuracy, the

$$R_\infty = \frac{m_e c \alpha^2}{2b} \qquad (4)$$

$$R_\infty = \frac{\alpha}{4\pi a_0} \qquad (5)$$

Rydberg constant gives precise relations between these constants that must be satisfied by any set of constants that may be adopted.

Measurement. The Rydberg constant is determined by measuring the wavelength or frequency of a spectral line of a hydrogenlike atom or ion. The energy levels of other atoms are also proportional to the Rydberg constant, but the difficulty of many-body calculations precludes computations with the same precision as is possible for simple hydrogenic atoms.

Conventional measurements. Prior to 1974, the value of the Rydberg constant was based on wavelength measurements of spectral lines (hydrogen, deuterium, tritium, and singly ionized helium) observed in the emission of gas discharges. The wavelengths are measured interferometrically by comparing the fringe patterns they produce with that of a known wavelength standard. The major limitation has been the Doppler broadening of the spectral lines due to the random thermal motion of the gas atoms. This broadening makes it impossible to resolve the different fine-structure components (**illus.** *a*) even when a deuterium discharge is cooled to liquid-nitrogen temperature (illus. *b*), and it has prevented any significant improvement of such wavelength measurements beyond 1 part in 10^7. A Rydberg value $R_\infty = 10{,}973{,}731.77 \pm 0.83$ m^{-1}, based on conventional measurements, was adopted in the 1973 adjustment of the fundamental constants. *See* DOPPLER EFFECT; WAVELENGTH STANDARDS.

Laser spectroscopy. Major improvements in the accuracy of the Rydberg constant became possible when techniques of laser spectroscopy were introduced which could overcome the Doppler broad-
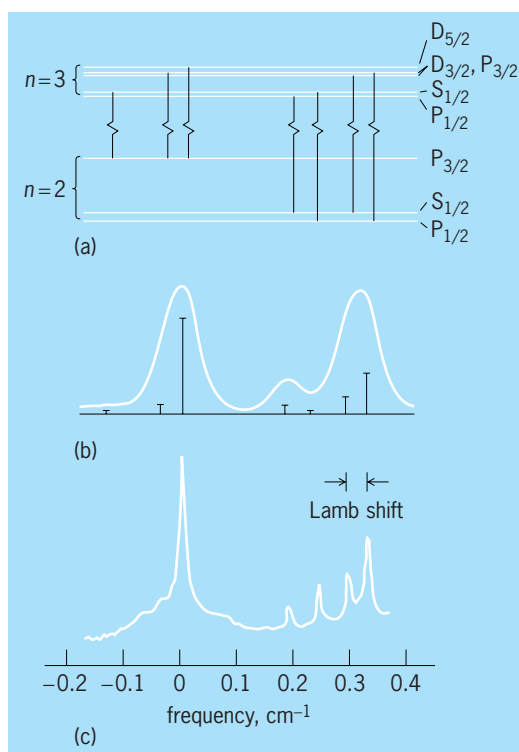
Deuterium Balmer line D$_\alpha$. (*a*) Energy levels with fine-structure transitions. (*b*) Emission-line profile of a cooled deuterium gas discharge (temperature 50 K or −370°F) and theoretical fine-structure lines with relative transition probabilities (*from B. P. Kibble et al., J. Phys., B6:1079–1089, 1973*). (*c*) Saturation spectrum with optically resolved Lamb shift (*from T. W. Hänsch et al., Precision measurement of the Rydberg constant by laser saturation spectroscopy of the Balmer α line in hydrogen and deuterium, Phys. Rev. Lett., 32:1336–1340, 1974*).

ening of hydrogen spectral lines. In the first laser measurement of the Rydberg constant in 1974, saturation spectroscopy was introduced to resolve single fine-structure components of the red Balmer-α line (illus. *c*). An intense, highly monochromatic beam from a tunable dye laser was used, in effect, to label atoms of slow velocity in a glow discharge, by exciting and thus removing them temporarily from the absorbing lower level. A signal from these atoms was registered as a bleaching effect for a weaker counterpropagating probe beam. An almost tenfold better Rydberg value, $R_\infty = 10{,}973{,}731.406 \pm 0.100$ m^{-1}, was obtained.

The highest resolution and accuracy has been achieved by the method of Doppler-free two-photon spectroscopy, which permits the observation of very sharp resonance transitions between long-living states. The atoms are excited by two counterpropagating laser beams so that two photons, one from each beam, provide the required excitation energy. From a moving atom, the two beams show opposite Doppler shifts, so that the Doppler effect is canceled to first order. This technique was first applied around 1975 to the transition from the hydrogen 1S ground state to the metastable 2S state, which has a natural linewidth of only 1.3 Hz. Such laser experiments were the basis for a Rydberg value,

$R_\infty = 10{,}973{,}731.534 \pm 0.013 \text{ m}^{-1}$, that was adopted in the 1986 adjustment of the fundamental constants.

Precision measurements. Since 1986, several Rydberg constant measurements have been performed in different laboratories. Aside from improvements in spectroscopic techniques, the experimental precision has benefited from the development of methods to phase-coherently measure and compare optical frequencies. With this increased precision, the accurate knowledge of quantum-electrodynamic contributions and corrections due to the finite size of the proton has been increasingly important in the determination of the Rydberg constant. These contributions are summarized in the Lamb shift of an energy level. While hydrogenic Lamb shifts, notably the $2S_{1/2}$–$2P_{1/2}$ splitting (illus. *c*), were traditionally measured by radio-frequency spectroscopy, more accurate Lamb-shift values have been obtained by comparison of different optical transition frequencies of the hydrogen atom. By themselves, Lamb-shift measurements provide stringent tests of quantum-electrodynamic theory, approaching the limits imposed by the uncertainty of the proton charge radius.

In one series of experiments, the hydrogen 1*S*-2*S* transition was excited in a cold beam of ground-state atoms in a collinear excitation geometry with the frequency-doubled light of a continuous-wave dye laser. In one of these experiments, the laser frequency was compared to the standard radio frequency of a cesium atomic clock with the help of a wide comb of precisely equidistant laser frequencies emitted by a mode-locked laser. This enabled a determination of the frequency of the hydrogen 1*S*-2*S* transition to a precision of 1.8 parts in 10^{-14}. In other experiments, Doppler-free two-photon spectroscopy on a metastable hydrogen beam was used to observe the transitions 2*S*-8*S*, 2*S*-8*D*, and 2*S*-12*D*. With a harmonic frequency chain, accuracies near 8 parts in 10^{-12} were obtained for the absolute frequencies of the transitions 2*S*-8*D* and 2*S*-12*D*. *See* ATOMIC CLOCK; MOLECULAR BEAMS; PHYSICAL MEASUREMENT.

By combining the result of, for example, the hydrogen 1*S*-2*S* two-photon transition frequency with that of 2*S*-8*D*, values for both the Rydberg constant and the 1*S* Lamb shift are obtained. The 2002 adjustment of the fundamental constants, taking into account different measurements, adopted the value $R_\infty = 10{,}973{,}731.568{,}525 \pm 0.000{,}073 \text{ m}^{-1}$ for the Rydberg constant. The measurements provide an important cornerstone for fundamental tests of basic laws of physics. *See* ATOMIC STRUCTURE AND SPECTRA; FUNDAMENTAL CONSTANTS; LASER; LASER SPECTROSCOPY.   Theodor W. Hänsch; Martin Weitz

Bibliography. G. F. Bassani, M. Inguscio, and T. W. Hänsch (eds.), *The Hydrogen Atom*, 1989; T. W. Hänsch, A. L. Schawlow, and G. W. Series, The spectrum of atomic hydrogen, *Sci. Amer.*, 240(3):94–110, 1979; P. J. Mohr and B. N. Taylor, CODATA recommended values of the fundamental physical constants: 2002, *Rev. Mod. Phys.*, 77:1–107, 2005; M. Niering et al., Measurement of the hydrogen 1*S*-2*S* transition frequency by phase coherent comparison with a microwave cesium fountain clock, *Phys. Rev. Lett.*, 84:5496–5499, 2000; C. Schwob et al., Optical frequency measurement of the 2*S*-12*D* transitions in hydrogen and deuterium: Rydberg constant and Lamb shift determinations, *Phys. Rev. Lett.*, 82:4960–4963, 1999; G. W. Series (ed.), *The Spectrum of Atomic Hydrogen: Advances*, 1988; *Units and Fundamental Constants in Physics and Chemistry*, subvol. b, *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology*, 1992.

Rye

A winter-hardy and drought-resistant cereal plant, *Secale cereale*, in the grass family (Graminae). It resembles wheat, with which it intercrosses to a limited extent. Rye is propagated almost completely by cross-pollination; it is partly sterile if a plant is made to self-pollinate. The inflorescence is a spike or ear (**Fig. 1***a*). Spikelets are arranged flatwise against a zigzag rachis (**Fig. 2**); they usually have two flowers, enclosed by a lemma and palea with two adjacent glumes. The young florets contain three stamens and a pistil. The fertilized pistil develops into a naked grain, or kernel (Fig. 1*b* and *c*), that is easily threshed. There are several recognized species of *Secale*, most of which have shattering spikes and small kernels. There are both perennial and winter-annual species of rye, with winter forms being favored over spring types for production. The only commercially cultivated species is the nonshattering *S. cereale*. Ergot sclerotia often are evident in the field and in threshed grain (Fig. 1*d*). *See* CYPERALES; WHEAT.

Production. Rye is more important in Europe and Asia than in the Western Hemisphere. Russia is the leading world producer, followed by Poland and Germany. Canada and Argentina produce significant
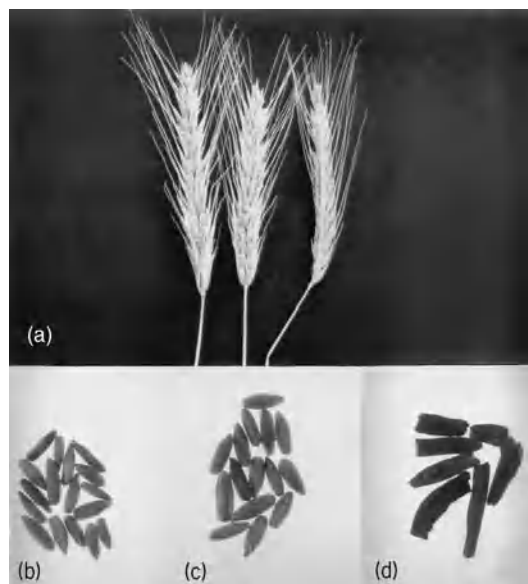

Fig. 1. Rye. (*a*) Spikes or ears. (*b*) Diploid kernels. (*c*) Tetraploid kernels. (*d*) Ergot.
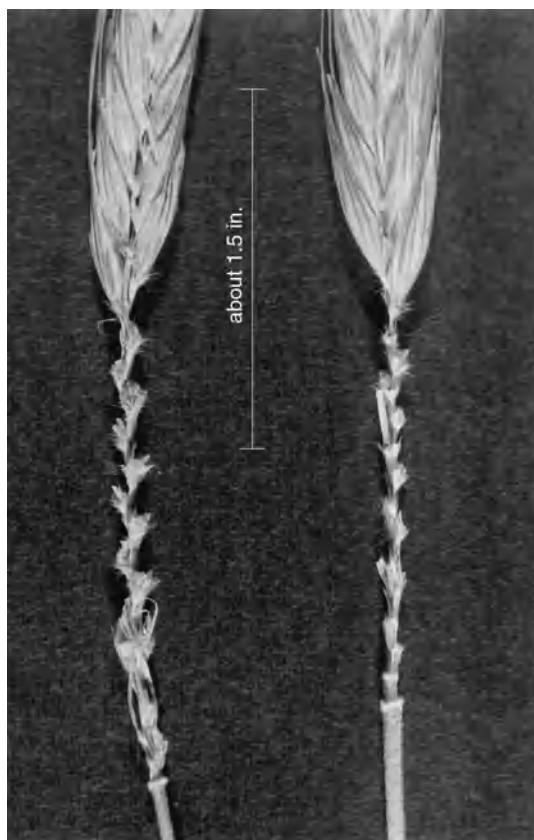
Fig. 2. Florets and zigzag rachis of rye. (*Wisconsin Agricultural Experiment Station*)

amounts, and Switzerland and northwest Europe have high yields. Rye production in the United States is mostly in South Dakota, North Dakota, Minnesota, and Georgia.

Uses. Rye grain is used for animal feed, human food, and production of spirits. Ground rye is mixed with other feeds for livestock. It is often fall-sown to provide soil cover and pasturage for livestock. Egg yolks of chickens and butter from cows fed on rye have a rich yellow color. *See* DISTILLED SPIRITS.

Next to wheat, rye has the most desirable gluten for breadmaking. Rye bread is a staple food in certain sandy land areas of Europe and is favored by peoples of Slavic origin. In the United States, rye bread is made from blends of rye and wheat flour. Rye doughs lack gas retention, resulting in compact loaves or heavy bread. Rye flour contains several B vitamins and as many as 18 amino acids. Rye of different production conditions has 30% more lysine than spring wheats and durums, grown mostly in North Dakota, and somewhat more lysine than Triticales. Dark flours contain a higher content of amino acids than the lighter, refined flours.

Origin. N. I. Vavilov, a Russian plant scientist, believed that rye was introduced into cultivation simultaneously and independently at many localities in central Asia or Asia Minor. Great botanical diversity is found in Transcaucasia, Iran, Turkestan, and Asia Minor, where rye persists without being cultivated. Weedy rye types have closely investing lem-

mas with partially shattering spikes. Others resemble cultivated rye and are said to have high grain protein; their kernels are not well filled. Rye probably spread as mixtures of wheat.

Varieties. Compared to other small grains, rye has a fewer number of cultivars (agricultural varieties). Some common varieties grown in Europe and Russia are Belta, Danae, Dominant, Golden Dankow, Petkus, Petkus short straw, Sangaste, Toivo, Von Lochow, Vyatka, and Vyatka 2. Some common varieties in North America are Abruzzi, Adams, Cougar, Gator, Rymin, Von Lochow, and Weser. Short-strawed types are gaining favor. Plant and kernel characteristics of rye are variable, partly because of cross-pollination. Height may range from 4 to 6 ft (120 to 180 cm) under moderately fertile conditions. Kernel color may be amber, gray, green, blue, brown, or black.

Tetraploid forms, whose chromosome number has been doubled, are available. These kernels are large (Fig. 1*c*) and the straw is usually stiff, yet this type has not found much commercial use. Tetraploid wheat and rye have been hybridized and chromosomes doubled to form Triticales, which is increasing in usage. *See* BREEDING (PLANT).

Hybrid rye. Heterosis in rye, though not readily usable, has long been recognized. Attempts to find cytoplasmic male sterility in cultivated types have been made in Russia and Europe. Fertility restorers have been found. Hybridizing proper combinations of cytoplasmic male sterility lines and re storers could enhance productivity and use of this crop. *See* HETEROSIS.

Cultural practices. Culture of rye is similar to that of wheat. Seed-bed preparation is kept to a minimum, usually utilizing sandy soil. Plowing the land is preferable, but rye is sometimes drilled directly into stubble. Seed can be sown by hand or with a regular drill at a rate of 56–85 lb/acre (63–95 kg/ha). Rye seed deteriorates during storage; therefore, it should be tested for germinability before sowing. Sowing in the Northern Hemisphere may be done from August to November, but before sustained freezing temperatures. Spring-sown rye is less important than winter rye. Rye may be plowed under for green manure, harvested for hay, or threshed for grain and straw. Grain should be 13.5% or lower in moisture and stored under dry conditions. *See* AGRICULTURAL SOIL AND CROP PRACTICES; GRAIN CROPS.   H. L. Shands

Diseases. Ergot is the most important disease of rye (occurs also in wheat) because of the presence of toxic alkaloids in the sclerotia that may be present in the grain. Grain containing ergot sclerotia is discounted heavily when marketed in the United States and many other countries. Grain-cleaning equipment can be used to remove the sclerotia to within tolerance levels of marketing grades and pure food laws. The disease-causing organism, *Claviceps purpurea*, is usually most severe on rye and Triticale, but is also found on wheat, barley, and such grasses as *Bromus* and *Agropyron* species. The disease is recognized by the purple-black, curved sclerotia which protrude from the glumes of infected flowers. *See* ERGOT AND ERGOTISM.

Fig. 3. Ergot of rye. Sclerotium is "germinating" to produce perithecial stromata.

The organism is carried from one season to the next by the sclerotia (ergot bodies), which fall to the ground before harvest or are planted with the seed grain, usually in the fall. In the next spring, when the host plants are flowering, the sclerotia in the ground "germinate" to produce stalks and spherical heads (**Fig. 3**) from which spores (ascospores) are ejected. These spores are carried by air currents to open grain flowers, where infection of the ovaries occurs. As a result of this infection, a second spore stage (conidia) is produced in honeylike droplets. Insects are attracted to these droplets and then visit other grain flowers, secondarily spreading the disease organism. Eventually, the sclerotia replace the kernels in infected flowers. The sclerotia may be up to 2 in. (5 cm) in length and are horny in texture.

Rye is particularly susceptible to ergot because it is normally cross-pollinated, and the flowers tend to be open longer than in self-pollinated grains such as barley and wheat. The longer the flowers are open, the more chance there is for the disease to develop. This is particularly true in thin plantings (where each plant develops many tillers that flower over a long period of time) or in plantings of mixed varieties (which flower at different times).

Control of ergot therefore involves a relatively heavy planting rate with seed of a pure variety. Seed should not contain sclerotia, and rotation of fields may be needed to avoid sclerotia in the soil. Deep plowing would bury the sclerotia so that primary inoculum would not be produced. Sclerotia formed in grasses in fence rows or meadows adjacent to rye fields may also serve as sources of inoculum. Mowing of these grasses to prevent flowering will eliminate this source.

Several other diseases caused by fungi are found on rye and, although usually of minor importance, may be problems at times. These include rusts of stems and leaves (caused by *Puccinia graminis* and *P. recondita*), scab (caused by *Gibberella zeae*) on heads and kernels, and leaf scald (caused by *Rhynchosporium secalis*). Control of these diseases involves resistance, rotation, and seed treatment. Bacteria and nematode diseases have also been described. *See* PLANT PATHOLOGY.   Deane C. Arny

Processing. Rye grain is milled into flour in a manner similar to that used for wheat flour. Several noted variations are made based on the compositional and structural differences between these grains. The careful sizing of rye kernels prior to breaking is important due to greater size and shape variability of this grain. Rye grain of sound, plump character without presence of mold or ergot is essential to yield a high-quality flour. Cleaning operations for rye are more extensive than those required for wheat. Removal of dirt, stones, and foreign material is accomplished by gravity separation and washing. Washed grain is commonly tempered to a moisture content of approximately 14.5% prior to milling.

Break roll operations typically release the greatest portion of flour. The break roll operation results in a limited quantity of rye middlings (intermediate-size rye particles) due to the softness and starchiness of the endosperm. Middling purifiers are not utilized extensively in rye milling. Sifting operations for rye require greater sifting surface area than is typically found in comparable wheat flour sifting. Reduction rolls are also corrugated to minimize flaking of rye break streams (rye particles produced during the first break roll operation) which would inhibit release of flour. The yield of rye flour will generally range from 75 to 85% extraction.

Grades of rye flour are based primarily on color rather than on ash or protein content. Rye flours are commonly described by color, ranging from white (light) to cream (medium) to dark, as the extraction rate increases from 60% to 85%. The desirable rye flavor also increases with extraction rate; thus high-extraction flours possess desirable attributes which have increasingly been utilized in variety breads.

Rye flour doughs lack mixing tolerance and do not develop strong gluten. Rye bread production requires blending of rye flours with wheat flours (commonly clear grades are utilized because color is generally not a limitation) to provide sufficient dough strength. Rye breads are commonly produced from doughs ranging from 20 to 40% rye flour without appreciable loaf volume depression. Specialty varieties of rye breads are classified according to ethnic origins or as sweet or sour doughs. Sour rye breads may be developed from natural lactic fermentations or through the incorporation of cultured milk.

Swedish rye crisp breads are generally prepared from whole ground meal. A mixture of whole meal, yeast, salt, and water may be fermented, thinly sheeted, and cut prior to baking. A short-time–high-temperature baking at 450°F (230°C) for 10 min, followed by additional drying at 200°F (93°C) for 3 h, results in the firm crisp product. *See* FOOD ENGINEERING.   Mark A. Uebersax

Bibliography. W. Bushuk, *Rye: Production, Chemistry and Technology*, 2d ed., 2001; J. G. Dickson, *Diseases of Field Crops*, 2d ed., 1956; R. Gair, J. E. Jenkins, and E. Lester *Cereal Pests and Diseases*, 4th ed., 1987; H. H. Geiger and F. W. Schnell, Cytoplasmic male sterility in rye (*Secale cereale* L.), *Crop Sci.*, 10:590–593, 1970; B. Godon and C. Willin (eds.), *Primary Cereal Processing: A Comprehensive*

Sourcebook, 1993; N. L. Kent, Rye, triticale, rice: Processing, nutritional attributes, technological uses, in *Technology of Cereals with Special Reference to Wheat*, 2d ed., pp. 256–276, 1975; N. L. Kent and A. D. Evers, *Technology of Cereals*: *An Introduction for Students of Food Science and Agriculture*, 4th ed., 1994; W. H. Leonard and J. H. Martin, Rye, in *Cereal Crops*, 1963; K. Lorenz and K. Kulp (eds.), *Handbook of Cereal Science and Technology*, 1990; H. H. Schopmeter, Rye and rye milling, *Cereal Sci. Today*, 7(5):138, 1962; M. V. Wiese, *Compendium of Wheat Diseases*, 2d ed., 1987.